(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
15 May 2025 (15.05.2025)

WIPO | PCT

(10) International Publication Number
WO 2025/101482 A1

(51) International Patent Classification:
G06N 3/08 (2023.01) *G16B 15/00* (2019.01)
G16B 40/00 (2019.01) *G06N 3/0455* (2023.01)
G16B 15/30 (2019.01) *G16C 20/50* (2019.01)

(21) International Application Number:
PCT/US2024/054511

(22) International Filing Date:
05 November 2024 (05.11.2024)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
63/547,493 06 November 2023 (06.11.2023) US

(71) Applicant: **UNIVERSITY OF SOUTHERN CALIFORNIA** [US/US]; 1150 South Olive Street, Suite 2300, Los Angeles, California 90015 (US).

(72) Inventors: **WELLER, Jesse**; c/o University of Southern California, 1150 South Olive Street, Suite 2300, Los Angeles, California 90015 (US). **ROHS, Remo**; c/o University of Southern California, 1150 South Olive Street, Suite 2300, Los Angeles, California 90015 (US).

(74) Agent: **GERSTING, Jason** et al.; 600 Anton Boulevard, Suite 1400, Costa Mesa, California 92626 (US).

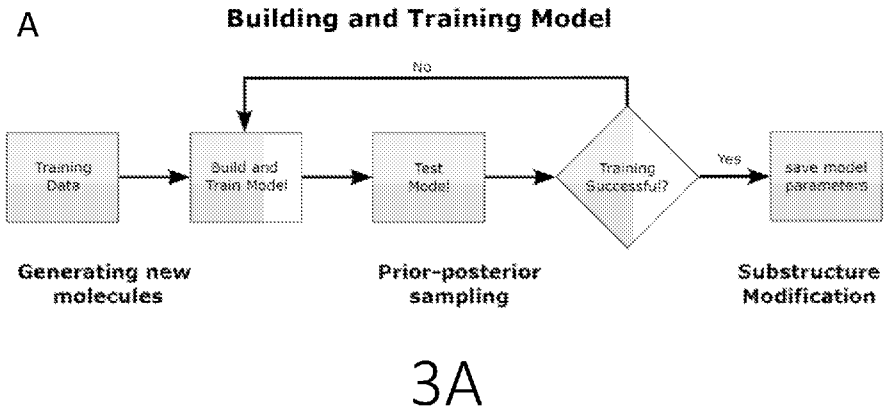
(81) Designated States *(unless otherwise indicated, for every kind of national protection available)*: AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CV, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IQ, IR, IS, IT, JM, JO, JP, KE, KG, KH, KN, KP, KR, KW, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, MG, MK, MN, MU, 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, CV, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SC, 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, ME, 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).

Declarations under Rule 4.17:
— *as to applicant's entitlement to apply for and be granted a patent (Rule 4.17(ii))*

Published:
— *with international search report (Art. 21(3))*

(54) Title: DEEP HIERARCHICAL VARIATION AUTOENCODER FOR THE STRUCTURE-BASED DESIGN OF DRUG-LIKE MOLECULES



3A

(57) Abstract: Provided for herein are computer implemented methods and a corresponding system comprising a generative model with a hierarchical variational autoencoder (HVAE) architecture for generating a plurality of new molecular structures based on the model being trained with one or more sources of training data. In several embodiments, the output molecules have improved characteristics of choice, such as enhanced binding affinity for a target receptor.

WO 2025/101482 A1

DEEP HIERARCHICAL VARIATION AUTOENCODER FOR THE STRUCTURE-BASED DESIGN OF DRUG-LIKE MOLECULES

CROSS-REFERENCE TO RELATED APPLICATIONS

[0001] This application claims the benefit of priority to United States Provisional Patent Application No. 63/547493, filed November 6, 2023, the entire contents of which is incorporated by reference herein.

STATEMENT AS TO FEDERALLY SPONSORED RESEARCH

[0002] This invention was made with government support under grant no. R35GM130376, awarded by the (NIH/NIGMS) National Institute of General Medical Sciences. The government has certain rights in the invention.

FIELD

[0003] Recently, the remarkable growth of available crystal structure data and libraries of commercially available or readily synthesizable molecules have unlocked previously inaccessible regions of chemical space for drug development. Paired with improvements in virtual ligand screening methods, these expanded libraries are having a notable impact on early drug design efforts. Yet screening-based methods still face scalability limits, due to computational constraints and the sheer scale of drug-like space.

SUMMARY

[0004] Machine learning approaches can serve to overcome the current limitations in drug design by learning the fundamental intra- and intermolecular relationships in drug-target systems from existing data. Provided for herein is a system, apparatus and/or method for generating a plurality of molecular structures having a predictable binding affinity for a target protein (e.g., a target receptor), where a processor (e.g., a hierarchical variational autoencoder (HVAE) architecture) is disclosed herein to perform the machine learning algorithms.

[0005] In several embodiments, there is provided a computer-implemented method to identify new molecular structures having a desired set of characteristics that are beneficial and/or improved over known compounds and that bind a target protein, such as a target receptor. In several embodiments, the method comprises a) receiving a set of training data, wherein the training data comprises one or more of: (i) crystal structure data of a target protein (e.g., a receptor); (ii) a

predicted three-dimensional structure of a target protein; and/or (iii) a chemical structure of a known ligand for the target protein, b) training, using the set of training data, a model to generate, in response to input, output comprising a plurality of molecular structures having a predictable binding affinity for a target receptor, wherein the model comprises a ligand encoder, a protein encoder, and a ligand decoder, c) generating, based on (i) and (ii), a prior molecular density grid representing prior information related to the atomic density of (i) and (ii), wherein the atomic density comprises atomic occupancy data and atomic features data, wherein atomic occupancy channels in the prior molecular density grid are sampled at full resolution and atomic feature channels in the prior molecular density grid are downsampled using a downsampling layer, d) generating, based on (iii), a posterior molecular density grid representing posterior information related to the atomic density of (iii), wherein the atomic density comprises atomic occupancy data and atomic features data, wherein atomic occupancy channels in the prior molecular density grid are sampled at full resolution and atomic feature channels in the prior molecular density grid are downsampled using a downsampling layer, e) generating, using the model, a plurality of candidate ligands by: i) inputting a target protein structure into the protein encoder, ii) inputting a randomly selected set of values from the prior molecular density grid into the ligand encoder, and iii) generating, using the protein encoder and the ligand encoder, a first plurality of molecular density grids in a hierarchical latent space, representing one or more encodings of the target protein structure and one or more encodings of the randomly selected set of values, and concurrently: i) inputting the target protein structure into the protein encoder a second time, ii) inputting a randomly selected set of values from the posterior molecular density grid into the ligand encoder, and iii) generating, using the protein encoder and the ligand encoder, a second plurality of molecular density grids in the hierarchical latent space, representing one or more encodings of the target protein structure and one or more encodings of the randomly selected set of values from the previously encoded ligand, and, in combination, f) generating, based on the first and second plurality of molecular density grids, using the ligand decoder, an output molecular density grid representing atomic densities of the candidate ligand, wherein the atomic density comprises atomic occupancy data and atomic features data, wherein the atomic occupancy data from the first and second plurality of molecular density gradients is sampled at full resolution and atomic feature data from the first and second plurality of molecular density gradients is up-sampled to full resolution, g) iteratively placing atomic coordinates within the output molecular density grid, h)

assigning atomic features to each atom using density from the upsampled atomic feature data, and i) fitting chemical bonds between the assigned atom coordinates based on the geometry and features of the atomic density data, thereby generating a molecular structure having a predictable binding affinity for the target receptor.

[0006] In several embodiments, the hierarchal latent space further comprises a set of tensors. Depending on the embodiment, each tensor spatially represents a reference ligand-receptor complex latent representation formed from at least one existing ligand and the target receptor or representing a reference ligand-receptor complex formed from a molecular structure that is generated and the target receptor.

[0007] In several embodiments, the generation of the output molecular density grid further comprises a prior-posterior sampling process. In several embodiments, the prior-posterior sampling process comprises interpolating toward a randomly sampled set of tensors in the latent space beginning from a reference ligand-receptor complex sampled from a latent distribution. In several embodiments, the prior-posterior sampling is configured to optimize a binding affinity (or another desired characteristic) of a molecular structure generated by the method and the target receptor. In several embodiments, the prior-posterior sampling configured to optimize the binding affinity further comprises a decoding of the interpolation into new molecular densities.

[0008] In several embodiments, iteratively placing atomic coordinates within the output molecular density grid further comprises application of an atom-fitting algorithm configured to iteratively place atomic coordinates for a chosen atom type within the output molecular density grid to minimize a residual occupancy density. In several embodiments, the fitting of chemical bonds between the assigned atom types within the atomic occupancy densities comprises application of a bond fitting algorithm.

[0009] In several embodiments, while generating the prior molecular density grid, each atom of the ligand is represented by a multi-channel pseudo-Gaussian density with channels. In several embodiments, each channel represents a different atomic feature of each atom.

[0010] In several embodiments, wherein the atomic features comprise the elements C, N, O, F, P, S, Cl, Br, and I, and combinations thereof. Any other desired element from the periodic table may be used in addition to, or in place of those listed. In several embodiments, the features of each element comprise status as H-bond donor, H-bond acceptor, aromaticity, positive (+) charge, negative (-) charge, or neutral (0) charge.

[0011] In several embodiments, at least one of the prior molecular density grid, the posterior molecular density grid, the first plurality of molecular density grids, second plurality of molecular density grids, and the output molecular density grid measures 24 x 24 x 24 Angstroms and the molecular density grid can be an input and/or output data structure. Other molecular density grid sizes are used in other embodiments, depending on target protein size, desired ligand size, computing processing resources, or combinations thereof.

[0012] In several embodiments, at least one of the prior molecular density grid, the posterior molecular density grid, the first plurality of molecular density grids, second plurality of molecular density grids, and the output molecular density grid measures is a molecular graph, a point cloud, and/or point set where the molecular graph is comprised of a set of atomic coordinates (vertices), bonds (edges), atom (vertex) types, and bond (edge) types and the point cloud or point set is a graph without edges or edge types.

