



(51) International Patent Classification:

G06G 7/58 (2006.01)

G01N 33/566 (2006.01)

(21) International Application Number:

PCT/US2021/034637

(22) International Filing Date:

27 May 2021 (27.05.2021)

(25) Filing Language:

English

(26) Publication Language:

English

(30) Priority Data:

63/031,957

29 May 2020 (29.05.2020)

US

(71) Applicant: THE BOARD OF TRUSTEES OF THE LE-
LAND STANFORD JUNIOR UNIVERSITY [US/US];Building 170, Third Floor, Main Quad, P.O. Box 20386,
Stanford, CA 94305-2038 (US).(72) Inventors: DROR, Ron, O.; Building 170, Third Floor,
Main Quad, P.O. Box 20386, Stanford, CA 94305-2038(US). PAGGI, Joseph, M.; Building 170, Third Floor,
Main Quad, P.O. Box 20386, Stanford, CA 94305-2038
(US).(74) Agent: HANS, Christian, S.; KPPB LLP, 2190 S. Towne
Centre Place, Suite 300, Anaheim, CA 92806 (US).

(81) Designated States (unless otherwise indicated, for every kind of national protection available): AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, IT, JO, JP, KE, KG, KH, KN, KP, KR, KW, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, WS, ZA, ZM, ZW.

(84) Designated States (unless otherwise indicated, for every kind of regional protection available): ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, ST, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, RU, TJ, TM), European (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG).

(54) Title: SYSTEMS AND METHODS TO PREDICT STRUCTURES AND PROPERTIES OF BIOMOLECULE-LIGAND COMPLEXES AND USES THEREOF

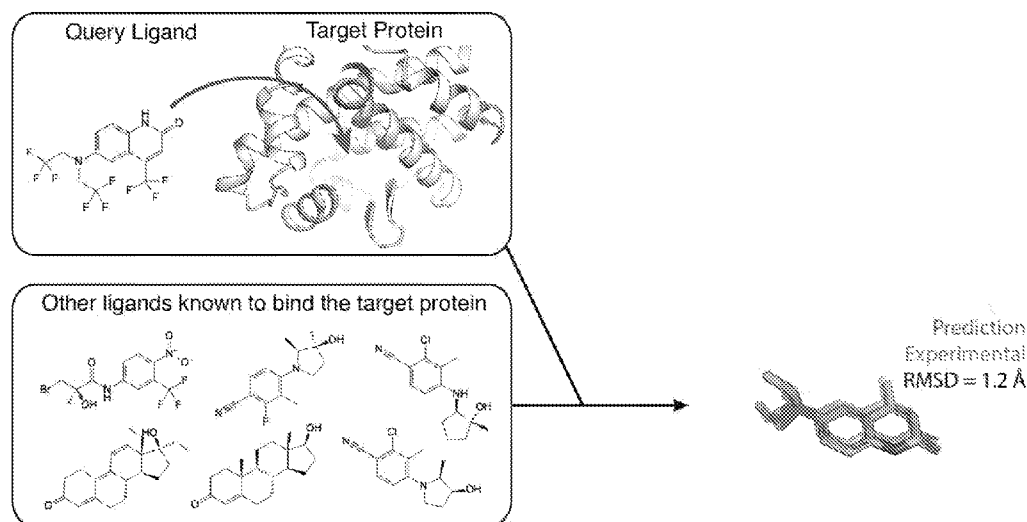


Fig. 1B

(57) Abstract: Embodiments herein describe systems and methods to predict biomolecule-ligand complexes and uses thereof. Many embodiments generate candidate poses and docking scores for ligands known to interact with a target protein as well as a candidate ligand. Based on similarities present in the candidate poses and docking scores, a pose for the candidate ligand is identified. Many embodiments determine whether a candidate ligand can bind to a target and, if so, determine an affinity. Many embodiments can be used for virtual screening or lead optimization.

Published:

- *with international search report (Art. 21(3))*
- *before the expiration of the time limit for amending the claims and to be republished in the event of receipt of amendments (Rule 48.2(h))*

SYSTEMS AND METHODS TO PREDICT STRUCTURES AND PROPERTIES OF BIOMOLECULE-LIGAND COMPLEXES AND USES THEREOF

FIELD OF THE INVENTION

[0001] The current application claims priority to U.S. Provisional Patent Application No. 63/031,957 entitled “Systems and Methods to Predict Protein-Ligand Complexes and Uses Thereof” filed May 29, 2020, the disclosure of which is hereby incorporated by reference in its entirety.

FIELD OF THE INVENTION

[0002] The present invention relates to biomolecule-ligand interactions. More specifically, the present invention relates to a methods and systems to predict biomolecule-ligand complexes showing higher accuracy in determining pose and/or binding affinity.

BACKGROUND

[0003] Binding of small-molecule ligands to proteins is one of the most fundamental processes in biology, and the great majority of drugs are ligands that exert their effects by binding to a target protein. Predicting properties of protein–ligand interactions—including three-dimensional (3D) structures, binding affinities, binding kinetics, selectivity, and functional effects—is critical both to the rational design of effective medicines and to solving major problems in molecular biology. An enormous amount of work has thus focused on the development of computational methods to predict these properties. (See e.g., W. Yu, A. D. MacKerell, Jr., Computer-Aided Drug Design Methods. *Methods Mol Biol* 1520, 85-106 (2017); and G. Sliwoski, S. Kothiwale, J. Meiler, E. W. Lowe, Jr., Computational methods in drug discovery. *Pharmacol Rev* 66, 334-395 (2013); the disclosures of which are incorporated herein by reference in their entireties.)

[0004] Such computational methods generally fall into two categories. “Physics-based” approaches use a 3D structure of the target protein and exploit an understanding of the physics of protein–ligand interactions. (See e.g., J. Li, A. Fu, L. Zhang, An Overview of Scoring Functions Used for Protein-Ligand Interactions in Molecular Docking. *Interdiscip Sci* 11, 320-328 (2019); the disclosure of which is incorporated herein by

reference in its entirety.) “Ligand-based” approaches use experimental measurements for many ligands of a given property (e.g., affinity) at a given target and employ pattern matching to predict the corresponding property for other ligands. (See e.g., S. P. Leelananda, S. Lindert, Computational methods in drug discovery. Beilstein J Org Chem 12, 2694-2718 (2016); and A. Cherkasov et al., QSAR modeling: where have you been? Where are you going to? Journal of medicinal chemistry 57, 4977-5010 (2014); the disclosures of which are incorporated herein by reference in their entireties.) However, such methodologies generally produce inaccurate results, thus are insufficient for broad use.

SUMMARY OF THE INVENTION

[0005] This summary is meant to provide some examples and is not intended to be limiting of the scope of the invention in any way. For example, any feature included in an example of this summary is not required by the claims, unless the claims explicitly recite the features. Various features and steps as described elsewhere in this disclosure may be included in the examples summarized here, and the features and steps described here and elsewhere can be combined in a variety of ways.

[0006] In one embodiment, a method for predicting ligand-protein interaction includes obtaining a set of ligands known to bind to a target protein, generating a set of candidate poses for each ligand in the set of ligands and a candidate ligand, and determining a pose for the candidate ligand based on a similarity of the candidate poses for the candidate ligand and the candidate poses for each ligand in the set of ligands.

[0007] In a further embodiment, generating a set of candidate poses is accomplished by docking at least one ligand from the set of ligands using docking software.

[0008] In another embodiment, the method further includes generating a set of scores for each ligand is using docking software.

[0009] In a still further embodiment, the docking software is selected from Glide, DOCK, AutoDock VINA, FlexX, GOLD, and Surflex.

[0010] In still another embodiment, at least one ligand in the set of ligands has an experimentally determined pose.

[0011] In a yet further embodiment, the pose for the at least one ligand is set to the experimentally determined pose.

[0012] In yet another embodiment, the similarity is based on pairwise comparisons between each ligand in the set of ligands and every other ligand in the set of ligands and pairwise comparisons between the candidate ligand and every ligand in the set of ligands.

[0013] In a further embodiment again, the similarity is further based on pairwise comparisons of substructures between each ligand in the set of ligands and every other ligand in the set of ligands and pairwise comparisons between the candidate ligand and every ligand in the set of ligands.

[0014] In another embodiment again, the similarity is further based on pairwise comparisons of a protein-ligand interaction between each ligand in the set of ligands and every other ligand in the set of ligands and pairwise comparisons between the candidate ligand and every ligand in the set of ligands.

[0015] In a further additional embodiment, the protein-ligand interaction is selected from a hydrogen bond, a salt bridge, and a hydrophobic interaction.

[0016] In another additional embodiment, determining a pose for the candidate ligand includes determining a pose for each ligand in the set of ligands that minimizes the potential of all ligands.

[0017] In a still yet further embodiment, the potential is determined by:

$$E(L) = (n - 1)E^{\text{dock}}(L) + E^{\text{pairwise}}(L),$$

where $L = (\ell_1, \dots, \ell_n)$ indexes the candidate poses for each ligand, n is the number of ligands, E^{dock} represents the standard docking scores for the set of ligand poses, and E^{pairwise} represents a pairwise contribution to the total score.

[0018] In still yet another embodiment, $E^{\text{pairwise}}(L) = \sum_{(i,j), i \neq j} e^{\text{pairwise}}(\ell_i, \ell_j)$.

[0019] In a still further embodiment again, e^{pairwise} is determined using a machine learning algorithm.

[0020] In still another embodiment again, the machine learning algorithm is fit to a benchmark set of pairs of correct poses and pairs of incorrect poses for a diverse set of proteins.

[0021] In a still further additional embodiment, determining a pose for each ligand in the set of ligands includes utilizing an expectation-maximization algorithm.

[0022] In still another additional embodiment, the method further includes altering a pose of a ligand in the set of ligands or the candidate ligand to optimize that ligand's contribution to the potential function.

[0023] In a yet further embodiment again, a method for virtual screening includes obtaining a set of ligands known to bind to a target protein, determining a pose for each ligand in the set of ligands by generating a set of candidate poses for ligands in the set of ligands, where the pose is based on a similarity of the candidate poses for each ligand in the set of ligands, and determining a binding affinity of at least one candidate ligand by determining whether the candidate ligands binds to the target protein based on similarities to each ligand in the set of ligands and determining an energy if the candidate ligand binds to the target ligand.

[0024] In yet another embodiment again, determining a binding affinity of a candidate ligand determines a binding affinity of at least two candidate ligands.

[0025] In a yet further additional embodiment, the method further includes obtaining a library of candidate ligands, and determining a binding affinity of at least one candidate ligand determines a binding affinity of at least one candidate ligand in the library of candidate ligands.

[0026] In yet another additional embodiment, the method further includes fixing a pose for each ligand in the set of ligands prior to determining a binding affinity of at least one candidate ligand.

[0027] In a further additional embodiment again, the similarity is based on pairwise comparisons between each ligand in the set of ligands and every other ligand in the set of ligands.

[0028] In another additional embodiment again, the similarities are further based on pairwise comparisons of substructures between each ligand in the set of ligands and every other ligand in the set of ligands and pairwise comparisons between the candidate ligand and every ligand in the set of ligands.

[0029] In a still yet further embodiment again, the similarities are further based on pairwise comparisons of a protein-ligand interaction between each ligand in the set of ligands and every other ligand in the set of ligands and pairwise comparisons between the candidate ligand and every ligand in the set of ligands.

[0030] In still yet another embodiment again, the protein-ligand interaction is selected from a hydrogen bond, a salt bridge, and a hydrophobic interaction.

[0031] In a still yet further additional embodiment, determining a pose for the candidate ligand includes determining a pose for each ligand in the set of ligands that minimizes the potential of all ligands.

[0032] In still yet another additional embodiment, the potential is determined by:

$$E(L) = (n - 1)E^{\text{dock}}(L) + E^{\text{pairwise}}(L),$$

where $L = (\ell_1, \dots, \ell_n)$ indexes the candidate poses for each ligand, n is the number of ligands, E^{dock} represents the standard docking scores for the set of ligand poses, and E^{pairwise} represents a pairwise contribution to the total score.

[0033] In a yet further additional embodiment again, $E^{\text{pairwise}}(L) = \sum_{(i,j), i \neq j} e^{\text{pairwise}}(\ell_i, \ell_j)$.

[0034] In yet another additional embodiment again, e^{pairwise} is determined using a machine learning algorithm.

[0035] In a still yet further additional embodiment again, the machine learning algorithm is fit to a benchmark set of pairs of correct poses and pairs of incorrect poses for a diverse set of proteins.

[0036] In still yet another additional embodiment again, determining a pose for each ligand in the set of ligands includes utilizing an expectation-maximization algorithm.

[0037] In another further embodiment, the method further includes altering a pose of a ligand in the set of ligands to optimize that ligand's contribution to the potential function.

[0038] In still another further embodiment, a method for lead optimization includes obtaining a set of ligands known to bind to a target protein, determining a pose for each ligand in the set of ligands by generating a set of candidate poses and scores for ligands in the set of ligands, where the pose is based on a similarity of the candidate poses and scores for each ligand in the set of ligands, determining a binding affinity of a candidate ligand by determining whether the candidate ligands binds to the target protein based on similarities to each ligand in the set of ligands and determining an energy if the candidate ligand binds to the target ligand, and modifying a location of the candidate ligand.

[0039] In yet another further embodiment, the method further includes identifying a modifiable location on the candidate ligand, and the modifying a location step includes modifying the modifiable location.

[0040] In another further embodiment again, modifying the location increases binding affinity of the candidate ligand.

[0041] Another further additional embodiment, the method further includes determining a binding affinity of the modified candidate ligand to the target protein based on similarities to each ligand in the set of ligands and determining an energy if the candidate ligand binds to the target ligand.

[0042] In a further embodiment, the similarity is based on pairwise comparisons between each ligand in the set of ligands and every other ligand in the set of ligands.

[0043] In another embodiment, the similarities are further based on pairwise comparisons of substructures between each ligand in the set of ligands and every other ligand in the set of ligands and pairwise comparisons between the candidate ligand and every ligand in the set of ligands.

[0044] In a still further embodiment, the similarities are further based on pairwise comparisons of a protein-ligand interaction between each ligand in the set of ligands and every other ligand in the set of ligands and pairwise comparisons between the candidate ligand and every ligand in the set of ligands.

[0045] In still another embodiment, the protein-ligand interaction is selected from a hydrogen bond, a salt bridge, and a hydrophobic interaction.

[0046] In a yet further embodiment, determining a pose for the candidate ligand includes determining a pose for each ligand in the set of ligands that minimizes the potential of all ligands.

[0047] In yet another embodiment, the potential is determined by:

$$E(L) = (n - 1)E^{\text{dock}}(L) + E^{\text{pairwise}}(L),$$

where $L = (\ell_1, \dots, \ell_n)$ indexes the candidate poses for each ligand, n is the number of ligands, E^{dock} represents the standard docking scores for the set of ligand poses, and E^{pairwise} represents a pairwise contribution to the total score.

[0048] In a further embodiment again, $E^{\text{pairwise}}(L) = \sum_{(i,j), i \neq j} e^{\text{pairwise}}(\ell_i, \ell_j)$.

[0049] In another embodiment again, $e^{pairwise}$ is determined using a machine learning algorithm.

[0050] In a further additional embodiment, the machine learning algorithm is fit to a benchmark set of pairs of correct poses and pairs of incorrect poses for a diverse set of proteins.

[0051] In another additional embodiment, determining a pose for each ligand in the set of ligands includes utilizing an expectation-maximization algorithm.

[0052] In a still yet further embodiment, the method further includes altering a pose of a ligand in the set of ligands to optimize that ligand's contribution to the potential function.

[0053] In still yet another embodiment, a method for predicting ligand-biomolecule interaction includes obtaining a set of ligands known to bind to a target biomolecule, generating a set of candidate poses for each ligand in the set of ligands and a candidate ligand, and determining a pose for the candidate ligand based on a similarity of the candidate poses for the candidate ligand and the candidate poses for each ligand in the set of ligands.

[0054] In a still further embodiment again, generating a set of candidate poses is accomplished by docking at least one ligand from the set of ligands using docking software.

[0055] In still another embodiment again, the method further includes generating a set of scores for each ligand using docking software.

[0056] In a still further additional embodiment, the docking software is selected from Glide, DOCK, AutoDock VINA, FlexX, GOLD, and Surflex.

[0057] In still another additional embodiment, at least one ligand in the set of ligands has an experimentally determined pose.

[0058] In a yet further embodiment again, the pose for the at least one ligand is set to the experimentally determined pose.

[0059] In yet another embodiment again, the similarity is based on pairwise comparisons between each ligand in the set of ligands and every other ligand in the set of ligands and pairwise comparisons between the candidate ligand and every ligand in the set of ligands.

[0060] In a yet further additional embodiment, the similarity is further based on pairwise comparisons of substructures between each ligand in the set of ligands and every other ligand in the set of ligands and pairwise comparisons between the candidate ligand and every ligand in the set of ligands.

[0061] In yet another additional embodiment, the similarity is further based on pairwise comparisons of a biomolecule-ligand interaction between each ligand in the set of ligands and every other ligand in the set of ligands and pairwise comparisons between the candidate ligand and every ligand in the set of ligands.

[0062] In a further additional embodiment again, the biomolecule-ligand interaction is selected from a hydrogen bond, a salt bridge, and a hydrophobic interaction.

[0063] In another additional embodiment again, determining a pose for the candidate ligand includes determining a pose for each ligand in the set of ligands that minimizes the potential of all ligands.

[0064] In a still yet further embodiment again, the potential is determined by:

$$E(L) = (n - 1)E^{\text{dock}}(L) + E^{\text{pairwise}}(L),$$

where $L = (\ell_1, \dots, \ell_n)$ indexes the candidate poses for each ligand, n is the number of ligands, E^{dock} represents the standard docking scores for the set of ligand poses, and E^{pairwise} represents a pairwise contribution to the total score.

[0065] In still yet another embodiment again, $E^{\text{pairwise}}(L) = \sum_{(i,j), i \neq j} e^{\text{pairwise}}(\ell_i, \ell_j)$.

[0066] In a still yet further additional embodiment, e^{pairwise} is determined using a machine learning algorithm.

[0067] In still yet another additional embodiment, the machine learning algorithm is fit to a benchmark set of pairs of correct poses and pairs of incorrect poses for a diverse set of proteins.

[0068] In a yet further additional embodiment again, determining a pose for each ligand in the set of ligands includes utilizing an expectation-maximization algorithm.

[0069] In yet another additional embodiment again, the method further includes altering a pose of a ligand in the set of ligands or the candidate ligand to optimize that ligand's contribution to the potential function.

[0070] In a still yet further additional embodiment again, the biomolecule is selected from a protein, a carbohydrate, a DNA molecule, an RNA molecule, and combinations thereof.

[0071] Other features and advantages of the present invention will become apparent from the following detailed description, taken in conjunction with the accompanying drawings which illustrate, by way of example, the principles of the invention.

BRIEF DESCRIPTION OF THE DRAWINGS

[0072] The description and claims will be more fully understood with reference to the following figures and data graphs, which are presented as exemplary embodiments of the invention and should not be construed as a complete recitation of the scope of the invention.

[0073] Figures 1A-1B illustrate ligand docking in accordance with various embodiments of the invention.

[0074] Figure 2 illustrates how ligands adopt similar binding poses accordance with various embodiments of the invention.

[0075] Figure 3 illustrates a flow chart of a method to determine binding pose in accordance with various embodiments of the invention.

[0076] Figure 4A-4B illustrate exemplary data of interaction and substructure similarity in accordance with various embodiments of the invention.

[0077] Figures 5A-5C illustrate exemplary data showing how certain embodiments identify key interactions in accordance with various embodiments of the invention.

[0078] Figures 6A-6B illustrate exemplary data demonstrating performance of certain embodiments against a diverse benchmarking set in accordance with various embodiments of the invention.

[0079] Figure 7 illustrates a flow chart of a method to determine binding affinity in accordance with various embodiments of the invention.

[0080] Figures 8A-8B illustrate exemplary data of virtual screening performance in accordance with various embodiments of the invention.

[0081] Figures 9A-9D illustrate exemplary data demonstrating prediction and validation of binding poses in accordance with various embodiments of the invention.

DETAILED DESCRIPTION

[0082] Turning now to the drawings, systems and methods to predict protein-ligand complexes and uses thereof are provided. Certain embodiments provide methods that substantially improves the accuracy of binding pose prediction by exploiting a widely available source of non-structural information: a list of other ligands that bind the same target. Many embodiments quantify and leverage a chemist's intuition that even very different ligands tend to form similar interactions with a target protein. Further embodiments describe virtual screening methods to identify molecules that bind to targets. Some embodiments further determine a binding affinity for such molecules.

[0083] Binding of small-molecule ligands to proteins is one of the most fundamental processes in biology, and the great majority of drugs are ligands that exert their effects by binding to a target protein. Determining a ligand's binding pose—the three-dimensional coordinates of the ligand's atoms when bound to the target—is important for structure-based optimization of the ligand's pharmacological properties, as well as to understanding how it influences its target. Knowledge of a ligand's binding pose is so advantageous that researchers in industry and academia often spend months or years to solve an experimental structure of a particular ligand in complex with a target protein. Knowing binding pose information provides many uses, including rationally designing ligands with desired signaling profiles, increasing potency, sub-type selectivity, and bridging multiple binding sites (e.g., bivalent ligands), among a broad range of other applications. While most commonly used intuitively by chemists, further uses can include structure-based virtual screening, free energy calculations, and/or molecular dynamics simulations.

[0084] Additionally, as medicinal chemistry progresses, many current and future therapeutics utilize other macromolecules, including DNA, RNA, peptides, proteins, and carbohydrates as therapeutics or as target molecules of ligands. While much of the description discloses the use on ligand-protein interactions, many embodiments are directed to ligand-macromolecule interactions more generally. Macromolecules can include biomolecules, including (but not limited to) DNA, RNA, carbohydrate, and proteins. Even more embodiments are directed to using biomolecules as ligands themselves, such that DNA, RNA, peptides, proteins, or carbohydrates interact with proteins or other biomolecules. Additionally, several embodiments utilize ligands

comprising small molecules, including organic and/or inorganic molecules. One of skill in the art would understand the following disclosure as enabling for such modifications to allow for such extensions.

[0085] Because experimental structure determination is time-consuming, expensive, and sometimes impossible, tremendous effort has been invested in the development of *in silico* “docking” methods for predicting ligand binding poses. (See e.g., Allen et al., 2015 DOCK 6: Impact of new features and current docking performance. J Comput Chem 36, 1132-1156; Friesner et al., 2004 Glide: a new approach for rapid, accurate docking and scoring. 1. Method and assessment of docking accuracy. Journal of medicinal chemistry 47, 1739-1749; Friesner et al., 2006 Extra precision glide: docking and scoring incorporating a model of hydrophobic enclosure for protein-ligand complexes. Journal of medicinal chemistry 49, 6177-6196; Jain, 2003 Surflex: fully automatic flexible molecular docking using a molecular similarity-based search engine. Journal of medicinal chemistry 46, 499-511; Jones et al., 1997 Development and validation of a genetic algorithm for flexible docking. Journal of molecular biology 267, 727-748; Kuntz et al., 1982 A geometric approach to macromolecule-ligand interactions. J Mol Biol 161, 269-288; Platzer et al., 1972 CONFORMATIONAL ENERGY CALCULATIONS OF ENZYME-SUBSTRATE INTERACTIONS. II. Computation of the Binding Energy for Substrates in the Active Site of α -Chymotrypsin. International Journal of Peptide and Protein Research 4, 201-219; Rarey et al., 1996 A fast flexible docking method using an incremental construction algorithm. Journal of molecular biology 261, 470-489; Trott and Olson, 2010 AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. J Comput Chem 31, 455-461; and Venkatachalam et al., 2003 LigandFit: a novel method for the shape-directed rapid docking of ligands to protein active sites. J Mol Graph Model 21, 289-307; the disclosures of which are hereby incorporated by reference herein in their entireties.) Such methods are considered physics-based or “per-ligand” approaches—i.e., given a structure of the target protein, these docking methods sample many candidate poses of a query ligand and rank these poses using scoring functions that approximate the energetic favorability of each pose. Such scoring functions are typically designed to capture interatomic interactions such as hydrogen bonds and van der Waals forces (Figure 1A).

[0086] Despite the development of dozens of docking software packages over the past 40 years, their performance at pose prediction leaves much room for improvement. Binding pose predictions are typically correct less than half the time for ligands substantially different from those in the experimental structures used for docking.

[0087] Many embodiments are directed to methods and systems that substantially improve binding pose prediction by using a widely available type of non-structural data, such as the identities of other ligands known to bind the same target (Figure 1B). Such data is routinely collected in drug development campaigns, both through initial screens to identify lead compounds and through assays of analogs synthesized during lead optimization, and is available in public databases such as ChEMBL for the vast majority of recognized drug targets. (See e.g., Gaulton et al., 2017 The ChEMBL database in 2017. *Nucleic Acids Res* 45, D945-D954; the disclosure of which is hereby incorporated by reference herein in its entirety.)