[0013] In several embodiments, there is provided a system comprising one or more computing devices associated with a processor and a memory for executing computer-executable instructions configured to a) receive a set of training data, the training data comprising one or more of: (i) crystal structure data of a target protein (e.g., a receptor), (ii) a predicted three-dimensional structure of a target protein, and/or (iii) a chemical structure of a known ligand for the target protein, b) train, using the set of training data, a model to generate, in response to input, output comprising a plurality of molecular structures having a predictable binding affinity for a target receptor, wherein the model comprises a ligand encoder, a protein encoder, and a ligand decoder, c) generate, based on (i) and (ii), a prior molecular density grid representing prior information related to the atomic density of (i) and (ii), wherein the atomic density comprises atomic occupancy data and atomic features data, wherein atomic occupancy channels in the prior molecular density grid are sampled at full resolution and atomic feature channels in the prior molecular density grid are downsampled using a downsampling layer, d) generate, based on (iii), a posterior molecular density grid representing posterior information related to the atomic density of (iii), wherein the atomic density comprises atomic occupancy data and atomic features data, wherein atomic occupancy channels in the prior molecular density grid are sampled at full resolution and atomic feature channels in the prior molecular density grid are downsampled using a downsampling layer, e) generate, using the model, a plurality of candidate ligands by: i) inputting a target protein structure into the protein encoder, ii) inputting a randomly selected set of values from the prior

molecular density grid into the ligand encoder, and iii) generating, using the protein encoder and the ligand encoder, a first plurality of molecular density grids in a hierarchical latent space, representing one or more encodings of the target protein structure and one or more encodings of the randomly selected set of values, and concurrently: i) inputting the target protein structure into the protein encoder a second time, ii) inputting a randomly selected set of values from the posterior molecular density grid into the ligand encoder, and iii) generating, using the protein encoder and the ligand encoder, a second plurality of molecular density grids in the hierarchical latent space, representing one or more encodings of the target protein structure and one or more encodings of the randomly selected set of values from the previously encoded ligand; and, in combination, f) generate, based on the first and second plurality of molecular density grids, using the ligand decoder, an output molecular density grid representing atomic densities of the candidate ligand, wherein the atomic density comprises atomic occupancy data and atomic features data, wherein the atomic occupancy data from the first and second plurality of molecular density gradients is sampled at full resolution and atomic feature data from the first and second plurality of molecular density gradients is up-sampled to full resolution, g) iteratively place atomic coordinates within the output molecular density grid, h) assign atomic features to each atom using density from the upsampled atomic feature data, and i) fit chemical bonds between the assigned atom coordinates based on the geometry and features of the atomic density data, thereby generating a molecular structure having a predictable binding affinity for the target receptor.

[0014] In several embodiments, the hierarchal latent space further comprises a set of tensors. In several embodiments, each tensor spatially represents a reference ligand-receptor complex latent representation formed from at least one existing ligand and the target receptor. In several embodiments, each tensor spatially represents a reference ligand-receptor complex formed from a molecular structure thus generated and the target receptor.

[0015] In several embodiments, generating the output molecular density grid further comprises a prior-posterior sampling process. In several embodiments, the prior-posterior sampling process comprises interpolating toward a randomly sampled set of tensors in the latent space beginning from a reference ligand-receptor complex sampled from a latent distribution. In several embodiments, the prior-posterior sampling is configured to optimize a binding affinity (or other desired characteristic) of a molecular structure generated by the method and the target receptor.

[0016] In several embodiments, the prior-posterior sampling further comprises a decoding of the interpolation into new molecular densities.

[0017] In several embodiments, the iterative placement of atomic coordinates within the output molecular density grid further comprises application of an atom-fitting algorithm configured to iteratively place atomic coordinates for a chosen atom type within the molecular density grid to minimize a residual occupancy density.

[0018] In several embodiments, the fitting of chemical bonds between the assigned atom types within the atomic occupancy densities comprises application of a bond fitting algorithm.

[0019] In several embodiments, when generating the prior molecular density grid, each atom of the ligand is represented by a multi-channel pseudo-Gaussian density with channels, each channel representing a different atomic feature of each atom.

[0020] In several embodiments, wherein the atomic features comprise the elements C, N, O, F, P, S, Cl, Br, and I, and combinations thereof. Any other desired element from the periodic table may be used in addition to, or in place of those listed. In several embodiments, the features of each element comprise status as H-bond donor, H-bond acceptor, aromaticity, positive (+) charge, negative (-) charge, or neutral (0) charge.

[0021] In several embodiments, at least one of the prior molecular density grid, the posterior molecular density grid, the first plurality of molecular density grids, second plurality of molecular density grids, and the output molecular density grid measures 24 x 24 x 24 Angstroms and the molecular density grid can be an input and/or output data structure. Other molecular density grid sizes are used in other embodiments, depending on target protein size, desired ligand size, computing processing resources, or combinations thereof.

[0022] In several embodiments, at least one of the prior molecular density grid, the posterior molecular density grid, the first plurality of molecular density grids, second plurality of molecular density grids, and the output molecular density grid measures is a molecular graph, a point cloud, and/or point set where the molecular graph is comprised of a set of atomic coordinates (vertices), bonds (edges), atom (vertex) types, and bond (edge) types and the point cloud or point set is a graph without edges or edge types.

[0023] In several embodiments, there is provided a non-transitory computer-readable medium having instructions stored thereon that, when executed by one or more processors, cause the one or more processors to perform operations comprising a) receiving a set of training data, the training

data comprising one or more of: (i) crystal structure data of a target (e.g., receptor), (ii) a predicted three-dimensional structure of a target protein, and/or (iii) a chemical structure of a known ligand for the target protein, b) training, using the set of training data, a model to generate, in response to input, output comprising a plurality of molecular structures having a predictable binding affinity for a target receptor, wherein the model comprises a ligand encoder, a protein encoder, and a ligand decoder, c) generating, based on (i) and (ii), a prior molecular density grid representing prior information related to the atomic density of (i) and (ii), wherein the atomic density comprises atomic occupancy data and atomic features data, wherein atomic occupancy channels in the prior molecular density grid are sampled at full resolution and atomic feature channels in the prior molecular density grid are downsampled using a downsampling layer, d) generating, based on (iii), a posterior molecular density grid representing posterior information related to the atomic density of (iii), wherein the atomic density comprises atomic occupancy data and atomic features data, wherein atomic occupancy channels in the prior molecular density grid are sampled at full resolution and atomic feature channels in the prior molecular density grid are downsampled using a downsampling layer, e) generating, using the model, a plurality of candidate ligands by: i) inputting a target protein structure into the protein encoder, ii) inputting a randomly selected set of values from the prior molecular density grid into the ligand encoder, and iii) generating, using the protein encoder and the ligand encoder, a first plurality of molecular density grids in a hierarchical latent space, representing one or more encodings of the target protein structure and one or more encodings of the randomly selected set of values, and concurrently: i) inputting the target protein structure into the protein encoder a second time, ii) inputting a randomly selected set of values from the posterior molecular density grid into the ligand encoder, and iii) generating, using the protein encoder and the ligand encoder, a second plurality of molecular density grids in the hierarchical latent space, representing one or more encodings of the target protein structure and one or more encodings of the randomly selected set of values from the previously encoded ligand, and, in combination, f) generating, based on the first and second plurality of molecular density grids, using the ligand decoder, an output molecular density grid representing atomic densities of the candidate ligand, wherein the atomic density comprises atomic occupancy data and atomic features data, wherein the atomic occupancy data from the first and second plurality of molecular density gradients is sampled at full resolution and atomic feature data from the first and second plurality of molecular density gradients is up-sampled to full resolution, g) iteratively placing

atomic coordinates within the output molecular density grid, h) assigning atomic features to each atom using density from the upsampled atomic feature data, and i) fitting chemical bonds between the assigned atom coordinates based on the geometry and features of the atomic density data, thereby generating a molecular structure having a predictable binding affinity for the target receptor.

BRIEF DESCRIPTION OF THE FIGURES

[0024] Figure 1 depicts a non-limiting schematic of a system for generating a plurality of molecular structures having a predictable binding affinity for a target receptor.

[0025] Figures 2A-2B schematically depict conceptual representations of generative models. Figure 2A shows a non-limiting schematic of a training dataset of molecular structures that are processed by a generative model to output novel molecular structures. Figure 2B shows a non-limiting schematic of a training dataset of drug-target complexes that are processed by a generative model to output novel ligand predicted to bind the target.

[0026] Figures 3A-3D show non-limiting flowcharts depicting applications of generative models as provided for herein. Figure 3A shows a non-limiting flowchart depicting the building and training of a generative model as provided for herein. Figure 3B shows a non-limiting flowchart depicting use of a generative model as provided for herein to generate new molecules with desired properties. Figure 3C shows a non-limiting flowchart depicting use of a generative model as provided for herein to implement prior-posterior sampling to new chemical structures with a high degree of control over variation in molecular details. Figure 3D shows a non-limiting flowchart depicting use of a generative model as provided for herein to accomplish substructure modification.

[0027] Figures 3E-3G show non-limiting flowcharts depicting applications of generative models as provided for herein. Figure 3E shows a non-limiting flowchart depicting use of a generative model as provided for herein to accomplish pattern replacement in a desired region of a molecular structure. Figure 3F shows a non-limiting flowchart depicting use of a generative model as provided for herein to design linkers that will function to link two (or more) portions of a molecular structure. Figure 3G shows a non-limiting flowchart depicting use of a generative model as provided for herein to optimize drug properties.

[0028] Figures 4A-4G show a non-limiting schematic of an embodiment of a generative model as provided for herein. Figure 4A shows various types of input data. Figure 4B shows a structural aspect of the model, a hierarchical structure, with varying spatial scales utilized. Figure 4C shows one embodiment of input of molecular information into the model as multi-channel atomic density grids, with ligand grids input into a ligand encoder and receptor grids input into a protein encoder. Figure 4D shows components of the generative models provided for herein which comprise a ligand encoder, a protein encoder, and a ligand decoder. Information from the protein encoder is pass into both the ligand encoder and ligand decoder. Figure 4E shows output from the model, atomic occupancy and atom-type grids, which are fit with a molecular structure. Figure 4F shows non-limiting examples of various atomic properties that can be used confirm the fit of a molecule. Figure 4G shows an example of dual resolution outputs that reduce memory consumption and grid sparsity on the machine running the model.