[0088] Medicinal chemists have long recognized that distinct ligands tend to bind a given protein in similar poses; even ligands sharing no common substructure often form similar interactions with the target protein (Figure 2). This intuition has a sound basis in physics. For example, the energetic favorability of a protein–ligand hydrogen bond depends on the mobility of the protein atoms involved and their ability to form hydrogen bonds with water in the absence of the ligand. (See e.g., Bissantz et al., 2010 A medicinal chemist's guide to molecular interactions. *Journal of medicinal chemistry* 53, 5061-5084; the disclosure of which is hereby incorporated by reference herein in its entirety.) These factors contribute similarly to binding of different ligands but are difficult to predict from a static protein structure alone.

Methods to Identify Ligand-Protein Interaction

[0089] Turning to Figure 3, various embodiments are directed to methods to identify ligand-protein interaction. The interaction between a ligand and a protein can be identified by its pose, or how a molecule is arranged within the target protein. Specifically, Figure 3 illustrates an exemplary method 300 to predict binding pose of a ligand. At 302, many embodiments identify a target protein and query ligand. Various embodiments select a target protein and query ligand that are known to interact. In numerous embodiments, the

target protein is a drug target. During 302, many embodiments obtain structures for the target protein and query ligand with molecular coordinates, such as through known means or databases, including ChEMBL, PDB, etc.

[0090] At 304, additional embodiments obtain a set of structures of ligands known to bind to or interact with the target protein ("helper ligands"). A set, in accordance with various embodiments, can contain any number of ligands, such as at least 1 ligand, at least 2 ligands, at least 5 ligands, at least 10 ligands, at least 20 ligands, at least 50 ligands, at least 100 ligands, or more. In some embodiments, a binding pose for a ligand in the set of ligands is known, such as a binding pose discovered via experimental means.

[0091] In some embodiments, the known-binding ligands are obtained from a public database, such as ChEMBL. Various embodiments curate the set of known-binding ligands based on factors such as binding affinity, including high affinity binding and/or sharing a large substructure with the query ligand (congeneric).

[0092] At 306, various embodiments predict initial binding poses of the query ligand and set of known-binding ligands to the target protein. Certain embodiments use per-ligand docking software to generate a list of candidate poses for each ligand. Exemplary docking software used in various embodiments includes (but is not limited to) one or more of Glide, DOCK, AutoDock VINA, FlexX, GOLD, and/or Surflex. However, many embodiments can utilize any per-ligand scoring function and pose sampling strategy, including those implemented in any other per-ligand docking package.

[0093] Various embodiments utilize the docking software to generate a set of poses for each ligand, including the query ligand and the set of ligands known to bind the target protein. In some embodiments, the set of poses are the highest ranking poses as determined by a scoring function within the docking software. In certain embodiments, the set of ranked poses includes at least 10 poses, at least 25 poses, at least 50 poses, at least 100 poses, or more poses for each ligand.

[0094] To dock ligands in some embodiments, a single target protein structure is selected from PDB. Additional embodiments prepare the target protein structure by removing waters, optimizing hydrogen bonds, and performing energy minimization with non-hydrogen atoms constrained to an RMSD of less than 0.3 Å from the initial structure.

[0095] Because various embodiments can use any per-ligand docking method for pose generation and scoring of individual ligands, it will be able to take advantage of improvements to these methods. For example, various embodiments use machine learning to fit scoring functions, and others allow for binding pocket flexibility when generating candidate poses.

[0096] At 308, many embodiments quantify similarities of binding poses for the query ligand and known-interacting ligand or ligands. Many embodiments solve for all their binding poses simultaneously. Many embodiments quantify similarities based on three-dimensional space, based on coordinates of each atom within the molecules (e.g., ligands and protein). To maximize a scoring function in various embodiments, the per-ligand docking software generates candidate poses and assigns a per-ligand score to each pose. Many embodiments then select a set of poses—one for each ligand—that optimizes the scoring function. In many embodiments, the scoring function considers both the favorability of each ligand's pose, as assessed by a per-ligand scoring function, and the likelihood that a set of poses sharing a given level of similarity is correct or incorrect.

[0097] Many embodiments perform pairwise comparisons between every ligand by computing similarities between every pose in a set of poses for one ligand and every pose in a set of poses for a second ligand. Numerous embodiments compute such similarities for every pair of ligands.

[0098] When an experimentally determined pose is available for one or more ligands, some embodiments “fix” the pose for a ligand having an experimentally determined pose to the “correct” pose, rather than maintaining a set of candidate poses, where a pose is considered correct if it is within a threshold distance of an experimentally determined pose—for example, within 1 Å, 2 Å, 5 Å, 10 Å, root mean squared deviation (RMSD) to the experimentally determined pose.

[0099] Further embodiments evaluate pose similarity for different types of protein-ligand interactions, including hydrogen bonds, salt bridges, and hydrophobic contacts. Quantitative measures used in various embodiments are designed to assess the presence of these interactions between the ligand and a given protein residue. In many embodiments, the hydrogen bond and salt bridge interaction measures are designed to give a value of 1 for interactions meeting established criteria. A soft boundary is added to

give borderline cases values between 0 and 1, in order to prevent discontinuities in some embodiments. The hydrophobic contact measure approximates the hydrophobic surface contact area by considering the number of protein-ligand atom pairs in contact with each other. Again, a soft boundary (in this case, between an atom pair being or not being in contact) can be used to prevent very similar poses from leading to very different values. Various embodiments denote the interaction value for interaction type k , for pose ℓ_i of ligand i , with protein residue r as $X_r^{(k)}(\ell_i)$.

[00100] Interaction similarities for a pair of poses (for two different ligands bound to the same target protein) are computed separately for each interaction type in various embodiments. The interactions made between the ligand and each residue of the target protein residue can be tabulated and then the similarity between the resulting lists for each pose can be measured by the Tanimoto coefficient. The Tanimoto coefficient can be modified by the addition of pseudo counts, which serve to make the metric well defined if neither ligand forms a particular type of interaction and to reward poses that share larger numbers of interactions in absolute terms. Certain embodiments define the interaction similarity, for interaction type k between a pair of poses ℓ_i, ℓ_j (for ligands i and j , respectively), as

$$s^{(k)}(\ell_i, \ell_j) = \frac{1 + \sum_{r \in R} \sqrt{X_r^{(k)}(\ell_i) X_r^{(k)}(\ell_j)}}{2 + \sum_{r \in R} [X_r^{(k)}(\ell_i) + X_r^{(k)}(\ell_j)] - \sum_{r \in R} \sqrt{X_r^{(k)}(\ell_i) X_r^{(k)}(\ell_j)}}$$

where R is the set of all protein residues.

[00101] In various embodiments, when computing hydrogen bond similarity, a case where a given protein residue acts as a hydrogen bond donor for one ligand and a hydrogen bond acceptor for another ligand is not considered a shared interaction.

[00102] Figure 4A illustrates exemplary data illustrating interaction similarity in correct and highly ranked poses, where pairs of correct poses form more similar interactions than other pairs of poses ranked highly by the per-ligand scoring function.

[00103] Additional embodiments define a substructure similarity as the RMSD of atom positions of the largest chemical substructure shared by a pair of ligands. Certain embodiments evaluate substructure similarity for pairs of ligands that shared a substructure at least half the size of the smaller ligand. Figure 4B illustrates exemplary

data showing that for this similarity metric, too, the native distribution exhibits higher similarity than the reference distribution, indicating that the common substructure tends to be more similarly positioned in pairs of correct poses than in other pairs of poses ranked highly by the per-ligand scoring function.

[00104] Many embodiments compute the substructure similarity for a pair of candidate poses. To compute the substructure similarity, the maximum common substructure of the two ligands can be identified using available software (e.g., Canvas from Schrodinger LLC) and then mapped onto each candidate pose. Finally, the RMSD between these two sets of atoms can be computed and used as the measure of substructure similarity. Certain embodiments define custom atom and bond types for computation of the common scaffold. Scaffold similarity is not considered for pairs of ligands with a maximum common substructure of less than half the size of the smaller ligand in certain embodiments. Various embodiments do not include hydrogen atoms in the substructure nor when determining the total number of atoms in each ligand.

[00105] The scoring function of various embodiments is derived by quantifying the medicinal chemist's intuition that different ligands tend to adopt structurally similar poses when binding the same target protein. By quantifying such assumptions, these embodiments seek not only to measure the similarity of correct poses of different ligands, but also to compare the similarity of correct poses to that of other poses ranked highly by per-ligand docking software. Additionally, two notions of similarity are considered: similarity of protein–ligand interactions and similarity in position of common ligand substructures.

[00106] Using a benchmark set of protein–ligand complex structures, many embodiments characterize the extent to which distinct ligands binding a common target adopt similar poses, as quantified by the interaction and substructure similarity metrics described above.

[00107] Some embodiments use a curated set of protein-ligand complex structures representing major drug targets. (See e.g., Santos et al., 2017 A comprehensive map of molecular drug targets. *Nature reviews Drug discovery* 16, 19-34; the disclosure of which is hereby incorporated by reference herein in its entirety.) Some of these embodiments use one or more of manual curation and adaption of a PDBbind refined set. (See e.g.,

Wang et al., 2005 The PDBbind database: methodologies and updates. Journal of medicinal chemistry 48, 4111-4119; the disclosure of which is hereby incorporated by reference herein in its entirety.)

[00108] Some embodiments generate a probability distribution based on the “correct pose” similarities, and various embodiments store the probability distribution as a native distribution. Additional embodiments determine quality of docked poses. In certain embodiments, the accuracy of each pose is quantified by the non-hydrogen-atom RMSD from the experimentally determined pose. Various embodiments compute the RMSD by aligning each complex to the structure used for docking based on non-hydrogen-atoms within 15 Å of the ligand, and an RMSD is then computed between the docked pose and the same ligand’s pose in the aligned complex. Some embodiments consider poses with within 2.0 Å RMSD to be “near-native” or “correct”.

[00109] When computing these statistics, various embodiments identify docked poses that are near-native among the candidate poses ranked in the top 100 by the docking software. Certain embodiments use these docked poses, as opposed to the experimentally determined pose, in order to ensure that the statistics will be applicable to the scoring of candidate poses generated by docking software. Additional embodiments compute the empirical distribution of each similarity type (e.g., substructure, interaction, etc.) across all pairs of near-native poses using a Gaussian kernel density estimate with standard deviation of 0.03 for interaction similarities and 0.18 for substructure similarities. To reduce bias near the boundaries, reflected boundary conditions are applied in some embodiments. Some embodiments cap substructure similarities at 6 Å (that is, substructure similarities greater than 6 Å were set to 6 Å), as the sparsity of near-native pose pairs for higher values led to overly rough distributions. However, additional embodiments can cap substructure similarities to another threshold, as optimized for a particular use. Various embodiments denote the similarity distribution over near-native poses for interaction type k as $f_k(x; \text{Native})$.

[00110] Many embodiments compute equivalent similarity distributions using all pairs of candidate poses produced by the docking software, regardless of whether they are near-native. Certain embodiments denote the resulting distributions as $f_k(x; \text{Reference})$.

[00111] To combine the distributions for the four similarity types into a single joint distribution, many embodiments assume that the interaction types are conditionally independent and express the joint distribution as a product of the distributions for each interaction type. That is:

$$f(s(\ell_i, \ell_j); \text{Native}) = \prod_k f_k(s^{(k)}(\ell_i, \ell_j); \text{Native})$$

$$f(s(\ell_i, \ell_j); \text{Reference}) = \prod_k f_k(s^{(k)}(\ell_i, \ell_j); \text{Reference})$$

where $s(\ell_i, \ell_j)$ is the vector of $s^{(k)}(\ell_i, \ell_j)$'s for each similarity type k .

[00112] Further embodiments can use any additional pairwise pose similarity metric. Various embodiments could potentially improve performance by using more fine-grained interaction descriptors or by using similarity metrics based on field-based methods developed for virtual screening.

[00113] Certain embodiments derive a statistical potential for sets of binding poses. Many embodiments use the similarity distributions (e.g., reference and native) to derive a statistical potential that—instead of acting on features of a single pose, as in previous docking software—acts on a set of hypothesized poses, one for each ligand, known to bind the target protein. In a number of embodiments, potential is determined by the equation:

$$E(L) = (n - 1)E^{\text{dock}}(L) + E^{\text{pairwise}}(L),$$

where $L = (\ell_1, \dots, \ell_n)$ indexes the candidate poses for each ligand, n is the number of ligands, E^{dock} represents the standard docking scores for the set of ligand poses, and E^{pairwise} represents the pairwise contribution to the total score. In some embodiments,

$$E^{\text{pairwise}}(L) = \sum_{(i,j), i \neq j} e^{\text{pairwise}}(\ell_i, \ell_j).$$

Various embodiments determine e^{pairwise} is determined utilizing one or more machine learning models. Some machine learning models fit to a benchmark set of pairs of correct poses and pairs of incorrect poses for a diverse set of proteins.

[00114] Additional embodiments utilize the following equation to determine potential:

$$E^{\text{ComBind}}(\text{Poses for a set of ligands}) = (n - 1) \sum_{\text{ligands}} E^{\text{dock}}(\text{pose}) + \sum_{\text{ligand pairs}} \sum_{\text{similarity types}} -\log \frac{\text{frequency of pose pair similarity in native distribution}}{\text{frequency of pose pair similarity in reference distribution}},$$

where n is the total number of ligands.

[00115] In the above equation, the first component evaluates the energetic favorability of each ligand's pose individually using a per-ligand scoring function E^{dock} (e.g., a scoring function used in Glide or another docking software package). The summation is over ligands known to bind the target protein, with "pose" referring to the hypothesized pose for each ligand.

[00116] The second component rewards sets of poses with a degree of similarity that is more often observed in correct poses than in other poses ranked highly by the per-ligand scoring function. Here, the outer summation is over pairs of distinct ligands known to bind the target protein, and the inner summation is over the interaction and substructure similarity measures (see Figures 4A-4B): hydrogen bond similarity, salt bridge similarity, hydrophobic contact similarity, and substructure similarity. "Pose pair similarity" refers to the calculated similarity value of the given type for the hypothesized poses of the given ligand pair. The "native distribution" and "reference distribution" for each similarity type are determined as described above. The resulting negative log likelihood ratios have the mathematical properties of an energy, namely that an additive decrease in energy corresponds to a multiplicative increase in likelihood ratio, allowing for straightforward integration with standard per-ligand docking scores, which are typically in units of energy. For pairs of ligands that do not share a substructure at least half the size of the smaller ligand, the substructure similarity term can be excluded in the summation.

[00117] The second component acts as a correction to the first. If the per-ligand scoring function were perfect, in the sense that it could perfectly distinguish correct poses from incorrect ones, the terms in the second component would consistently assume values of

zero. Because per-ligand scoring functions remain imperfect—and, in particular, tend to underpredict the likelihood that a set of ligands will adopt similar poses—the second component typically assumes non-zero values.

[00118] The statistical potential used by various embodiments is sufficiently general that the method could be extended to exploit other types of data, ranging from multiple experimental structures of the protein in complex with different ligands to effects of protein mutation on ligand binding. Likewise, various embodiments exploit the affinity of each known binder.

[00119] It should be noted that various embodiments may perform various steps simultaneously, multiple times, and/or omit steps as appropriate for a particular use. For example. Some embodiments may obtain multiple query ligands and/or multiple sets of known-binding ligands for use within an embodiment of method 300. Additionally, certain embodiments iteratively change a pose for a ligand to optimize that ligand's contribution to the potential function, such that multiple features of method 300 may be repeated.

[00120] Turning to Figures 5A-5C, exemplary data showing the improvement of many embodiments is illustrated. Specifically, Figure 5A illustrates a comparison between various ligands known to bind to the β_1 -adrenergic receptor (β_1 AR). Typical docking software (e.g., Glide) evaluate each ligand independently. In contrast, many embodiments evaluate multiple poses for multiple ligands simultaneously to provide poses (circled) that minimize potential. When the poses identified by the docking software and exemplary embodiment were compared to experimentally determined poses, Figure 5B illustrates that Glide (X-axis) correctly identified only 4 out of 11 poses (left of 2 Å dashed line), while the exemplary embodiment (Y-axis) correctly identified 10 of the 11 poses (below 2 Å dashed line). Additionally, Figure 5C illustrates a comparison of experimentally determined protein-ligand interactions between Glide (Per-ligand Docking) and the exemplary embodiment, showing that the exemplary embodiment is more similar to the experimentally determined interactions.

[00121] Similarly, Figures 6A-6B illustrate exemplary data of the improvements provided by embodiments over per-ligand docking software (e.g., Glide). Specifically, when benchmarked against a set of 245 protein-ligand complexes representing major families of drug targets. Figure 6A illustrates an improvement provided by the

embodiment (dots) over Glide (open circles) for the various protein families. On average, these exemplary embodiments select a correct binding pose 30% more frequently. For G-protein-coupled receptors (GPCRs), which are of particular interest both because they represent the largest family of drug targets and because their structures are notoriously difficult to determine experimentally, various embodiments select a correct binding pose over 60% more frequently, increasing the probability of correct prediction from 47% to 76%. Figure 6B illustrates an improvement provided by the known-binding, or helper, ligands. As illustrated in Figure 6B, increasing the size of the set of known-binding (helper) ligands improves performance of some embodiments.

Virtual Screening

[00122] Further embodiments are directed to methods to identify ligands that bind to a target protein from a query set, utilizing prior knowledge of one or more molecules that bind the target protein. Some of these embodiments identify which molecules from a query set that bind to the target protein, then identify a binding affinity for any molecule that binds to the target protein. Some embodiments further provide rankings or other information for any molecules that are determined to bind to a target protein. Such embodiments can assist in many uses, such as pharmaceutical lead optimization and virtual screening of molecules.

[00123] Some embodiments are used in drug discovery to identify drug candidates that may interact with target proteins. However additional embodiments are used in cellular signaling (e.g., binding of hormones and neurotransmitters), sensation (e.g., binding of odorants and flavorants), enzyme function (e.g., binding of nutrients and other metabolic substrates), and defense mechanisms (e.g., binding of toxins and antibiotics). Pose prediction is used to understand the effects of genetic variation on responses to both naturally occurring ligands and drugs, which can benefit personalized medicine. In each of these cases, multiple ligands are typically known to bind the targets of interest, and certain embodiments may thus be used to improve binding pose prediction.

[00124] Figure 7 illustrates an exemplary method 700 for determining a binding affinity for a molecule in accordance with various embodiments. Many embodiments obtain a structure of a target protein at 702. In numerous embodiments, the target protein is a drug

target. Some embodiments obtain a structure for the target protein with molecular coordinates, such as through known means or databases, including ChEMBL, PDB, etc.

[00125] Many embodiments obtain a set of structures of ligands known to bind to or interact with the target protein (“helper ligands”) at 70. A set, in accordance with various embodiments, can contain any number of ligands, such as at least 1 ligand, at least 2 ligands, at least 5 ligands, at least 10 ligands, at least 20 ligands, at least 50 ligands, at least 100 ligands, or more. In some embodiments, a binding pose for a ligand in the set of ligands is known, such as a binding pose discovered via experimental means.

[00126] In some embodiments, the known-binding ligands are obtained from a public database, such as ChEMBL. Various embodiments curate the set of known-binding ligands based on factors such as binding affinity, including high affinity binding and/or sharing a large substructure with the query ligand (congeneric).

[00127] At 706, various embodiments identify a pose for each ligand in the set of helper ligands. Such methods are similar to the docking and pose prediction as described in method 300 (see Figure 3 and associated text), although, method 700 may not include simultaneous docking of a query ligand or molecule. Many embodiments “fix” or hold a single pose for each helper ligand as a permanent binding pose.

[00128] At 708, many embodiments obtain a set of query molecules. The set of query molecules can include any number of molecules, including 1 molecule, 2 molecules, 3 molecules, 4 molecules, 5 molecules, 10 molecules, 15 molecules, 20 molecules, 25 molecules, 50 molecules, 75 molecules, 100 molecules, or more. Many embodiments obtain structures for the query molecules including coordinates for each atom in the molecule.

[00129] At 710, many embodiments identify if each query molecule binds to the target protein and generates a binding affinity for each binding molecule. Numerous embodiments utilize a similar procedure as described in method 300 to determine pose. Further embodiments select a pose that minimizes the potential with respect to the helper ligands. In doing so, the minimized potential demonstrates an increased binding affinity between the query molecule and the target protein.

[00130] Further embodiments perform lead optimization of one or more query molecules at 712-716. In various embodiments, a modifiable location is identified on the query ligand at 712. The modifiable position can be any position that may allow for additional modification that allows for a change in chemical group, including groups that may sit internal to a binding site that could increase binding affinity, while some embodiments may identify a location that may not contribute to binding, such that a modification could be used for increasing solubility, labeling, or conjugating additional molecules to the query molecule.

[00131] At 714, some embodiments alter the modifiable position. For example, some embodiments may alter the position to increase binding affinity via the inclusion of a chemical group that may form an interaction with the target protein, such as via a hydrogen bond, salt bridge, and/or hydrophobic interaction.

[00132] Additional embodiments determine a new binding affinity for the modified query molecule at 716. Such binding affinity is assessed similarly to 710, where the pose prediction and potential demonstrate a binding affinity for the modified query molecule.

[00133] It should be noted that various embodiments may perform various steps simultaneously, multiple times, and/or omit steps as appropriate for a particular use. For example. Some embodiments may obtain multiple query ligands and/or multiple sets of known-binding ligands for use within an embodiment of method 700.

[00134] Various embodiments perform better at identifying ligands from query sets than current methods. Figures 8A-8B illustrate exemplary performance metrics of various embodiments against a per-ligand docking software. Specifically, Figure 8A illustrates a performance increase with increasing numbers of helper ligands. Helper ligands include In Figure 8A, the query molecules possessed a Tanimoto Coefficient of < 0.20 relative to the helper ligands, while in Figure 8B, the query molecules possessed a Tanimoto Coefficient of > 0.20 and < 0.30 relative to the helper ligands. In Figures 8A-8B, "chem. similarity" is based on the presence of similar chemical groups between the query molecule and helper ligands.

EXEMPLARY EMBODIMENTS

[00135] Although the following embodiments provide details on certain embodiments of the inventions, it should be understood that these are only exemplary in nature, and are not intended to limit the scope of the invention.

Example 1: Structure prediction informed by non-structural data

[00136] As an illustrative example, one embodiment was used to predict binding poses for ligands at the β_1 -adrenergic receptor (β_1 AR), the primary target of the beta blocker drugs that are widely used to treat heart attack, heart failure, and hypertension. Eleven diverse ligands known to bind β_1 AR were selected, including both beta blockers and beta agonists. The 11 distinct ligands were docked to a crystallographic β_1 AR structure using Glide, producing up to 100 candidate poses for each ligand. This embodiment solved for a set of poses—one per ligand—that minimizes the potential (Figure 5A). The crystallographic β_1 AR structure used for docking was determined in complex with a ligand distinct from any of the 11 docked ligands. Crystallographic ligand poses were not used in any way by Glide or this embodiment.

[00137] Glide's top-ranked pose was correct for 4 of 11 ligands, whereas the pose selected by this embodiment was correct for 10 of 11 ligands (Figure 5B). In the selected poses for this embodiment—as in experimentally determined poses—most of the ligands form a salt bridge with D121 and hydrogen bonds with S211 and N329 (Figure 5C). In comparison, the poses ranked most highly by Glide's per-ligand scoring function exhibited more varied hydrogen bonds and salt bridges (Figure 5C).

[00138] It is emphasized that this embodiment did not require that all ligands adopt similar poses or form similar interactions. This embodiment correctly predicted, for example, that two of these β_1 AR ligands do not form a hydrogen bond with S211.

[00139] In another exemplary embodiment, 245 protein–ligand complexes representing all major families of drug targets were used to benchmark this embodiment. Additionally, several steps to mimic a real-world use case for this embodiment. First, when predicting the binding pose of a query ligand with this embodiment, helper ligands selected from the public ChEMBL database were used. No experimental information on poses of helper ligands were used; indeed, for the vast majority of helper ligands selected, poses have

not been determined experimentally. Second, target protein structure determined in the presence of a ligand that shares a chemical scaffold with the query ligand were not used either. Finally, when predicting ligand binding poses at a given target protein, all structures involving that protein when constructing the distributions used to define the potential were omitted.

[00140] When comparing this embodiment to Glide, two helper ligand selection criteria were used: (1) a diverse set of ligands with the highest binding affinity (“high affinity”), and (2) the ligands sharing the largest substructure with the query ligand (“congeneric”). Both of these selection criteria lead to substantial performance improvements over Glide, indicating that this embodiment could be applied effectively using either a diverse set of ligands identified from a high-throughput screen or a congeneric series of ligands generated during lead optimization.

[00141] Additionally, performance improves with the use of more helper ligands (Figure 6B). Interestingly, this embodiment substantially outperforms Glide even when using only a single helper ligand.

Example 2: Benchmarking against state-of-the-art software

[00142] *Methods:* The performance of this embodiments was evaluated on 30 target proteins. Only ligands that have less than 50% scaffold overlap with the ligand that was originally present in the experimental structure used for docking were considered. It was found that ligands with higher scaffold overlap were substantially easier to dock, likely due to the binding pocket being well shaped to accommodate the similar ligand. Additionally, only consider for which there is at least one correct candidate pose were considered, since only in these cases is it possible for either this embodiment or Glide to make a correct prediction.

[00143] For each of the 245 unique ligands meeting these criteria, other ligands known to bind the respective target protein from the ChEMBL database were identified and then used this embodiment to jointly predict their binding poses. Importantly, when evaluating the performance of this method on a particular target protein, the data for that target protein from the similarity statistics were excluded.