[0029] Figure 5A shows a non-limiting schematic of one application of a generative model as provided for herein, namely substructure modification.

[0030] Figures 5B-5C show non-limiting schematics of applications of a generative model as provided for herein. Figure 5B shows a schematic depiction of use of a generative model as provided for herein for linker design. Figure 5C shows a schematic depiction of use of a generative model as provided for herein for fragment growing.

[0031] Figures 5D-5E shows a non-limiting schematic of one application of a generative model as provided for herein, namely pattern replacement. Figure 5D schematically depicts the region of the molecule queried for a pattern of interest and regions that are modified (solid circles) and those held constant (dashed circles). Figure 5E shows data related to the resultant Vina scores for the generated molecules with the initial ligand score shown as the black dashed line.

[0032] Figure 6A depicts various data related to the use of a generative model as provided for herein for ligand optimization, using a ligand (PDB Chemical ID 57G) bound to human bromodomain-containing protein 4 receptor (PDB ID 5D3H). Circled values are indicative of successful improvement in the indicated parameter.

[0033] Figures 6B-6D depict additional data related to the use of a generative model as provided for herein for ligand optimization. Figure 6B shows the molecular structure of the initial molecule to be optimized. Figure 6C shows data related to optimization of hydrophobicity. Figure 6D shows data related to average diversity and similarity of ligands in the test set.

[0034] Figures 6E-6F depict additional data related to the use of a generative model as provided for herein for ligand optimization. Figure 6E shows data related to virtual screening of efficiency (generation of high affinity molecules) compared to random samples of molecules drawing from the ZINC drug-like database. Figure 6F shows a schematic of the prior-posterior sampling procedure.

[0035] Figure 6G depicts data related to ligand optimization, in particular, generation of molecules with optimized selectivity for human ataxia-telangiectasia mutated (ATM) kinase and against human vacuolar protein sorting (Vps34) kinase.

[0036] Figure 6H depicts various data related to the use of a generative model as provided for herein for ligand optimization, using the ligand (S)-1-[N2-(1-carboxy-3-phenylpropyl)-L lysyl]-L-proline dihydrate bound to the human angiotensin-converting enzyme (ACE) receptor (PDB ID 1086). This compound is an FDA approved ACE inhibitor in humans used treat hypertension, heart failure, and myocardial infarction. Circled values are indicative of successful improvement in the indicated parameter.

[0037] Figures 6I-6K depict additional data related to the use of a generative model as provided for herein for ligand optimization. Figure 6I shows the molecular structure of the initial molecule to be optimized. Figure 6J shows data related to optimization of binding affinity. Figure 6K shows data related to the drug likeness of the generated molecules.

[0038] Figures 6L-6M depict additional data related to the use of a generative model as provided for herein for ligand optimization. Figure 6L shows data related to the synthesizability of the generated molecules. Figure 6M shows data related to optimization of hydrophobicity.

[0039] Figures 6N-6O depict additional data related to the use of a generative model as provided for herein for ligand optimization. Figure 6N shows data related to the higher average diversity and lower molecular similarity to the reference (starting molecule). Figure 6O shows data related to the distribution of binding affinities that shifts away from that of the reference ligand and significantly widens.

[0040] Figure 7A schematically depicts various applications of generative models as provided for herein, such as substructure modification, linker design, fragment growing, and pattern replacement. In each instance circled regions (or otherwise unmarked) are targeted for modification while the other darkly shaded regions are maintained.

[0041] Figure 7B shows a non-limiting schematic of the prior-posterior sampling procedure.

[0042] Figures 7C-7D show data related to the various properties of the output molecules. Figure 7C shows data for each molecule modified at a target region with respect to predicted binding (*Vina*), hydrophobicity (*ALogP*), drug-likeness (*QED*), and topological polar surface area (*TPSA*).

[0043] Figures 8A-8B show information and data related to molecular generation of molecules based on AlphaFold-predicted structures. Figure 8A shows a graphical overview of the pipeline for creating the dataset for AlphaFold (AF) test structures. Figure 8B shows data related to a comparison of the AF-generated molecules versus crystal structure-based (PDB) generated molecules (left) and correlation of predicted *Vina* scores between molecules docked to AF and PDB structures (right).

[0044] Figures 8C-8D show additional information and data related to molecular generation of molecules based on AlphaFold-predicted structures. Figure 8C depicts data related to the evolutionary optimization of ligand bound to the BRD4 receptor (PDF ID 5D3H) using the AF-predicted receptor structure, in particular the distribution of *Vina* scores at each step of the optimization process. Figure 8D shows random examples of molecules generated for PDB ID 1CET using the PDB (lower) or AF (upper) receptor.

[0045] Figures 8E-8F show additional information and data related to molecular generation of molecules based on AlphaFold-predicted structures. Figure 8E depicts data related to the evolutionary optimization of ligand (S)-1-[N2-(1-carboxy-3-phenylpropyl)-L lysyl]-L-proline dihydrate bound to the human angiotensin-converting enzyme (ACE) receptor (PDB ID 1086) using the AF-predicted receptor structure, in particular the distribution of *Vina* scores at each step of the optimization process. Figure 8F shows random examples of molecules generated for PDB ID 1CET using the PDB (lower) or AF (upper) receptor.

[0046] Figure 9 shows Property distributions of generated molecules for different versions of the generative model provided for herein (pdb: model trained using PDBbind dataset from compared to training datasets. z: model trained using ZINC dataset.) compared to a random subsample of the ZINC dataset and the crystal ligands from the (PDBbind) test set. Molecules generated without a receptor are labeled with no rec. (mol. wt.: molecular weight, SA: synthetic accessibility, QED: drug-likeness, rot. bonds: rotatable bonds, HBD: H-bond Donors, HBA: H-bond Acceptors, TPSA: topological polar surface area, alogP: hydrophobicity)

[0047] Figures 10A-10B depict data related to the binding affinity of a molecule versus the number of atoms in the molecule. Figure 10A shows data correlating the number of molecules as a function of the number of atoms. Figure 10B shows scatterplots showing the correlation between predicted affinity (Vina) score and number of atoms in the ligand.

[0048] Figures 10C-10E depict additional data related to the binding affinity of a molecule versus the number of atoms in the molecule. Figure 10C shows boxplots showing the distributions of predicted affinity (Vina) score vs. number of atoms. Figure 10D is a table summarizing an evaluation of molecules generated for test set of 100 target receptors chosen from the PDBbind dataset. Values (mean ± std. dev.) are reported for competing models and the test set crystal ligands (Ref).

[0049] Figures 11A-11C show data related to force field optimization. Figure 11A shows relative change in binding affinity score (Δ Vina) as a function of random displacement for each of the ligands. Mean (black) and 95% confidence interval (grey shadow) also shown. Figure 10B shows distortion of molecular structure shown for example ligand (PDB ID: 2HHN) along with predicted affinity (Vina) score values. Figure 10C shows a plot of predicted affinity (Vina) score and internal energy (computed with MMFF94 force field) as a function of random displacement for example ligand (PDB ID: 2HHN).

[0050] Figures 12A-12C show data related to the latent representation and associated molecular properties, in particular molecular densities generated from an embodiment of the generative models provided for herein with varying temperature factor settings. Figure 12A shows randomly generated densities with all temperature factors set to zero ($t_i = 0, \forall_i$). Figure 12B shows randomly generated densities with all temperature factors set to zero except for the j^{th} latent resolution indicated on the right $(t_j = 1)$. Figure 12C shows schematics depicting the latent resolution for each row of generated densities.

[0051] Figures 13A-13B show data related to properties of ligands generated from AlphaFold vs PDB receptors. Distributions of properties for molecules generated for the AF test set receptors using the AF predicted structure (AF) and the PDB crystal structure (PDB). For comparison, property distributions also shown for the crystal ligands (Ref) along with medians of crystal ligand distributions (grey dashed line). (mol. wt.: molecular weight, SA: synthetic accessibility, QED: drug-likeness, rot. bonds: rotatable bonds, HBD: H-bond Donors, HBA: H-bond Acceptors, TPSA: topological polar surface area, alogP: hydrophobicity). Figure 13A shows data for molecular

weight, SA, QED, and rotational bonds. Figure 13B shows data for HBA, HBA, TPSA, and ALogP.

[0052] Figure 14A shows a random set of generated molecules targeted at PDB crystal receptors sampled from the prior of an embodiment of a generative model provided for herein.

[0053] Figure 14B shows a random set of generated molecules targeted at PDB crystal receptors sampled from the posterior (β=0) of an embodiment of a generative model provided for herein.

[0054] Figure 14C shows a random set of generated molecules targeted at PDB crystal receptors sampled from an embodiment of a generative model provided for here using prior-posterior sampling with interpolation factor $\beta = 0.5$, starting from the crystal ligand.

[0055] Figure 14D shows a random set of generated molecules with no receptor sampled from the prior of an embodiment of a generative model provided for herein.

[0056] Figure 15 depicts a schematic of a general architecture of a computing system that generates a plurality of molecular structures having a predictable binding affinity for a target receptor.

DETAILED DESCRIPTION

[0057] The detailed description of non-limiting embodiments makes reference to the accompanying drawings, which show non-limiting embodiments by way of illustration and their best mode. While these non-limiting embodiments are described in sufficient detail to enable those skilled in the art to practice the invention, it should be understood that other embodiments may be realized and that logical, chemical, and mechanical changes may be made without departing from the spirit and scope of the inventions. Thus, the detailed description is presented for purposes of illustration only and not of limitation. For example, unless otherwise noted, the steps recited in any of the method or process descriptions may be executed in any order and are not necessarily limited to the order presented. Furthermore, any reference to singular includes plural embodiments, and any reference to more than one component or step may include a singular embodiment or step. Also, any reference to attached, fixed, connected or the like may include permanent, removable, temporary, partial, full and/or any other possible attachment option. Additionally, any reference to without contact (or similar phrases) may also include reduced contact or minimal contact.