[00144] *Selecting Helper Ligands:* For all targets, K_i or IC_{50} data (whichever was more numerous) from ChEMBL were downloaded. Ligands were removed that did not meet the following criteria: a ChEMBL confidence score of 9 (the highest value), molecular weight < 800 Da, and K_i or IC_{50} < 1 μ M. Ligand structures were generated from the SMILES strings provided by ChEMBL.

[00145] Two criteria were benchmarked for selecting which ChEMBL ligands to use as helper ligands for each query ligand: (1) the highest affinity binders that do not share a chemical scaffold, and (2) the ligands that share the largest chemical substructure with the query ligand. To define the size of the common substructure, the same maximum common substructure definition as that used to compute substructure similarity were used. For selection method (1), helper ligands were added in order of affinity, not adding a ligand if it has greater than 80% substructure overlap with any ligand already in the selected set of helpers.

[00146] The benchmarking results presented in the figures were obtained using the following ligand selection criteria and number of helper ligands: Figure 6A: 20 helper ligands selected using criterion (1); Figure 6B: the indicated number of ligands selected using criterion (1). For a handful of targets, fewer than 20 helper ligands were available meeting our criteria. In these cases, we used the minimum of the indicated number of ligands and the number of available ligands. Targets with only one ligand are omitted from Figure 6A.

[00147] *Performance Evaluation:* An overall performance metric was developed to represent the expected performance in drug development campaigns. For each protein family, the average performance was computed, then weighted each by the fraction of FDA-approved drugs targeting the protein family.

[00148] *Results:* Turning to Figure 6A, benchmark results of an embodiment are illustrated. In particular, 20 helper ligands for each query ligand were used and selected from ChEMBL by the high-affinity criterion. When evaluating the ability of both per-ligand docking (e.g., Glide) and the one embodiment to select a correct pose from a list of candidate poses, query ligands with at least one correct pose in the list were considered. When computing overall results, target families were averaged across, the performance for each family by the fraction of FDA-approved drugs targeting that family were weighted.

[00149] On average, this embodiment selects a correct pose for 72% of ligands, a 30% improvement over Glide. This embodiment proves particularly effective on certain target families. But this embodiment improves pose prediction performance for all target families considered. Even at the individual target level, this embodiment hardly ever degrades performance: this embodiment only reduced performance for one of the 30 targets considered, and this performance reduction was minor.

[00150] Removing any of the similarity types from the potential reduced this embodiment's performance. In particular, both protein–ligand interaction similarity and substructure similarity contribute substantially to this embodiment's accuracy. Protein–ligand interaction similarity is the more important of the two, particularly when using a diverse set of helper ligands.

Example 3: Predicting binding poses of antipsychotics at the D₂ dopamine receptor

[00151] To illustrate the practical application of embodiments, it was predicted that the previously unknown binding poses of several antipsychotic drugs that target the D₂ dopamine receptor (D₂R): pimozide, benperidol and spiperone. Only the available structure of D₂R was used, which was determined in complex with risperidone, a ligand substantially different from any of those considered.

[00152] It was predicted that binding poses for pimozide, benperidol and spiperone, as well as the tool compound mespiperone, using both Glide and an embodiment. For spiperone and mespiperone, An embodiment and Glide predicted similar poses. For pimozide and benperidol, however, an embodiment's predictions were very different from Glide's: a fluorobenzene ring of each compound was positioned near the top of the binding pocket by Glide and near the bottom by this exemplary embodiment (Figures 9A-9B).

[00153] To test the embodiment's predictions, mutagenesis experiments were designed. First, a series of mutations of Ser193 (S193) were tested, which is positioned close to the second fluorobenzene ring of pimozide in the embodiment's predicted pose but not in Glide's (Figure 9C). Indeed, mutating S193 to a larger residue (Val or Leu) decreases pimozide's affinity, while mutating S193 to a smaller residue (Ala) increases pimozide's affinity. Such effects are not observed for benperidol, which is identical to

pimozide except that it lacks the fluorobenzene ring that contacts S193 in pimozide (Figure 9D). Indeed, benperidol's affinity actually increases when S193 is mutated to a larger residue. These results are consistent with predicted poses of embodiments but not with Glide's: Glide predicts that pimozide and benperidol position nearly identical chemical groups in essentially identical positions near S193.

[00154] Mutations to several residues surrounding the top and bottom of the binding pocket were also considered. Unlike Glide, the embodiment predicts that all four ligands will position a fluorobenzene ring at the bottom of the binding pocket, packing favorably against Trp386 (W386). Indeed, mutating W386 to a smaller residue (Phe) reduced affinity to a similar extent for all of the ligands, with a slightly smaller effect for pimozide, which packs less tightly against W386 according to the embodiment's prediction. At the top of the ligand binding pocket, near Val91 and Trp100, the embodiment predicts that the pimozide and benperidol will place identical functional groups that differ somewhat from those of spiperone and mespiperone. Indeed, mutation of these residues affects pimozide and benperidol slightly differently from spiperone and mespiperone.

[00155] Binding poses were predicted for the typical antipsychotics spiperone, mespiperone, benperidol, and pimozide at the D₂ dopamine receptor (D₂R). The ligands were prepared using the Schrodinger ligprep tool, considering the unprotonated tautomer and both inversions of the protonated tautomer. The same docking protocol was used as described above, except that the top 300 poses were considered by this embodiment, in order to account for the use of the 3 tautomeric states of the ligand.

Example 4: Binding score

[00156] This embodiment includes hypothesized set of binding poses of a set of n ligands as $L = \ell_1, \ell_2, \dots, \ell_n$, where ℓ_i specifies the hypothesized pose for ligand i .

[00157] Per-ligand scoring functions, which consider each ligand independently, would determine an optimal set of poses $\hat{L}$ by choosing the binding pose with minimum docking score for each ligand or, equivalently, by minimizing

$$E^{\text{dock}}(L) = \sum_{i=1}^n E^{\text{dock}}(\ell_i)$$

where $E^{\text{dock}}(\ell_i)$ is the output of a per-ligand scoring function (such as that reported by Glide) for pose ℓ_i of ligand i .

[00158] Pairwise terms that tend to favor sets of similar poses were added:

$$E^{\text{ComBind}}(L) = (n - 1)E^{\text{dock}}(L) + \sum_{(i,j), i \neq j} -\log \frac{f(s(\ell_i, \ell_j); \text{Native})}{f(s(\ell_i, \ell_j); \text{Reference})}.$$

[00159] Intuitively, these pairwise terms reward pose pairs with similarity values more often observed in near-native (correct) pose pairs than in reference pose pairs (*i.e.*, pose pairs chosen at random from among all candidates). The idea of comparing the distribution of features in correct solutions to the distribution in all possible solutions has been used in statistical potentials for biomolecular structure prediction and in the naïve Bayes machine learning model. The docking scores were weighted by the number of ligands minus 1, in order to hold the relative contribution of singleton and pairwise terms constant for different numbers of helper ligands.

[00160] Consistent with their reported units of kcal/mol, it was found that Glide scores have the mathematical properties of an energy; namely, the negative log likelihood ratio of a pose being near-native is linear in its Glide score. By construction, the pairwise terms we introduce in this study also have this property. This congruence implies that these singleton and pairwise terms can be additively combined (as this is the equivalent of multiplying likelihood ratios).

[00161] In general, it could be that the per-ligand docking scores need to be scaled by a constant factor in order to be consistent with the pairwise terms. For example, if the docking scores were on average 10 times the negative log likelihood ratio of a pose being near-native, they would need to be scaled by 1/10. This constant factor can be identified by performing logistic regression with the docking scores as features and whether each pose is near-native as the response. For Glide scores, the appropriate constant is close to 1 ($1/0.9 = 1.1$), and we chose to set it to 1 for simplicity.

Example 5: Optimization

[00162] This embodiment used coordinate descent to compute a set of poses that minimizes the binding score. At first, L is randomly initialized. L is then iteratively

improved by iterating through the ligands, in a random order, and updating the selected ligand's pose to the argument minimum of $E^{\text{ComBind}}(L)$ assuming that the other poses in L are correct. This procedure is repeated until no more updates can be made. Each update can be computed efficiently because it depends only on the partial contribution of the selected ligand's pose to the binding score:

$$\hat{\ell}_q = \underset{\ell_q}{\operatorname{argmin}} \left[(n-1)E^{\text{dock}}(\ell_q) + \sum_{i \neq q} -\log \frac{f(s(\ell_q, \ell_i); \text{Native})}{f(s(\ell_q, \ell_i); \text{Reference})} \right].$$

[00163] In order to account for the non-convex nature of the ComBind score, this algorithm was repeated from 500 initial configurations, explicitly including the initial configuration corresponding to the generic scoring function predictions at least once and return the best scoring configuration. Empirically this procedure converges to the same result over multiple runs.

DOCTRINE OF EQUIVALENTS

[00164] Having described several embodiments, it will be recognized by those skilled in the art that various modifications, alternative constructions, and equivalents may be used without departing from the spirit of the invention. Additionally, a number of well-known processes and elements have not been described in order to avoid unnecessarily obscuring the present invention. Accordingly, the above description should not be taken as limiting the scope of the invention.

[00165] Those skilled in the art will appreciate that the foregoing examples and descriptions of various preferred embodiments of the present invention are merely illustrative of the invention as a whole, and that variations in the components or steps of the present invention may be made within the spirit and scope of the invention. Accordingly, the present invention is not limited to the specific embodiments described herein, but, rather, is defined by the scope of the appended claims.

WHAT IS CLAIMED IS:

1. A method for predicting ligand-protein interaction comprising:
obtaining a set of ligands known to bind to a target protein;
generating a set of candidate poses for each ligand in the set of ligands and a candidate ligand; and
determining a pose for the candidate ligand based on a similarity of the candidate poses for the candidate ligand and the candidate poses for each ligand in the set of ligands.
2. The method of claim 1, wherein generating a set of candidate poses is accomplished by docking at least one ligand from the set of ligands using docking software.
3. The method of claim 1, further comprising generating a set of scores for each ligand is using docking software.
4. The method of claim 3, wherein the docking software is selected from the group consisting of: Glide, DOCK, AutoDock VINA, FlexX, GOLD, and Surflex.
5. The method of claim 1, wherein at least one ligand in the set of ligands has an experimentally determined pose.
6. The method of claim 5, wherein the pose for the at least one ligand is set to the experimentally determined pose.
7. The method of claim 1, wherein the similarity is based on pairwise comparisons between each ligand in the set of ligands and every other ligand in the set of ligands and pairwise comparisons between the candidate ligand and every ligand in the set of ligands.

8. The method of claim 7, wherein the similarity is further based on pairwise comparisons of substructures between each ligand in the set of ligands and every other ligand in the set of ligands and pairwise comparisons between the candidate ligand and every ligand in the set of ligands.

9. The method of claim 7, wherein the similarity is further based on pairwise comparisons of a protein-ligand interaction between each ligand in the set of ligands and every other ligand in the set of ligands and pairwise comparisons between the candidate ligand and every ligand in the set of ligands.

10. The method of claim 9, wherein the protein-ligand interaction is selected from the group consisting of: a hydrogen bond, a salt bridge, and a hydrophobic interaction.

11. The method of claim 1, wherein determining a pose for the candidate ligand comprises determining a pose for each ligand in the set of ligands that minimizes the potential of all ligands.

12. The method of claim 11, wherein the potential is determined by:

$$E(L) = (n - 1)E^{\text{dock}}(L) + E^{\text{pairwise}}(L),$$

where $L = (\ell_1, \dots, \ell_n)$ indexes the candidate poses for each ligand, n is the number of ligands, E^{dock} represents the standard docking scores for the set of ligand poses, and E^{pairwise} represents a pairwise contribution to the total score.

13. The method of claim 12, wherein $E^{\text{pairwise}}(L) = \sum_{(i,j), i \neq j} e^{\text{pairwise}}(\ell_i, \ell_j)$.

14. The method of claim 12, wherein e^{pairwise} is determined using a machine learning algorithm.

15. The method of 14, wherein the machine learning algorithm is fit to a benchmark set of pairs of correct poses and pairs of incorrect poses for a diverse set of proteins.