Definitions and Interpretations

[0058] Unless specifically stated or obvious from context, as used herein, the term "about" is understood as within a range of normal tolerance in the art, for example within 2 standard deviations of the mean. "About" can be understood as within 10%, 9%, 8%, 7%, 6%, 5%, 4%, 3%, 2%, 1%, 0.5%, 0.1%, 0.05%, or 0.01% of the stated value. Unless otherwise clear from context, all numerical values provided herein are modified by the term about.

[0059] As used in the specification and claims, the terms "comprises," "comprising," "containing," "having," and the like can have the meaning ascribed to them in U.S. patent law and can mean "includes," "including," and the like.

[0060] Unless specifically stated or obvious from context, the term "or," as used herein, is understood to be inclusive. Ranges provided herein are understood to be shorthand for all of the values within the range. For example, a range of 1 to 50 is understood to include any number, combination of numbers, or sub-range from the group consisting 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36. 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, or 50 (as well as fractions thereof unless the context clearly dictates otherwise).

[0061] As used herein, the term "fragment" shall be given its ordinary meaning and shall refer to a small molecule that typically consists of a low number of atoms (in various embodiments less than 10, 15, or 20 atoms, not including hydrogen) and represents a simplified version of a larger molecule. The term "fragment" can also refer to a contiguous portion of a ligand. In some contexts, a fragment may be below a size threshold, or a fragment may be defined as a part of a ligand that can include some of its chemical activity. As an example, "fragments" can be used in fragment-based drug discovery (FBDD) as a starting point to identify small molecules that can bind to a target protein or enzyme, and which can then be elaborated into larger and more potent drug-like compounds.

[0062] As used herein, the term "ligand" shall be given its ordinary meaning and shall refer to a molecule that binds to a specific target (e.g., molecule), typically a protein or enzyme. In many examples, a "ligand" (when bound to a target) can modulate the activity of a target (e.g., activation or inhibition of enzymatic activity, modulation of protein-protein interactions, or stabilization of protein conformation).

[0063] As used herein, "docking" shall be given its ordinary meaning and shall refer to computational simulation of a candidate ligand binding to a receptor.

[0064] As used herein, the term "molecular density" shall be given its ordinary meaning and, unless otherwise indicated, shall refer to the combination of atomic occupancy and atom feature (or atom-type) densities.

[0065] As used herein, the term "drug-like" shall be given its ordinary meaning and shall refer to any of the following, or a combination of any of the following: i) a compound that contains no more than 5 hydrogen bond donors, no more than 10 hydrogen bond acceptors, a molecular mass less than 600 g/mol, and a calculated octanol-water partition coefficient (C log P) that is 5 or less; ii) a compound that has a C log P between about -0.4 and about 5.6, a molecular refractivity from about 40 to about 130, a molecular mass between 180 and 600, and total number from about 20 to about 80; or iii) ten or less rotatable bonds and a polar surface area of 140 A^2 or less; iv) W logP less than 6 and polar surface area less than about 135 A^2; or v) a molecular mass between 200 and 600 g/mol, X log P between about -2 and about 5, polar surface area less than about 150 A^2, 7 or fewer rings, more than 4 carbons, more than 1 heteroatoms, less than 15 rotatable bonds, less than 10 hydrogen bond acceptors, and less than 5 hydrogen bond donors.

[0066] As used herein, a "rotatable bond" is any single bond, not in a ring, bound to a nonterminal heavy (e.g., non-hydrogen) atom and not including amide C-N bonds.

[0067] As used herein, the term "polar surface area" or "total polar surface area" (TPSA) refers to the area of a molecule's surface belonging to polar atoms, as described in, for example, Ertl, P. et al. J. Med. Chem. 2000, 43, 20, 3714-3717.

[0068] As used herein, the term "chemically stable" means a compound that does not further react, isomerize, or decompose under one or more of the following conditions: a) temperatures of 10 to 40 °C; b) a relative humidity of 30 to 100%; c) when formulated in a pharmaceutical composition containing one or more pharmaceutically acceptable excipients as described herein. In various embodiments, a chemically stable compound is stable for a period of one week to one year, or more. In various embodiments, a chemically stable compound is stable in the solid state and/or in a liquid formulation.

[0069] Success in the drug development process depends largely on the quality of initial candidate molecules identified in the early drug discovery process. Historically, a major obstacle to finding better candidates has been the limited size of drug-like molecular collections. Recently,

libraries of commercially available or readily synthesizable drug-like chemicals have grown into the billions of compounds. However, this represents only a miniscule fraction of the total drug-like chemical space - estimated to be as large as 10^{60} compounds. Similarly, structure based Virtual Ligand Screening (VLS) approaches have historically been limited to well below 10^{7} compounds. With the advent of virtually enumerated libraries and fragment based virtual screening techniques, this capacity has increased into the billions, possibly extending to the tera-scale in the future. Despite this increase, these techniques are fundamentally limited by the bruteforce nature of their search through chemical space and may not be trivially scalable even to a significant fraction of the almost infinite drug-like space. Further, their success hinges on the availability of high-resolution crystal structure information for each target and the speed and accuracy of virtual docking algorithms which are known to be limited. Other critical steps in the early drug development process are limited due to their manual nature, such as the systematic modification of candidates during hit-to-lead and lead optimization. Although progress has been made in leveraging computational tools using structure-based drug design (SBDD), these costly and time-consuming tasks still rely significantly on, and are therefore limited by, the expertise and intuition of working scientists. Overcoming such limitations and expanding the reach of computational drug design to greater scales of drug-like chemical space will require new approaches, such as the models provided for herein.

[0070] Recent advances in artificial intelligence and generative modeling, which have yielded impressive results on a number of previously intractable problems in the biological sciences, provide a promising new approach. Rather than further optimizing the exhaustive search of chemical space, deep learning models are capable of learning distributions over large molecular datasets and have already been successfully applied to large drug-like chemical datasets for de novo molecular generation. More recently, efforts have turned toward modeling drug-target interactions by training models on datasets of ligand-bound receptor crystal structures. By learning the joint probability distribution of ligands and their receptors, these models can generate new molecules conditioned on the receptor structure, effectively narrowing the regions of chemical space in which to search for new drugs. A major benefit of the generative modeling approach is the capacity to generalize the drug-target interactions learned over the dataset to new receptor structures. This is in stark contrast to the classical VLS approach in which the significant cost and computational effort of screening large libraries must be repeated with each new target.

[0071] Provided for herein is a deep structure-based Drug generating HIerarchical Variational autoencoder (referred to as DrugHIVE in some instances). Unlike previous encoder-decoder models, embodiments provided for herein employ a hierarchical latent structure (prior) which represents more naturally the distribution of molecular structure space. The use of a hierarchical prior leads to the encoding of input molecular structures at varying spatial scales and enables the generation of molecules with a high level of spatial control - a necessity for many drug design tasks.

[0072] Owing to its hierarchical latent structure, models provided for herein are capable of generating molecules in myriad ways and achieves state of the art in predicted binding affinity of de novo generated molecules conditioned on receptor targets. Several non-limiting examples of common drug design tasks are demonstrated to be accomplished by models provided for herein. For example, an embodiment of a model provided for herein was used to generate a vastly improved ability to optimize the drug-like properties and predicted affinity of molecules that are approved by the FDA. Applications demonstrated herein also include the automated replacement of Pan-Assay Interference Compounds (PAINS) patterns, the linking of fragments from a fragment screening experiment, substructure optimization (scaffold hopping), and the growing of molecular structures (fragment growing). Models provided for herein extend existing limits of generative molecular design beyond the currently available crystal structures by successfully generating and optimizing ligands with improved predicted binding affinity for receptors using only predicted receptor structures.

General Computing System Architecture

[0073] A system, apparatus and/or method for generating a plurality of molecular structures having a predictable binding affinity for a target receptor, where a processor (e.g., an autoencoder) is disclosed herein to perform the AI and machine learning algorithms and is further described as set forth below.

[0074] Referring now to Figure 1, a system 100 for generating a plurality of molecular structures having a predictable binding affinity for a target receptor is disclosed herein. The system 100 (e.g., a computing system) may include a computing apparatus 102. The computing apparatus 102 may include one or more processors 104, one or more memories 106 and/or one or more buses 112 and/or other mechanisms for communicating between the one or more processors 104. The

system 100 may be a cloud computing system including processors, servers, storage, databases, networking, software, analytics, and/or intelligence accessed or performed over or using the Internet ("the cloud"). The one or more processors 104 may be implemented as a single processor or as multiple processors. The one or more processors 104 may execute instructions stored in the memory 106 to implement the applications and/or detection of the system 100.

[0075] The one or more processors 104 may be coupled to the memory 106. The memory 106 may include one or more of a Random Access Memory (RAM) or other volatile or non-volatile memory. The memory 106 may be a non-transitory memory or a data storage device, such as a hard disk drive, a solid-state disk drive, a hybrid disk drive, or other appropriate data storage, and may further store machine-readable instructions, which may be loaded and executed by the one or more processors 104.

[0076] The memory 106 may include one or more of random-access memory ("RAM"), static memory, cache, flash memory and any other suitable type of storage device or computer readable storage medium, which is used for storing instructions to be executed by the one or more processors 104. The storage device or the computer readable storage medium may be a read only memory ("ROM"), flash memory, and/or memory card, that may be coupled to a bus 112 or other communication mechanism. The storage device may be a mass storage device, such as a magnetic disk, optical disk, and/or flash disk that may be directly or indirectly, temporarily, or semi-permanently coupled to the bus 112 or other communication mechanism and be electrically coupled to some or all the other components within the system 100 including the memory 106, the user interface 110 and/or the communications interface 108 via the bus 112.