16. The method of claim 11, wherein determining a pose for each ligand in the set of ligands comprises utilizing an expectation-maximization algorithm.
17. The method of claim 11, further comprising altering a pose of a ligand in the set of ligands or the candidate ligand to optimize that ligand's contribution to the potential function.
18. A method for virtual screening, comprising:
obtaining a set of ligands known to bind to a target protein;
determining a pose for each ligand in the set of ligands by generating a set of candidate poses for ligands in the set of ligands, wherein the pose is based on a similarity of the candidate poses for each ligand in the set of ligands; and
determining a binding affinity of at least one candidate ligand by determining whether the candidate ligands binds to the target protein based on similarities to each ligand in the set of ligands and determining an energy if the candidate ligand binds to the target ligand.
19. The method of claim 18, wherein determining a binding affinity of a candidate ligand determines a binding affinity of at least two candidate ligands.
20. The method of claim 19, further comprising obtaining a library of candidate ligands, and wherein determining a binding affinity of at least one candidate ligand determines a binding affinity of at least one candidate ligand in the library of candidate ligands.
21. The method of claim 18, further comprising fixing a pose for each ligand in the set of ligands prior to determining a binding affinity of at least one candidate ligand.
22. The method of claim 18, wherein the similarity is based on pairwise comparisons between each ligand in the set of ligands and every other ligand in the set of ligands.

23. The method of claim 22, wherein the similarities are further based on pairwise comparisons of substructures between each ligand in the set of ligands and every other ligand in the set of ligands and pairwise comparisons between the candidate ligand and every ligand in the set of ligands.

24. The method of claim 22, wherein the similarities are further based on pairwise comparisons of a protein-ligand interaction between each ligand in the set of ligands and every other ligand in the set of ligands and pairwise comparisons between the candidate ligand and every ligand in the set of ligands.

25. The method of claim 24, wherein the protein-ligand interaction is selected from the group consisting of: a hydrogen bond, a salt bridge, and a hydrophobic interaction.

26. The method of claim 18, wherein determining a pose for the candidate ligand comprises determining a pose for each ligand in the set of ligands that minimizes the potential of all ligands.

27. The method of claim 26 wherein the potential is determined by:

$$E(L) = (n - 1)E^{\text{dock}}(L) + E^{\text{pairwise}}(L),$$

where $L = (\ell_1, \dots, \ell_n)$ indexes the candidate poses for each ligand, n is the number of ligands, E^{dock} represents the standard docking scores for the set of ligand poses, and E^{pairwise} represents a pairwise contribution to the total score.

28. The method of claim 27, wherein $E^{\text{pairwise}}(L) = \sum_{(i,j), i \neq j} e^{\text{pairwise}}(\ell_i, \ell_j)$.

29. The method of claim 27, wherein e^{pairwise} is determined using a machine learning algorithm.

30. The method of 29, wherein the machine learning algorithm is fit to a benchmark set of pairs of correct poses and pairs of incorrect poses for a diverse set of proteins.

31. The method of claim 26, wherein determining a pose for each ligand in the set of ligands comprises utilizing an expectation-maximization algorithm.
32. The method of claim 26, further comprising altering a pose of a ligand in the set of ligands to optimize that ligand's contribution to the potential function.
33. A method for lead optimization, comprising:
obtaining a set of ligands known to bind to a target protein;
determining a pose for each ligand in the set of ligands by generating a set of candidate poses and scores for ligands in the set of ligands, wherein the pose is based on a similarity of the candidate poses and scores for each ligand in the set of ligands;
determining a binding affinity of a candidate ligand by determining whether the candidate ligands binds to the target protein based on similarities to each ligand in the set of ligands and determining an energy if the candidate ligand binds to the target ligand;
and
modifying a location of the candidate ligand.
34. The method of claim 33, further comprising identifying a modifiable location on the candidate ligand, and wherein the modifying a location step comprises modifying the modifiable location.
35. The method of claim 33, wherein modifying the location increases binding affinity of the candidate ligand.
36. The method of claim 33, further comprising determining a binding affinity of the modified candidate ligand to the target protein based on similarities to each ligand in the set of ligands and determining an energy if the candidate ligand binds to the target ligand.
37. The method of claim 33, wherein the similarity is based on pairwise comparisons between each ligand in the set of ligands and every other ligand in the set of ligands.

38. The method of claim 37, wherein the similarities are further based on pairwise comparisons of substructures between each ligand in the set of ligands and every other ligand in the set of ligands and pairwise comparisons between the candidate ligand and every ligand in the set of ligands.

39. The method of claim 37, wherein the similarities are further based on pairwise comparisons of a protein-ligand interaction between each ligand in the set of ligands and every other ligand in the set of ligands and pairwise comparisons between the candidate ligand and every ligand in the set of ligands.

40. The method of claim 39, wherein the protein-ligand interaction is selected from the group consisting of: a hydrogen bond, a salt bridge, and a hydrophobic interaction.

41. The method of claim 33, wherein determining a pose for the candidate ligand comprises determining a pose for each ligand in the set of ligands that minimizes the potential of all ligands.

42. The method of claim 41, wherein the potential is determined by:

$$E(L) = (n - 1)E^{\text{dock}}(L) + E^{\text{pairwise}}(L),$$

where $L = (\ell_1, \dots, \ell_n)$ indexes the candidate poses for each ligand, n is the number of ligands, E^{dock} represents the standard docking scores for the set of ligand poses, and E^{pairwise} represents a pairwise contribution to the total score.

43. The method of claim 42, wherein $E^{\text{pairwise}}(L) = \sum_{(i,j), i \neq j} e^{\text{pairwise}}(\ell_i, \ell_j)$.

44. The method of claim 42, wherein e^{pairwise} is determined using a machine learning algorithm.

45. The method of 44, wherein the machine learning algorithm is fit to a benchmark set of pairs of correct poses and pairs of incorrect poses for a diverse set of proteins.

46. The method of claim 41, wherein determining a pose for each ligand in the set of ligands comprises utilizing an expectation-maximization algorithm.
47. The method of claim 41, further comprising altering a pose of a ligand in the set of ligands to optimize that ligand's contribution to the potential function.
48. A method for predicting ligand-biomolecule interaction comprising:
obtaining a set of ligands known to bind to a target biomolecule;
generating a set of candidate poses for each ligand in the set of ligands and a candidate ligand; and
determining a pose for the candidate ligand based on a similarity of the candidate poses for the candidate ligand and the candidate poses for each ligand in the set of ligands.
49. The method of claim 48, wherein generating a set of candidate poses is accomplished by docking at least one ligand from the set of ligands using docking software.
50. The method of claim 48, further comprising generating a set of scores for each ligand using docking software.
51. The method of claim 50, wherein the docking software is selected from the group consisting of: Glide, DOCK, AutoDock VINA, FlexX, GOLD, and Surflex.
52. The method of claim 48, wherein at least one ligand in the set of ligands has an experimentally determined pose.
53. The method of claim 52, wherein the pose for the at least one ligand is set to the experimentally determined pose.

54. The method of claim 48, wherein the similarity is based on pairwise comparisons between each ligand in the set of ligands and every other ligand in the set of ligands and pairwise comparisons between the candidate ligand and every ligand in the set of ligands.

55. The method of claim 54, wherein the similarity is further based on pairwise comparisons of substructures between each ligand in the set of ligands and every other ligand in the set of ligands and pairwise comparisons between the candidate ligand and every ligand in the set of ligands.

56. The method of claim 54, wherein the similarity is further based on pairwise comparisons of a biomolecule-ligand interaction between each ligand in the set of ligands and every other ligand in the set of ligands and pairwise comparisons between the candidate ligand and every ligand in the set of ligands.

57. The method of claim 56, wherein the biomolecule-ligand interaction is selected from the group consisting of: a hydrogen bond, a salt bridge, and a hydrophobic interaction.

58. The method of claim 48, wherein determining a pose for the candidate ligand comprises determining a pose for each ligand in the set of ligands that minimizes the potential of all ligands.

59. The method of claim 58, wherein the potential is determined by:

$$E(L) = (n - 1)E^{\text{dock}}(L) + E^{\text{pairwise}}(L),$$

where $L = (\ell_1, \dots, \ell_n)$ indexes the candidate poses for each ligand, n is the number of ligands, E^{dock} represents the standard docking scores for the set of ligand poses, and E^{pairwise} represents a pairwise contribution to the total score.

60. The method of claim 59, wherein $E^{\text{pairwise}}(L) = \sum_{(i,j), i \neq j} e^{\text{pairwise}}(\ell_i, \ell_j)$.

61. The method of claim 59, wherein $e^{pairwise}$ is determined using a machine learning algorithm.
62. The method of 61, wherein the machine learning algorithm is fit to a benchmark set of pairs of correct poses and pairs of incorrect poses for a diverse set of proteins.
63. The method of claim 58, wherein determining a pose for each ligand in the set of ligands comprises utilizing an expectation-maximization algorithm.
64. The method of claim 58, further comprising altering a pose of a ligand in the set of ligands or the candidate ligand to optimize that ligand's contribution to the potential function.
65. The method of claim 48, wherein the biomolecule is selected from a protein, a carbohydrate, a DNA molecule, an RNA molecule, and combinations thereof.