[0077] The term "computer-readable medium" is used to define any medium that can store and provide instructions and other data to a processor, particularly where the instructions are to be executed by a processor and/or other peripheral of the processing system. Such medium can include non-volatile storage, volatile storage, and transmission media. Non-volatile storage may be embodied on media such as optical or magnetic disks. Storage may be provided locally and in physical proximity to a processor or remotely, typically by use of network connection. Non-volatile storage may be removable from computing system, as in storage or memory cards or sticks that can be easily connected or disconnected from a computer using a standard interface.

[0078] The system 100 may include a user interface 110. The user interface 110 may include an input/output device. The input/output device may receive user input, such as a user interface

element, hand-held controller that provides tactile/proprioceptive feedback, a button, a dial, a microphone, a keyboard, or a touch screen, and/or provides output, such as a display, a speaker, an audio and/or visual indicator, or a refreshable braille display. The display may be a computer display, a tablet display, a mobile phone display, an augmented reality display or a virtual reality headset. The display may output or provide molecular structures or molecular substructures.

[0079] The user interface 110 may include an input/output device that receives user input, such as a user interface element, a button, a dial, a microphone, a keyboard, or a touch screen, and/or provides output, such as a display, a speaker, headphones, an audio and/or visual indicator, a device that provides tactile/proprioceptive feedback or a refreshable braille display. The speaker may be used to output audio associated with the audio conference and/or the video conference. The user interface 110 may receive user input that may include configuration settings for one or more user preferences, such as a selection of joining an audio conference or a video conference when both options are available, for example.

[0080] The system 100 may have a network 116 connected to a server 114. The network 116 may be a local area network (LAN), a wide area network (WAN), a cellular network, the Internet, or combination thereof, that connects, couples and/or otherwise communicates between the various components of the system 100 with the server 114. The server 114 may be a remote computing device or system that includes a memory, a processor and/or a network access device coupled together via a bus. The server 114 may be a computer in a network that is used to provide services, such as accessing files or sharing peripherals, to other computers in the network.

[0081] The system 100 may include a communications interface 108, such as a network access device. The communications interface 108 may include a communication port or channel, such as one or more of a Dedicated Short-Range Communication (DSRC) unit, a Wi-Fi unit, a Bluetooth® unit, a radio frequency identification (RFID) tag or reader, or a cellular network unit for accessing a cellular network (such as 3G, 4G or 5G). The communication interface may transmit data to and receive data from the different components.

[0082] The server 114 may include a database. A database is any collection of pieces of information that is organized for search and retrieval, such as by a computer, and the database may be organized in tables, schemas, queries, reports, or any other data structures. A database may use any number of database management systems. The information may include real-time information, periodically updated information, or user-inputted information.

[0083] In various embodiments, the computing apparatus 102 can include a generative artificial intelligence ("AI") module 122. The generative AI module 122 can include the one or more processors 104. Stated another way, the generative AI module 122 can be run, or operated by the one or more processors 104. In various embodiments, the generative AI module 122 can perform the steps of the methods claimed herein and output or provide molecular structures or molecular substructures to the user interface 110.

[0084] In various embodiments, the generative AI module 122 is configured to create a molecular density grid from a database 124 as described further herein. In this regard, the generative AI module 122 is configured to update a generative AI model based on the molecular structures or molecular substructures and data received from the database 124. In various embodiments, the molecular structures or molecular substructures of the generative AI module 122, as described further herein can include the methods described herein.

[0085] Referring now to Figure 15, Figure 15 depicts a general architecture of a computing system (referenced as computing device 1500) that generates a plurality of molecular structures having a predictable binding affinity for a target receptor. The computing device 1500 may be an example of the computing apparatus 102 of Figure 1. The general architecture of the computing device 1500 depicted in Figure 15 includes an arrangement of computer hardware and software modules that may be used to implement aspects of the present disclosure. The hardware modules may be implemented with physical electronic devices. The computing device 1500 may include many more (or fewer) elements than those shown in Figure 15. Additionally, the general architecture illustrated in Figure 15 may be used to implement one or more of the other components illustrated in Figure 1. As illustrated, the computing device 1500 may include one or more processor(s) 1502, a network interface 1504, a computer readable medium drive 1506, and an input/output device interface 1508. The network interface 1504 may provide connectivity to one or more networks or computing systems. The processor(s) 1502 may thus receive information and instructions from other computing systems or services via a network, such as the network 116 of Figure 1. The processor(s) 1502 may also communicate to and from memory 1510. The processor(s) 1502 may further provide output information for an optional display (not shown) using the input/output device interface 1508. The input/output device interface 1508 may also accept input from an optional input device (not shown). The one or more processors 1502 may be implemented as a single processor or as multiple processors. The one or more processors 1502

may execute instructions stored in the memory 1510 to implement the applications and/or detection of the computing system.

[0086] The memory 1510 may contain computer program instructions that the processor(s) 1502 executes in order to implement one or more aspects of the present disclosure. The memory 1510 may include one or more of a Random Access Memory (RAM) or other volatile or non-volatile memory. The memory 1510 may be a non-transitory memory or a data storage device, such as a hard disk drive, a solid-state disk drive, a hybrid disk drive, or other appropriate data storage, and may further store machine-readable instructions, which may be loaded and executed by the one or more processors 1502.

[0087] The memory 1510 may include one or more of random-access memory ("RAM"), static memory, cache, flash memory and any other suitable type of storage device or computer readable storage medium, which is used for storing instructions to be executed by the one or more processors 1502. The memory 1510 may store an operating system 1514. The operating system 1514 can provide computer program instructions for use by the processor(s) 1502 in the general administration and operation of the computing device 1500. The memory 1510 may further include computer program instructions and other information for implementing aspects of the present disclosure. For example, in one embodiment, the memory 1510 includes a user interface unit 1512 that generates user interfaces (and/or instructions therefor) for display upon a computing device and an operating system 1514. As an example, the user interface unit 1512 may generate user interfaces via a navigation and/or browsing interface such as a browser or application installed on the computing device 1500. In addition, the memory 1510 may include and/or communicate with one or more data repositories, such as database 1516, for example, to access user program codes and/or libraries.

[0088] In addition to and/or in combination with the user interface unit 1512 and operating system 1514, the memory 1510 may include a generative AI module 122, such as the generative AI module 122 as shown and described in Figure 1, that may be executed by the processor(s) 1502. In one embodiment, the generative AI module 122 implements various aspects of the present disclosure. For example, the generative AI module 122 can represent code executable to perform the steps of the methods claimed herein and output molecular structures or molecular substructures.

Generative Models, Methods and Applications Thereof

[0089] According to the present disclosure, a system to design molecules (e.g., to treat a patient with a disease) is provided herein. In certain embodiments, the generated molecules are ligands, relatively small (e.g., generally <500 atomic molecular units) chemical compounds. Generally, such small molecule ligands treat a disease by binding with a specific protein target to inhibit, enhance, or otherwise alter its function in patients.

[0090] According to several embodiments, a computing system, such as that shown in Figure 1 and/or Figure 15 is used to receive an input dataset, operate a generative model and output new molecules. As shown in Figure 2A an input dataset can comprise a plurality of molecules. In some embodiments, the plurality of molecules is a collection of known ligands of a receptor(s). In some embodiments, the plurality of molecules is a collection of molecules of the same drug class and/or designed to treat a common disease or symptom. As discussed in more detail herein, in several embodiments, the models and methods provided herein maintain as a constant one or more regions of an input molecule, and generate variations at one or more specific regions of the input molecule, thereby allowing output of new molecules that can be screened for one or more desired characteristics, such as binding affinity, drug-like qualities, hydrophobicity or hydrophilicity, and/or ease of synthesizability. Similarly, Figure 2A shows a schematic input dataset that is a drug-target (e.g., ligand-receptor) complex. Depending on the embodiment, such input data can be based on a virtual docking of the ligand to a receptor, or an actual crystal structure of a ligand docked with its receptor. Moreover, in several embodiments, the input dataset can comprise a receptor without a docked ligand. In some embodiments, the input dataset comprises crystal structure data of a receptor. In some embodiments, the input dataset comprises a predicted structure (e.g., predicted crystal structure) of a receptor.

[0091] Figure 3A shows a schematic depicting a general framework for use and operation of the presently disclosed models. Input data (e.g., ligands, receptors (actual or predicted), or a combination thereof) is used as training data for the system. The training data is used to train the model and then the model is tested. If the training is successful (e.g., new viable molecules are output) then the model parameters are saved. If the training is not successful, the model build and training is refined and re-tested, until such time that the training is deemed successful. In several embodiments, the methods further comprise adding to or otherwise refining the training data. For example, in several embodiments the output data (new molecules) is added to the training data (in

whole or in part) to further iterate and refine the learned features within the model, leading to enhancement of the desired characteristics of the new molecules.

[0092] Figures 3B-3G depict schematic flow diagrams for various applications that the generative models disclosed herein can be used to accomplish. An overview of these is provided, followed by a more detailed description of the architecture of the model.

[0093] Figure 3B depicts a schematic flow diagram for use of the models provided herein for the generation of new molecules (e.g., those not within the training/input dataset). In this non-limiting embodiment, input data comprises one or more of protein receptor structure (based on crystal data and/or predicted structure) and coordinates to the center of the binding pocket (e.g., the localization of where the ligand binds). Samples are drawn from the latent distribution and decoded into a new atomic density. A molecule is subsequently fit to the generated density. The resultant latent representation is decoded into a new set of atomic densities and a new molecule is fit to the new set of atomic densities.

[0094] Figure 3C depicts a schematic flow diagram for use of the models provided herein for the prior-posterior sampling. As discussed below, prior sampling (generating new structured by inputting (1) a protein structure into the protein encoder and (2) a randomly sampled set of values from the prior into the ligand decoder) results in a randomly generated density that can then be fit with a molecular structure. This is combined with posterior sampling (randomly sampling from a previously encoded ligand structure) to generate densities of varying similarity to an initial ligand density. Input data in this context comprises protein receptor structure (based on crystal data and/or predicted structure), ligand structure, interpolation factors and interpolation patterns. The ligand-protein complex (either individually or when the ligand is docked) is encoded into the latent representation. Random samples are pulled from the latent distribution and interpolation is performed (based on the interpolation factors and pattern) between the encoded ligand-protein complex and the latent sample. The resultant latent representation is then decoded into a new set of atomic densities and a new molecule is fit to the new density.

[0095] Figure 3D depicts a schematic flow diagram for use of the models provided herein for the substructure modification. In this non-limiting embodiment, the input data comprises a protein receptor structure (based on crystal data and/or predicted structure), ligand structure, and the identification of one or more substructures to modify. Based on the ligand structure and the one or more substructures to modify, an interpolation pattern is generated. A new molecule is

generated through the performance of prior-posterior sampling (Figure 3C) with the interpolation pattern determined by which substructure(s) are chosen to be modified.

[0096] Figure 3E depicts a schematic flow diagram for use of the models provided herein for pattern replacement (exchange of a core or pendant structure with another structure having a similar atomic pattern). In this non-limiting embodiment, the input data comprises protein receptor structure (based on crystal data and/or predicted structure), ligand structure, and identification of the molecular pattern(s) or fragment(s) to be replaced. The ligand is scanned, based on the molecular pattern(s) or fragment(s), to identify the substructure of the ligand to me modified. A new molecule is generated by performing substructure modification sampling (Figure 3D).

[0097] Figure 3F depicts a schematic flow diagram for use of the models provided herein for linker design. In this non-limiting example, the input data comprises protein receptor structure (based on crystal data and/or predicted structure) and fragment structures (e.g., portions of a ligand that will bind to the receptor). An interpolation pattern is generated that encompasses the grid region between two or more fragment structures to be joined. A new molecule is then generated by performing prior-posterior sampling (Figure 3C) with the interpolation pattern being determined by the substructure to be modified.

[0098] Figure 3G depicts a schematic flow diagram for use of the models provided herein for drug property optimization. In this non-limiting example, the input data comprises protein receptor structure (based on crystal data and/or predicted structure) and fragment structures. In an initialization step, an initial population of N_c molecules are generated. A fitness function is applied, and the candidate molecules are clustered based on, for example, Taniomoto similarity, with the most fit individual from each cluster being kept (the N_p molecules). In an optimization step, a new generation of N_c "child" molecules are generated from the N_p molecules. The fitness function is applied, and the molecules are clustered, with the most fit individual molecules being kept from each cluster. In this process, any parent molecules that are less fit than a child molecule will be replaced by such a child molecule. An inquiry is made to determine if the parental (N_p) and child (N_c) populations have converged. If they have not, the optimization step is repeated. If they have, then the drug optimization process is complete, and the final population of optimized molecules is returned.

Data Sources

[0099] Turning to Figure 4, a schematic process flow and overlay of the architecture of generative models as provided for herein, Figure 4A shows non-limiting examples of input data. Input data is also referred to herein as training data, as it is used to train the generative models provided for herein. In some embodiments, the input data comprises output data from a prior learning cycle performed by the generative model. As discussed herein, in several embodiments, data, such as the data relating to a target protein or receptor of interest, is experimentally derived. For example, in some embodiments, the input data is derived from x-ray crystallographic structures of a protein, or a portion of interest of a protein. However, as discussed herein, the scope of the human proteome is far larger than the pool of proteins that have been crystallized and have had their structures empirically determined. Thus, in several embodiments, the protein or receptor input data can be based on predicted structure. Structure can be predicted by homology modeling, threading and fold recognition, ab initio structure prediction, secondary structure prediction, or any combination thereof. Homology modeling can be accomplished using, for example, IntFOLD, RaptorX, Biskit, ESyPred3D, FoldX, Phyre, or Phyre2, HHpredd, MODELLER, CONFOLD, Molecular Operating Environment (MOE), Robetta, BHAGEERATH-H, Swiss-model, Yasara, AWSEM-Suite. Threading and fold recognition can be accomplished using, for example, IntFOLD, RaptorX, HHpred, Phyre or Phyre2, or I-TASSER. Ab initio structure prediction can be accomplished using, for example, trRosetta, ROBETTA, Rosetta@home, Abalon, or C-Quark. Protein secondary structure prediction programs include, for example, RaptorX-SS8, GOR, Jpred, PredictProtein, PSIPRED. In several embodiments, AlphaFold is used to predict protein structure. In several embodiments, ligand structure is derived from the chemical formula of the ligand. In several embodiments, chemical structures can be input by sketching and querying a database, such as the RCSB PDB database, and output regarding the ligand structure (such as SMILES or International chemical Identifier (InChI) identity is generated.

Hierarchical Prior

[00100] Figure 4B shows a structural aspect of the model, a hierarchical structure, with varying spatial scales utilized. Input molecules are first converted to molecular density grids, where each atom is represented by a multi-channel pseudo-gaussian density with channels representing different atomic features. For each atom, an identical atomic density is placed in each grid channel

corresponding to that atom's features. In several embodiments, a total of 15 feature channels are used. In this non-limiting embodiment, they are accounting for 9 elements (C, N, O, F, P, S, Cl, Br, I) and 6 other atomic features (H-bond donor, H-bond acceptor, aromaticity, positive charge, neutral charge, negative charge). Various levels of resolution can be used, depending on the embodiment. As non-limiting embodiments, the grids shown in Figure 4B range from most detailed at 1 angstrom (Å) at grid 1, 2 Å at grid 2, 4 Å at grid 3, and 8 Å at grid 4. Greater or lesser numbers of grids can be used in the hierarchy, including 2, 3, 4, 5, 6, 7, 8, 9, 10 or more grids. Greater or lesser levels of resolution can be used as well, as computing power and computational resources allow for. In several embodiments, the resolution for a given grid ranges from about 0.1 to about 0.25 Å, about 0.25 to about 0.5 Å, about 0.5 to about 0.75 Å, about 0.75 to about 1 Å, about 1 to about 2 Å, about 2 to about 3 Å, about 3 to about 4 Å, about 4 to about 5 Å, about 5 to about 6 Å, about 6 to about 7 Å, about 7 to about 8 Å, about 8 to about 9 Å, about 9 to about 0 Å, or any value between those listed.

Ligand Encoder, Protein Encoder, and Ligand Decoder

[00101] The molecular grid of the hierarchical prior is input into the network via two pathways (see Figure 4C). A single-channel full resolution grid, representing atomic occupancy, is input into each of the ligand encoder and the protein encoder (see Figure 4D). All other atomic feature channels are downsampled by a factor of two and input into the network after the first downsampling layer. Other downsampling factors may be used in other embodiments, such as a downsampling factor of 3, 4, 5, 6, 7, 8, 9, 10 or greater. The atomic occupancy and atom feature data occupies the latent space representation from which random samples are pulled, depending on the embodiment. For known ligands, random samples are pulled from one or more grids of the posterior, depending on the embodiment. The protein encoder feeds into both the ligand encoder (to provide latent space representations of the protein and the ligand) and the ligand decoder (to determine the plausible fit of the newly decoded ligand to the protein of interest). Model output from the ligand decoder mirrors this input dual resolution arrangement, with atomic occupancy output at full resolution and all other atomic features being upsampled in an inverse fashion from their respective downsampling factor. Thus, upsampling can also occur at various factors, including, but not limited to a factor of 2, 3, 4, 5, 6, 7, 8, 9, 10 or greater. Advantageously, by segregating occupancy from atomic features in this fashion, atom position information is

maintained while significantly decreasing grid sparsity and memory load. Figure 4E shows output from the model, atomic occupancy and atom-type grids, which are fit with a molecular structure. The atomic occupancy data is the first level of data decoded by the ligand decoder, whereas the additional atomic characteristics are integrated by type, and bonds are subsequently fits. Finally, the optimized new molecule(s) that now comprises chemical structure fit within the generated atomic density data is output. Non-limiting examples of various atomic properties that can be used confirm the fit of a molecule include, but are not limited to, the element identity (and possible steric interaction with other nearby elements), the element charge (and how that charge interacts with the charge of surrounding elements), the element's tendency to be a hydrogen bond acceptor or a hydrogen bond donor, and/or the aromaticity of the element. Other atomic characteristics may be added into the model in addition to, or in place of, those listed.

[00102] To reiterate the advantageous nature of the multiple resolution data input and output, Figure 4G shows that the full resolution used on the atomic occupancy data (e.g., 1 x 48^3 data units) coupled with the downsampling of the other atomic characteristics (e.g., from C x 48^3 to C X 24^3) leads to an ~8X reduction in GPU memory and ~50% reduction in grid sparsity. Thus, generative models as provided for herein result in robust data output while the computing power and resources needed to generate that output are reduced.

Applications

[00103] Models as provided for herein can be used in a variety of applications. See, for example the methodology of Figures 3A-3G. In more detail, Figures 5A-5E graphically depict non-limiting examples of such application.

Substructure Modification

[00104] Figure 5A shows a non-limiting schematic of substructure modification (also referred to as scaffold hopping. In this approach, a portion (or portions) of a starting molecule is identified as having their structure retained. Likewise, one or more portions are identified as a substructure (e.g., a region not encompassing the entire molecule) to be modified. As shown in 5A, the region to be modified can be an external portion of the molecule, or an internal region. The various chemical structures below the boxes represent non-limiting examples of output molecules from embodiments of the model provided for herein, as evidenced by the shaded (retained) regions and

the different structures in the unshaded regions. This application allows iterative changes to be made to an existing group of a starting molecule, and once fitted to a new molecular structure based on the atomic density output of the model, molecules with desired characteristics (e.g., enhanced binding affinity, etc.) can be identified.

Linker Design

[00105] Figure 5B shows a non-limiting schematic of linker design. Many drugs have two or more substructures that are linked by an intermediate molecular structure. In this approach, new intermediate structures (e.g., linkers) are generated by the model based on identification of the regions to be retained and those to be modified. Linker design can advantageously be used to improve the starting drug by, for example, reducing potential steric hindrances that exist (e.g., original linker too short), or improving synergy between two substructures (e.g., original linker too long). Again, the molecular structures below the box represent non-limiting examples of new molecules that can be generated with improved functionality.