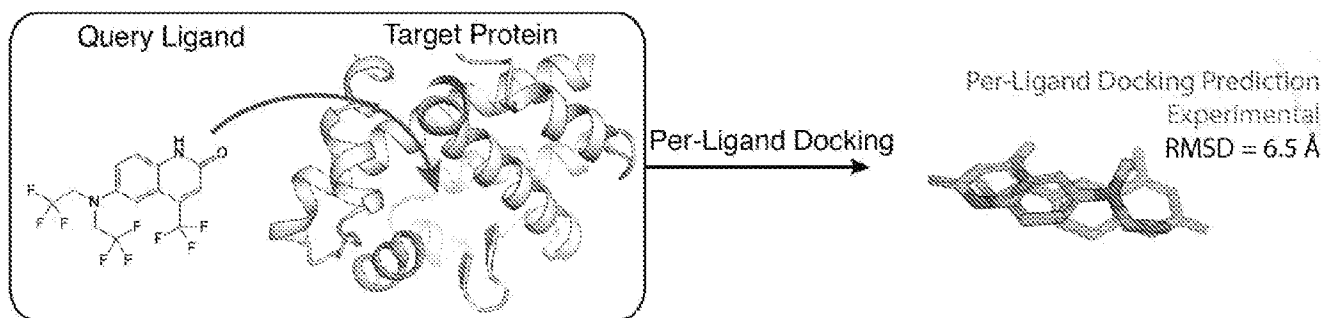


Fig. 1A

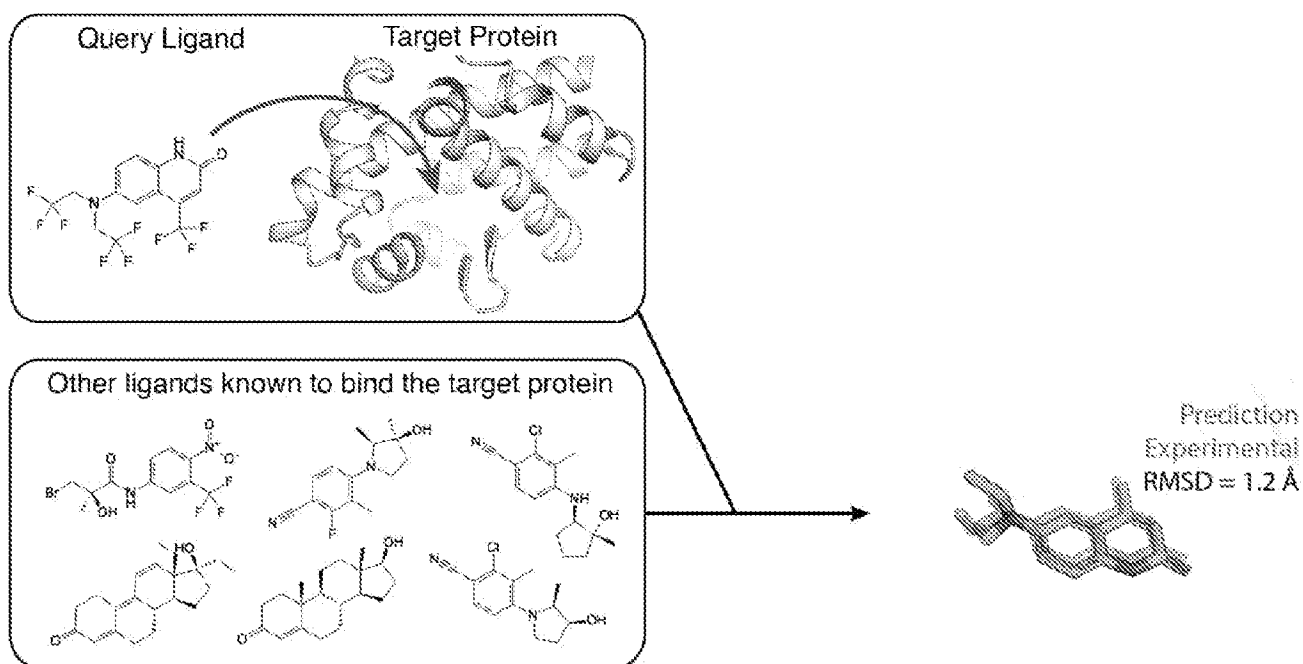


Fig. 1B

Distinct ligands tend to form similar interactions with the target

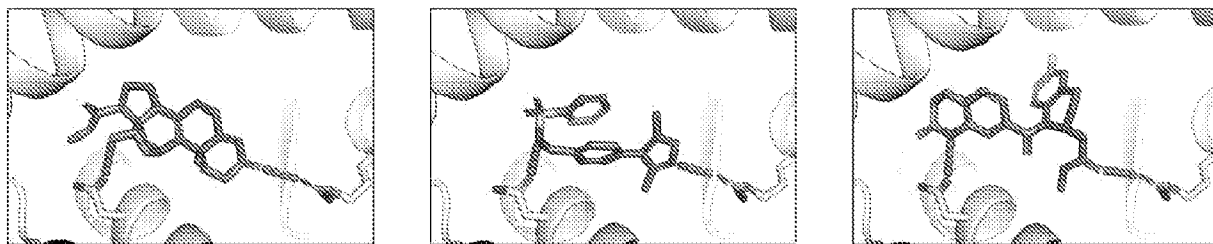
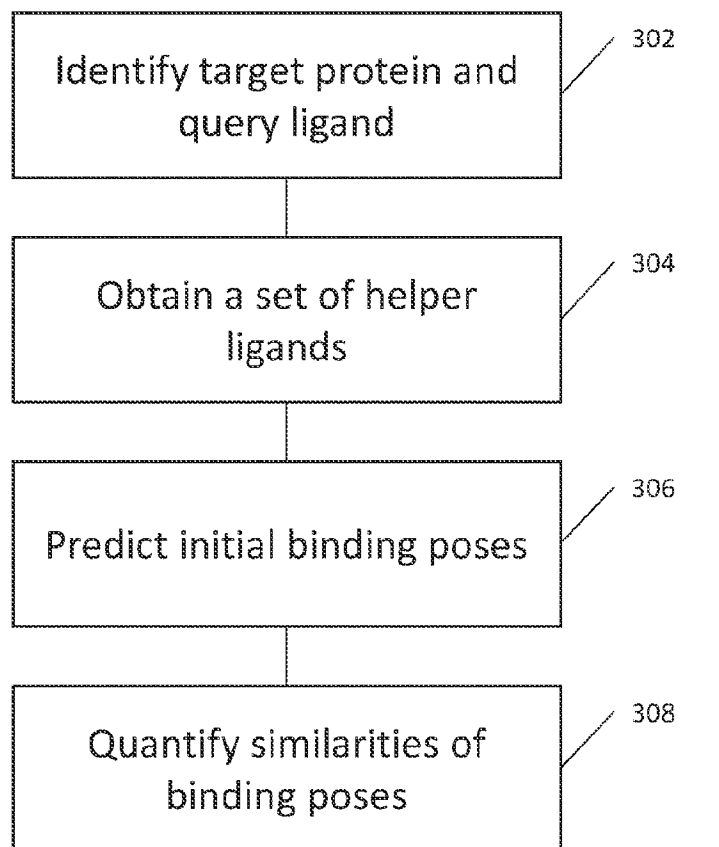


Fig. 2



300

Fig. 3

Per-ligand docking underestimates the structural similarity of distinct ligands' binding poses