Fragment Growing

[00106] Figure 5C shows a non-limiting schematic of fragment growing. Fragment-based drug design is initiated by using fragments, typically low molecular weight compounds with low affinity for a target and adding fragments that conform to various design constraints (e.g., size of binding pocket, etc.) and assessing the affinity. In conjunction with the present models, a starting molecule can have a region designated as constant (e.g., structure to be retained) and new fragments can be added to the constant portion in an iterative fashion, with only those exhibiting increased affinity (or other characteristic of choice) being kept in the candidate pool. According to some embodiments, the compounds that show no change, or decreased affinity are recycled into the training data, to further enhance the ability of the models provided for to learn patterns of molecular structure that impact affinity (or other characteristic of choice) and output candidates that show further gains in affinity (or other characteristic of choice).

Pattern Recognition

[00107] Figure 5D shows a non-limiting schematic of pattern recognition. One non-limiting application of pattern recognition is for identifying pan-assay interference compounds (PAINS compounds). PAINS compounds are those that result in false positives in the drug-discovery process. These compounds are biologically active compounds that are erroneously identified as potential drug candidates during initial high-volume screening used to search for possible new drugs. The false positives result from their ability to disrupt the assay technology used in the screening to report biological activity, but these compounds are not active against the intended biological target. In Figure 5D, a molecular pattern of interest serves as the query, as it is responsible for the false positive. That pattern can then be identified as a region to be modified (solid circles) and the other regions not matching the patter to be held constant (dashed circles). The models provided for herein can then use the queried pattern to generate new structures not matching the pattern in question, and thereby output molecular structures with biological activity against the target of interest (rather than just in the screening assay). As an example, Figure 5E shows data related to the resultant Vina (affinity) scores for the generated molecules with the initial ligand score shown as the black dashed line. Of the 200 molecules with the PAINS pattern replaced, 28% of generated molecules exhibited an improved binding affinity score. This pattern replacement capability is scalable to large collections of molecules, as the run time per molecule is on par with fast virtual docking algorithms.

EXAMPLES

[00108] The materials and methods disclosed herein are non-limiting examples that are employed according to certain embodiments disclosed herein. Additional experimental details can be found in Weller et. al., Structure-Based Drug Design with a Deep Hierarchical Generative Model, J. Chem. Inf. Model., V. 64, pgs.: 6450-6463 (2024), the entire contents of which is incorporated by reference herein.

Methods

[00109] According to several embodiments, models for new molecule generation are based on hierarchical variational autoencoder (HVAE) architecture (see Figure 4D), a deep learning implementation of hierarchical Bayesian methods. As with standard variational autoencoders

(VAEs), the objective is to learn a reversible encoding of input data into a latent space usually represented by a multidimensional normal distribution. Whereas a standard VAE has a single latent representation, a HVAE, such as provided for herein, has multiple levels with dependency between lower- and higher-level encodings. This structure, as provided for in the embodiments disclosed herein, is well-suited for data with inherent hierarchical relationships, such as molecular structures. In a successfully trained model, novel examples (e.g., molecular structures) can be generated by drawing and decoding random samples from this latent distribution. According to several embodiments, generative sampling produces new high-quality examples that are unobserved in the training data set.

[00110] In molecules, properties arise at different spatial scales, e.g., from single atoms to functional groups and moieties to larger structural motifs. Although standard VAEs can successfully learn distributions over molecular data, according to several embodiments, hierarchical models such as those disclosed herein advantageously more naturally represent the inherent hierarchical spatial relationships. As explained in more detail below, models as provided for herein allow for spatial-scale organization of the latent space. For example, attributes such as large-scale geometry are encoded at higher latent levels, whereas more fine-grained attributes such as atomic properties are primarily encoded a lower level (see, e.g., Figures 12A-12C). Note that the non-limiting models disclosed herein represent assessment of the spatial hierarchy of molecules in chemistry, though other attributes can be evaluated, in several embodiments, such as based on the hierarchical categorization of molecules based on therapeutic or natural product classes.

Model

[00111] A non-limiting schematic of the machine learning model is illustrated in Figures 4A-4G. Input molecules are first converted to molecular density grids, where each atom is represented by a multi-channel pseudo-gaussian density with channels representing different atomic features. For each atom, an identical atomic density is placed in each grid channel corresponding to that atom's features (Figure 4B). In several embodiments, a total of 15 feature channels are used. In this non-limiting embodiment, they are accounting for 9 elements (C, N, O, F, P, S, Cl, Br, I) and 6 other atomic features (H-bond donor, H-bond acceptor, aromaticity, positive charge, neutral charge, negative charge). The molecular grid is input into the network via two pathways (see

Figure 4C). A single-channel full resolution grid, representing atomic occupancy, is input into the top of the network. All other atomic feature channels are downsampled by a factor of two and input into the network after the first downsampling layer (other downsampling factors may be used in other embodiments). Model outputs mirror this arrangement. Advantageously, by segregating occupancy from atomic features in this fashion, atom position information is maintained while significantly decreasing grid sparsity and memory load.

[00112] As discussed herein, in several embodiments, the models provided for herein are based on HVAE architecture, which has been shown to outperform autoregressive models on image generation tasks. The goal is to learn the distribution of ligand density conditioned on receptor density from the training dataset. However, unlike for a standard VAE, the latent space of an HVAE retains spatial context (Figure 4B). Therefore, there is provided a generative model of the form $p(D_{\text{lig}}, D_{\text{rec}}, \mathbf{z}) = p(\mathbf{z})p(D_{\text{lig}} \mid \mathbf{z}, D_{\text{rec}})$, where $p(D_{\text{lig}} \mid \mathbf{z}, D_{\text{rec}})$ is the likelihood function (decoder), the prior is represented as $p(\mathbf{z}) = \prod_i p(z_i \mid z_{<i})$ and the approximate posterior as $q(\mathbf{z} \mid D_{\text{lig}}, D_{rec}) = \prod_i q(z_i \mid z_{<i}, D_{\text{lig}}, D_{rec})$. The likelihood function and approximate posterior as neural networks are represented by p_θ and q_ϕ, respectively. The latent representation is a set of variables partitioned into a set of disjoint groups $\mathbf{z} = \{z_1, \ldots, z_n\}$ where each group is conditioned on the lower groups in the hierarchy, $z_i \sim q(z_i \mid D_{\text{lig}}, D_{rec}, z_{<i})$. Because the individual latent variables retain spatial context at varying resolutions (as shown in Figure 4B), it is believed that the structure of the latent space is a better reflection of molecular space and therefore allows embodiments of the models provided for herein to generate novel molecular compounds that have desired attributes, such as affinity for a particular receptor that is greater than that of a known ligand that is used to train the model.

Pseudo-Gaussian Atom Representation

[00113] Atoms are represented as pseudo-Gaussian densities by first encoding them to a grid, using what is referred to herein as center-of-mass (COM) encoding, followed by a Gaussian convolution. Each atomic coordinate $c \in \mathbb{R}3c \in R3$ is considered to represent the center of a cube with the same dimensions as a single voxel. Then, to encode an atom to the grid, the proportion of the cube that is contained in each grid voxel is calculated and that value is added to the corresponding voxel value. Therefore, to acquire the COM encoding of a set of N atoms with grid

coordinates $c_j \in \{c_1, \cdots, c_N\}$ on a grid with spacing $a = 1$, the contribution of each atom to each grid voxel $g_i \in G_{COM}$ is summed:

$$g_i = \sum_j \begin{cases} \left\| (1 - d_{i,j}) \right\|^2 & \text{if} \quad \max(d_{i,j}) < 1 \\ 0 & \text{if} \quad \max(d_{i,j}) \geq 1 \end{cases}$$

[00114] where c_{gi} are the grid coordinates of the i^{th} voxel, c_j are the grid coordinates of the jth atom, and $d_{i,j} = |c_{gi} - c_j|$. Importantly, precise positional information is retained and remains recoverable after the encoding process. Given a set of adjacent grid points containing an encoded atom $g_i \in G_{cj}$, the original coordinate c_j can be exactly recovered using the COM equation,

$$c_j = \sum_i g_i \cdot c_{g_i}$$

[00115] where the value of g_i represents the mass of each grid point. This is in contrast to the simple one-hot encoding scheme, where each atom is represented by a single voxel and positional precision is lost.

[00116] Next, a convolution of the COM grid is performed with $K \in \mathbb{R}M3K \in RM3$, a Gaussian kernel with side length M, to obtain the density grid $D = K \times G_{COM}$. The value at each kernel point $k_i \in K$ is calculated as,

$$k_i = \begin{cases} e^{-|r_i|^2/2\sigma^2} & \text{if} \quad r \leq r_c \\ 0 & \text{if} \quad r > r_c \end{cases}$$

[00117] where r_i is the distance of the grid point from the grid center and r_c is a cutoff radius. The resulting density grid approximates an exact Gaussian encoding of each atom and can be summarized as the Gaussian blurring or diffusion of the initial COM encoding pattern. The advantage of this procedure is that the initial COM encoding is very fast to compute, and the Gaussian convolution operation can be represented as a standard convolutional neural network layer with a fixed Gaussian kernel. The choice of convolutional kernel is customizable, and the

whole process can be easily and efficiently implemented using common Python scientific and machine learning libraries.

Fitting Molecules to Densities

[00118] The output of the models provided for herein is a molecular density grid which must then be mapped to a molecular structure (Figure 4E). To achieve this, custom atom-fitting and bond-fitting algorithms are utilized. The atom-fitting algorithm first iteratively places atomic coordinates within the grid with the goal of minimizing the residual density,

$$C_{fit} = \arg\min_{C} ||D_{occ,gen} - d(C)||^2 \qquad (cfit)$$

[00119] where $D_{occ,gen}$ is the predicted atomic occupancy density and d is the mapping from coordinates to density grid D. Next, atomic features are assigned to each atom using the density in the atomic feature grid. Atomic number is assigned to the highest density value across all atomic number channels at each atom location. The remaining atomic features are assigned based on a density threshold value, with disputes between conflicting features (e.g., formal charge ± 1) resolved by selecting the feature with the highest density value. Next, bonds are added based on the geometry and features of the atom set to create a molecular structure.