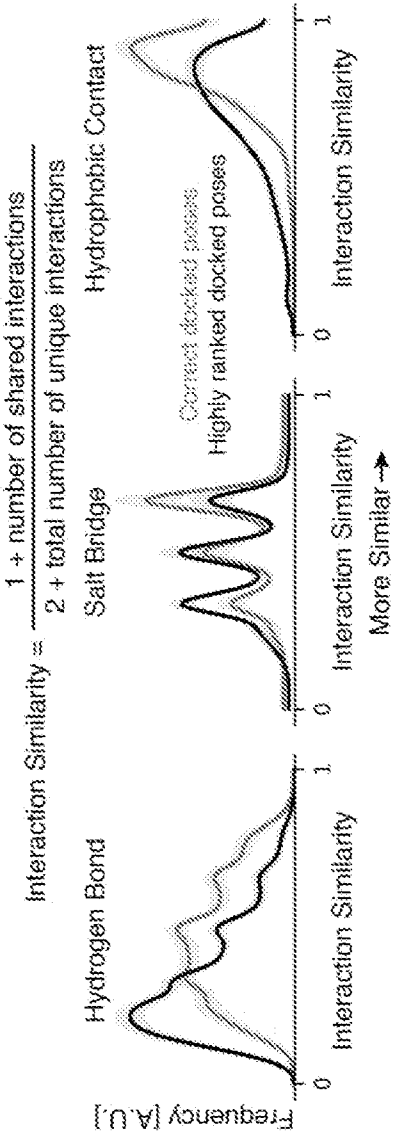


Fig. 4A

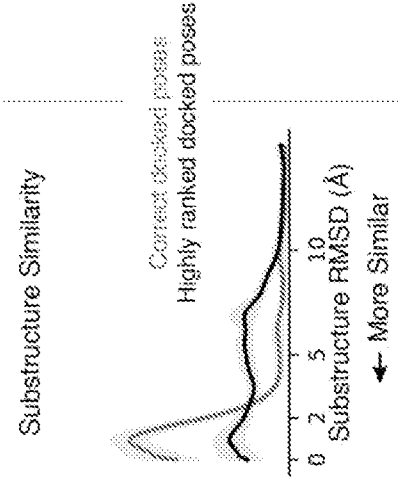


Fig. 4B

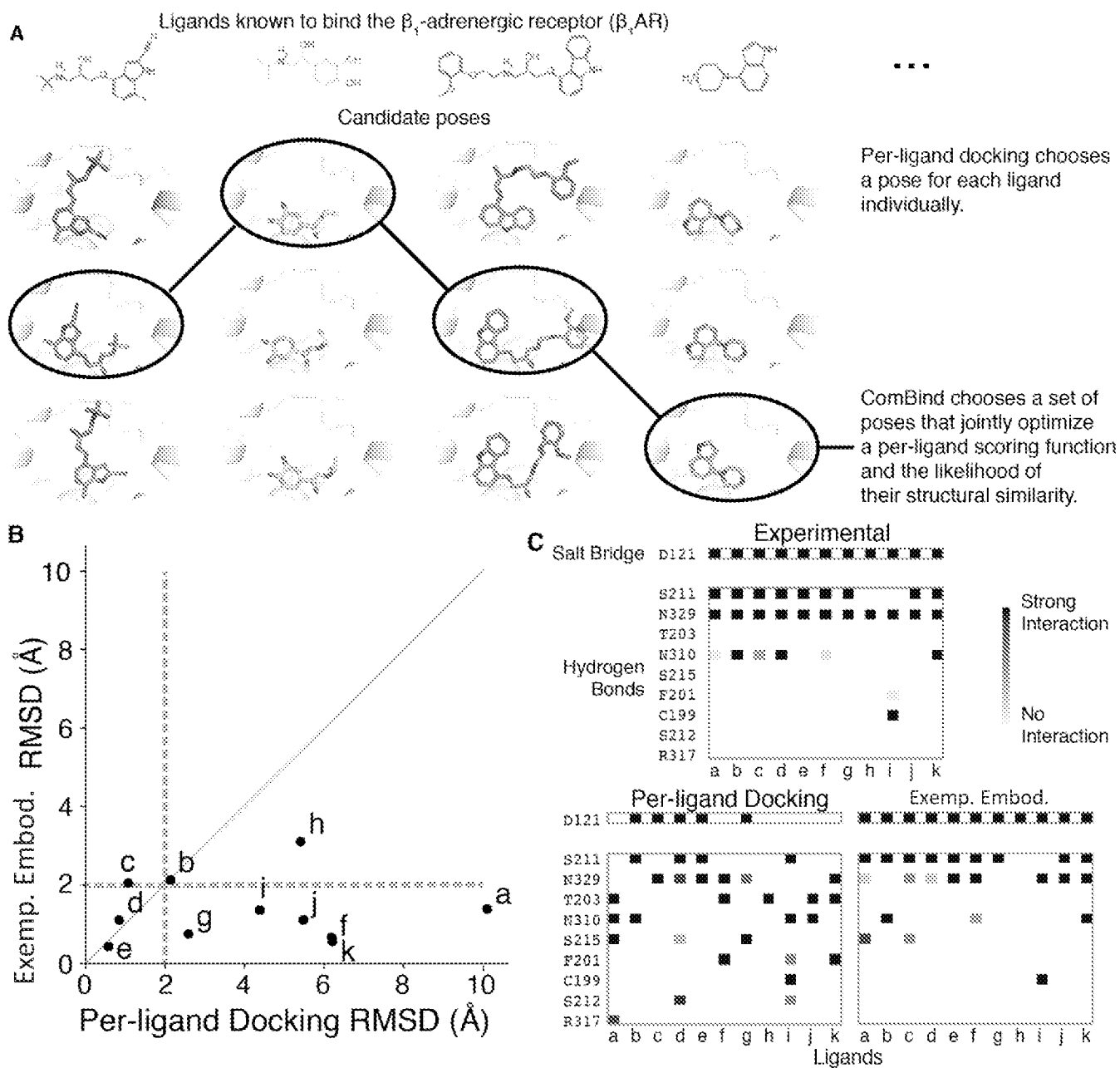


Fig. 5

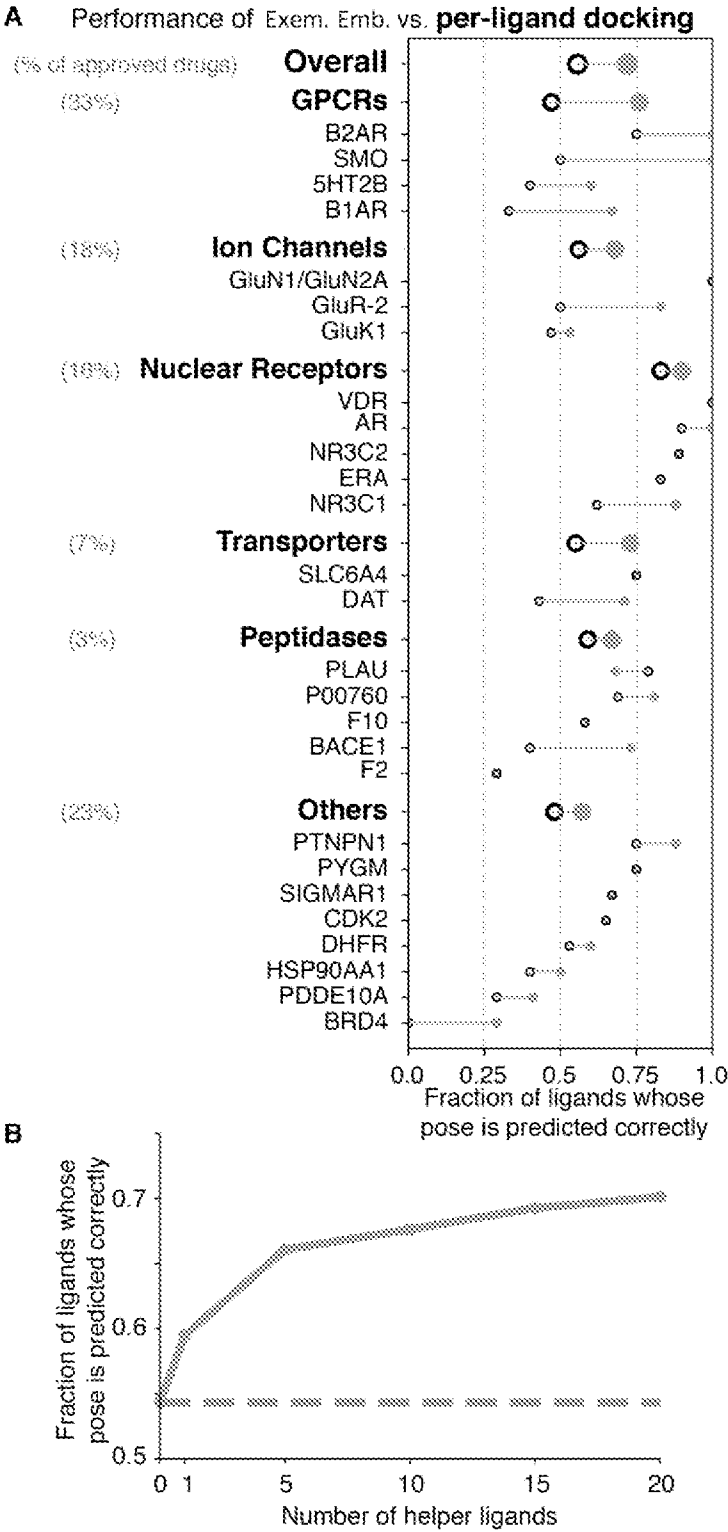
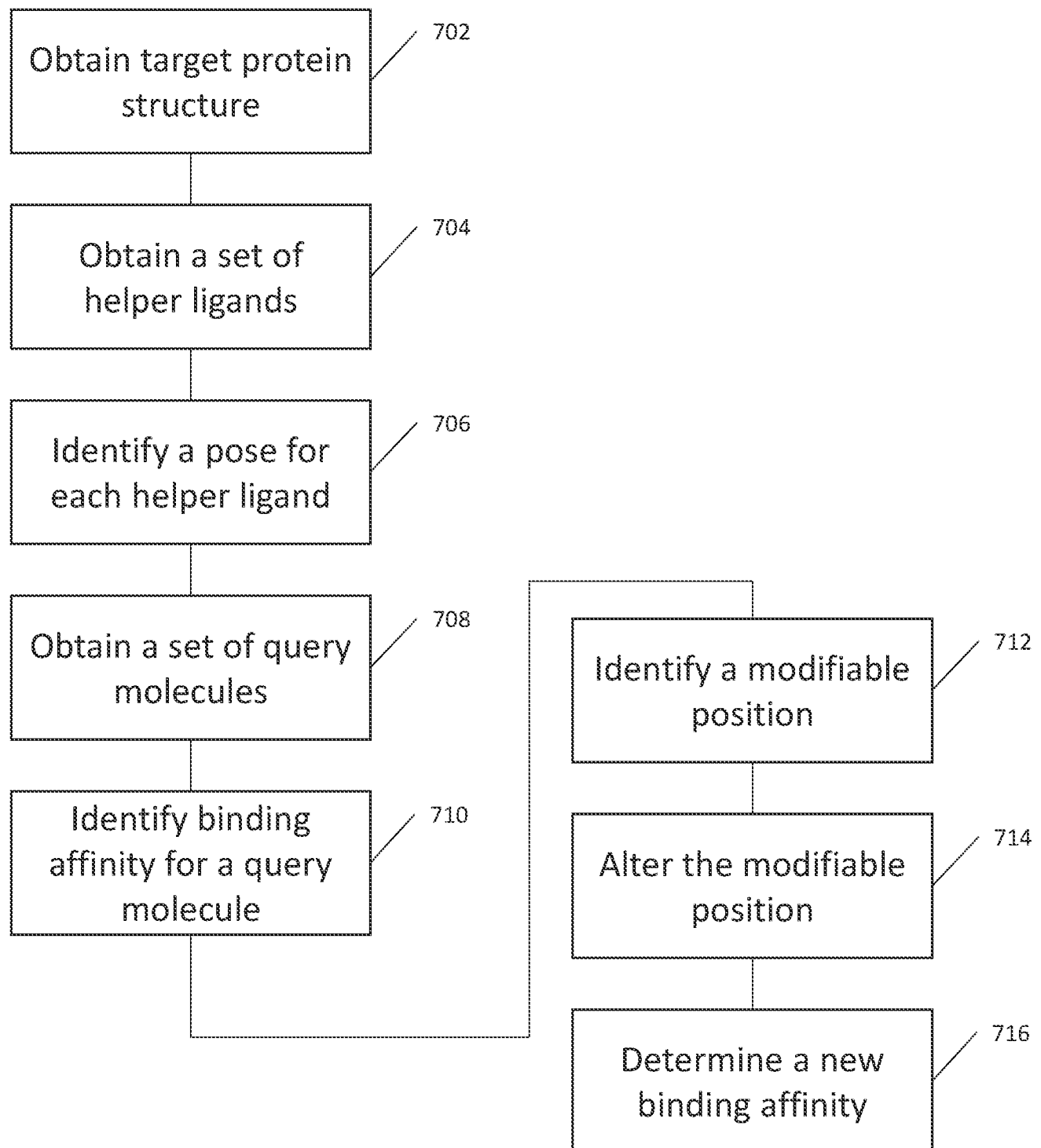


Fig. 6

**Fig. 7**

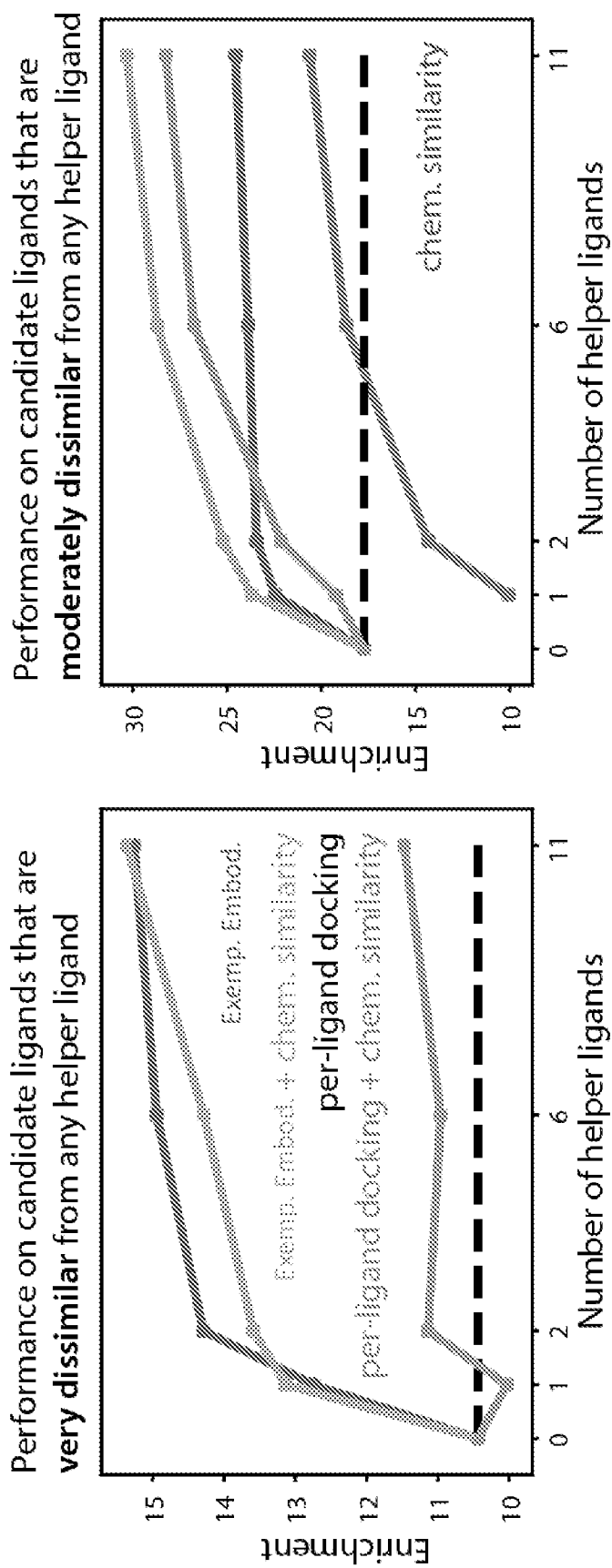


Fig. 8A

Fig. 8B

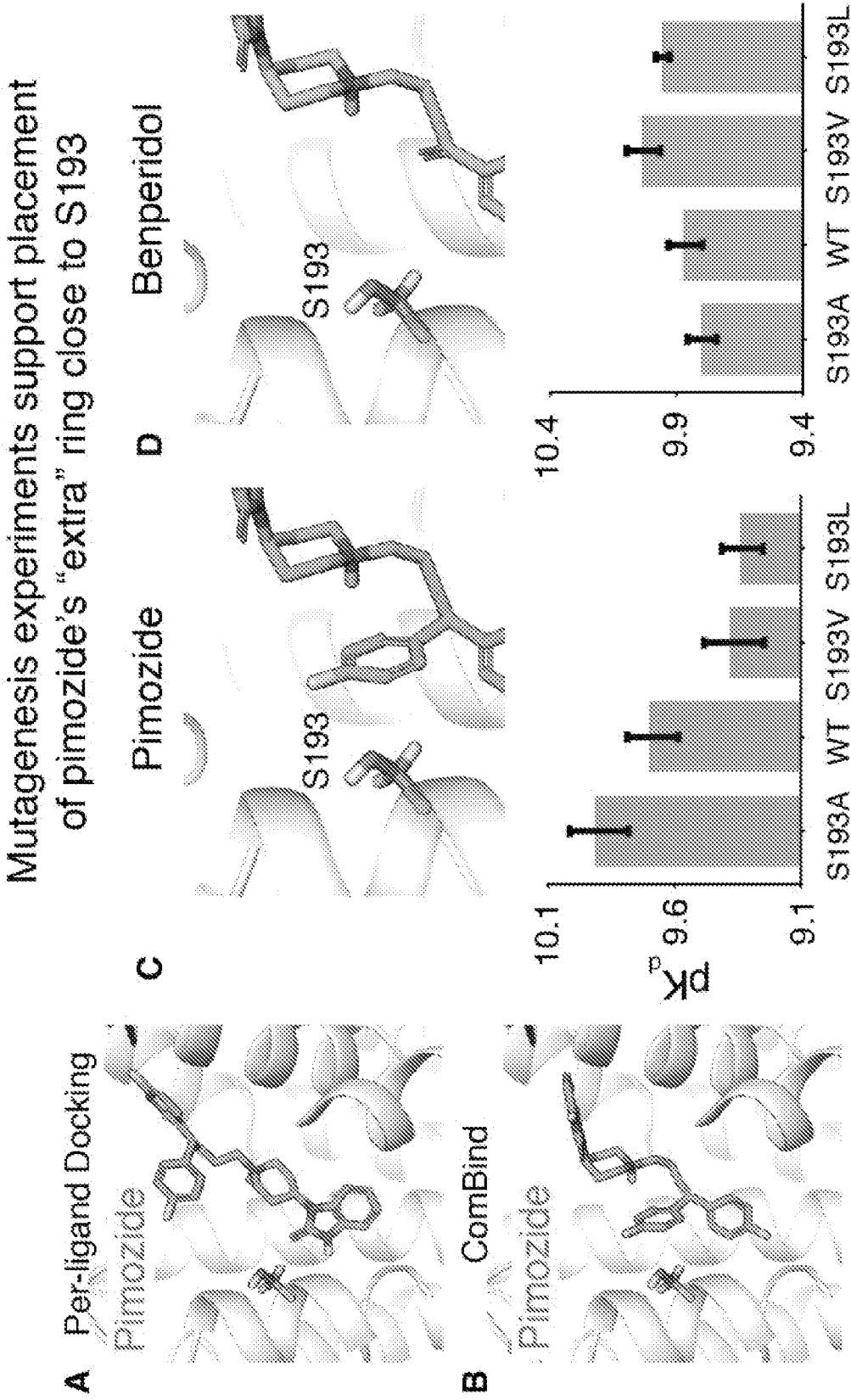


Fig. 9

INTERNATIONAL SEARCH REPORT

International application No.

PCT/US 21/34637

A. CLASSIFICATION OF SUBJECT MATTER

IPC - G06G 7/58, G01N 33/566 (2021.01)

CPC - G16B 15/00, G16B 15/30, G16C 20/50, G01N 33/6803, G16B 35/20

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

See Search History document

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

See Search History document

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

See Search History document

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
Y --- A	US 2010/0112724 A1 to Tovbin et al. (hereinafter Tovbin), 06 May 2010 (06.05.2010), entire document, especially Abstract; para [0011], [0128]	1-11, 16-26, 31-41, 46-58, 63-65 ----- 12-15, 27-30, 42-45, 59-62
Y -- A	Cleves et al. "Electrostatic-field and surface-shape similarity for virtual screening and pose prediction." In: Journal of Computer-Aided Molecular Design volume 33, pages 865-886 (2019), [online] [retrieved on 24 August 2021 (24.08.2021)] Retrieved from the Internet < URL: https://link.springer.com/article/10.1007/s10822-019-00236-6 >, entire document,	1-11, 16-26, 31-41, 46-58, 63-65 ----- 12-15, 27-30, 42-45, 59-62
A	US 2019/0369097 A1 (ALBERT EINSTEIN COLLEGE OF MEDICINE, INC.), 05 December 2019 (05.12.2019), entire document	1-65
A	US 2011/0112818 A1 (GODDARD, III et al.), 12 May 2011 (12.05.2011), entire document	1-65

☐ Further documents are listed in the continuation of Box C.☐ See patent family annex.

* Special categories of cited documents:

"A" document defining the general state of the art which is not considered to be of particular relevance

"D" document cited by the applicant in the international application

"E" earlier application or patent but published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art

"&" document member of the same patent family

Date of the actual completion of the international search

25 August 2021

Date of mailing of the international search report

SEP 30 2021

Name and mailing address of the ISA/US

Mail Stop PCT, Attn: ISA/US, Commissioner for Patents
P.O. Box 1450, Alexandria, Virginia 22313-1450

Facsimile No. 571-273-8300

Authorized officer

Kari Rodriguez

Telephone No. PCT Helpdesk: 571-272-4300