Training

[00120] The model is trained to minimize the objective,

$$\mathcal{L} = \alpha_{\text{recon}}\mathcal{L}_{\text{recon}} + \alpha_{KL}\mathcal{L}_{KL} + \alpha_{\text{steric}}\mathcal{L}_{\text{steric}} \qquad (loss_full)$$

[00121] where $\mathcal{L}_{\text{recon}}$ is the reconstruction loss, $\mathcal{L}_{KL}$ is the Kullback-Leibler (KL) divergence between the prior and posterior distributions, $\mathcal{L}_{\text{steric}}$ is the loss for steric clash between the ligand and receptor density, and each α is a scaling term.

$$\mathcal{L}_{KL} = -D_{KL}\left(q_{\phi}(\mathbf{z} \mid D_{\text{lig}}, D_{rec}) \| p(\mathbf{z})\right) \qquad (loss_kl)$$

$$\mathcal{L}_{\text{recon}} = -\log p_{\theta}\left(D_{\text{lig}} \mid D_{\text{rec}}, \mathbf{z}\right) \qquad (loss_kl)$$

$$\mathcal{L}_{\text{steric}} = \log p_{\theta}(D_{\text{rec}} \mid D_{rec}, \mathbf{z}) \qquad (loss_steric)$$

(loss_recon)

[00122] Because the grid representation is sparse, a mask is applied to scale the reconstruction loss for zero grid values in order focus the loss signal on non-empty grid locations. This is achieved by scaling the loss for zero grid values by a factor

$$\gamma_i = \begin{cases} \gamma & \text{if } D_{\text{lig},i} = 0 \\ 1 & \text{otherwise} \end{cases}$$

[00123] γ, yielding the scaled reconstruction loss,

$$\mathcal{L}_{\text{recon}} = \sum_i -\gamma_i \log p_\theta(D_{\text{lig},i} \mid D_{\text{rec},i}, \mathbf{z})$$

(loss_recon_scaled)

Data

[00124] Two datasets were used in the training and evaluation of DrugHIVE, PDBbind (v2020, refined set) and a subsample of the ZINC drug-like subset. The PDBbind refined set was chosen as it provides a set of high-quality crystal structures of ligands bound to protein receptors. Structures with ligands too wide to fit well into the density grid (> 16Å) were removed as well as those that have too many atoms (>43). A test set of 100 structures was selected from the dataset by first clustering all structures by 30% sequence identity and then randomly assigning clusters until the desired split is achieved. A subsample of the ZINC dataset as naked drug molecules was also selected to improve training and reduce overfitting to the relatively small number of example ligands in the PDBbind dataset. To acquire the subset, the ZINC drug-like subset was used and about 27 million molecules were selected uniformly by alogP and number of heavy atoms, filtering out the small number of structures with heavy atoms not in the current atomic element set ($Z \in \{C, N, O, F, P, S, Cl, Br, I\}$) (other elements are used in other embodiments).

[00125] For evaluation of generating from AlphaFold predicted structures, a dataset of 100 AF structures for which there is also the PDB structure were used. To create this dataset, all structures from the PDBbind dataset where the receptor pocket is only comprised of a single protein chain and that chain is available from the AlphaFold Protein Structure Database were identified. Sequence and structural alignment of the AF structure to the PDB structure using PyMOL align function, filtering out structures with less than 70% sequence identity was performed. The binding pocket was extracted by cutting off any excess residues based on sequence alignment and removing all residues with atoms > 24Å from the ligand. Any structures without a high confidence score (pLDDT < 70) for any residue within 5Å of the ligand were filtered out. Finally, a random sample of 100 of the remaining receptors for the AF test set was taken.

Sampling

[00126] Embodiments of the models provided for herein have multiple sampling modes. Once a model is trained, new molecular structures can be generated by inputting (1) a protein structure into the protein encoder and (2) a randomly sampled set of values from the prior into the ligand decoder. This mode is referred to as prior sampling, and the result is a randomly generated density that can then be fit with a molecular structure. Since known ligands can be encoded into a set of latent values, the same procedure can be repeated except, instead of randomly sampling from the prior, the latent values from a previously encoded ligand structure can be used. This mode is referred to as posterior sampling, and the result is a density that will be very similar to the true input ligand density.

[00127] These two sampling modes are combined into what is referred to herein as prior-posterior sampling (see Figure 6, especially 6F) to generate densities of varying similarity to an initial ligand density. To achieve this, the procedure for posterior sampling is repeated while also sampling a random set of values from the latent prior. Then starting with the values from the posterior, interpolation toward the values sampled from the prior is performed:

$$z_i = z_{i,post} + \beta_i\left(z_{i,prior} - z_{i,post}\right), \text{ (z_prior_post)}$$

[00128] with β_i being the interpolation factors. Since the latent has multiple scales, one can specify a different interpolation factor for each scale. Different molecular properties are encoded at different scales (see Figures 10A-10E), so this leads to a high degree of control over the variation of molecular details in generating new structures. Since each latent tensor retains spatial context, spatial prior-posterior sampling (Figure 6I) by performing prior-posterior sampling over a spatial subset of the density grid.

[00129] Another means of controlling the generation process is through a scaling (temperature) factor on the variance of each latent variable. The latent space of models provided for herein is a high dimensional normal distribution, with each latent value being sampled from a distribution, $z_i \sim N(\mu_i, \sigma_i\tau_i)$, where μ_i is the mean, σ_i is the variance, and τ_i is the temperature factor. As with the prior-posterior interpolation factor, the temperature factor can be varied across latent scales and spatial subsets. The default temperature factor during training is $\tau_i = 1$. Decreasing the temperature factor has the effect of lowering the variability of generated samples.

Spatial Sampling

[00130] The latent space of models provided for herein is comprised of a set of tensors, each representing information at a specific resolution of molecular space (Figure 4B). This varying resolution is the result of downsampling layers in the neural network. In models provided for herein, each latent tensor retains its spatial context; thus, it is possible to perform spatial prior-posterior sampling by performing prior-posterior sampling over a subset of the density grid (Figure 7). For this sampling mode, a region of the input grid was chosen on which to perform prior-posterior sampling. For each latent tensor, it was determined which values correspond to the modification region and which fall outside by mapping the tensor values back to the full-resolution grid. Prior-posterior sampling was then performed, setting the interpolation factors equal to zero for all latent values outside of the modification region. The result is a generated structure similar to the input ligand except for within the modification region, which will differ more or less according to interpolation factor value. This sampling mode is used for the applications of substructure modification, linker design, fragment growing, and pattern replacement (Figures 5A-5D), which are described below.

Substructure Modification

[00131] Given a substructure of a parent molecule to be modified as input, the molecule is first rotated such that the first three principal components of the substructure coordinates are aligned with the axes of the grid. The coordinates of the substructure atoms are used to determine a bounding box that contains every atom to be modified while excluding all other atoms in the parent molecule. That box is used to define the modification region and apply spatial prior-posterior sampling. Finally, a new substructure is fit to the resulting density within the modification region and connect that substructure to the unmodified atoms from the parent molecule using a bond-adding procedure as described herein.

Fragment Growing

[00132] As input, a molecular structure to be grown is selected and a region adjacent to the molecule where a new structure is to be added, as defined by the box coordinates. That box is used as the modification region and apply spatial prior-posterior sampling. A substructure is then fit to the resulting density and connected to the initial structure using a bond-adding procedure as

described herein. In some embodiments, sparsely seeding the modification region with atoms before the encoding step is performed to prevent the model from generating empty densities within that region. This random seeding does not determine the atom types or precise spatial distribution of the generated density, since this region is resampled. According to several embodiments, this step is necessary because empty regions are described nonlocally within the upper levels of the latent representation.

Linker Design

[00133] Given a structure representing two molecular fragments to be connected as input, the structure is first rotated such that its first three principal components are aligned to the axes of the grid. The modification region is then defined by determining a box containing the empty region between the two fragments and apply spatial prior-posterior sampling. As with certain embodiments of fragment growing, the empty modification region is seeded before the encoding step. Finally, a new substructure (the linker) is fit to the resulting density and connected to the initial fragments with a bond-adding procedure as described herein.

Pattern Replacement

[00134] As input, a set of SMILES or SMARTS patterns are selected and a set of molecules over which to search for and replace those patterns. For each pattern, iteration is performed through the set of molecules to determine if that pattern is present. If so, the substructure representing the pattern is found and the same procedure is performed as described in substructure modification to generate a new set of molecules with that pattern replaced. For the example of PAINS pattern replacement (Figured 5D and 5E), a set of filters was used to identify PAINS molecules in the PDB crystal ligands from the training data set and 200 molecules were generated with the PAINS pattern replaced for the ligand bound to phospholipase A2 (PDB ID 1Q7A).

Evolutionary Molecular Optimization

[00135] Evolutionary (genetic) algorithms are a commonly used tool employed in the multi-objective optimization of molecules. They mimic the process of evolution to efficiently search through spaces of high complexity, where ordinary search methods would be prohibitively inefficient or difficult to implement. According to embodiments provided for herein, an

evolutionary algorithm coupled with prior-posterior sampling is employed to search the latent space of the provided models and optimize desired molecular properties. In embodiments provided for herein, the genetic information is represented by the latent encoding of molecules and the mutation rate is represented by the prior-posterior interpolation factor (β).

[00136] Initialization step: Generate an initial population of N_c molecules. Apply the fitness function and cluster the molecules (Tanimoto similarity < 0.2), keeping the most fit individual from each cluster. Select the N_p best molecules from the current population as "parents".

[00137] Optimization step (repeat until converged): Generate a new population of N_c "child" molecules from parents. Apply the fitness function and cluster, keeping the most fit individual from each cluster. Replace least fit parents with child molecules having better fitness score.

[00138] During generation, molecules with configurations that lead to strained geometries and false positive docking results are filtered out. Additionally, molecules with consecutive double bonds, fused ring systems of more than four members or that form loops, and rings other than five- and six-membered rings are removed.

[00139] The fitness function will vary depending on desired outcomes but, in general, will be some measure of distance between the observed molecular property value(s) and the desired ones. For example, in the case of binding affinity optimization, the predicted binding affinity score is used as the fitness measure. For multiobjective optimization, the fitness function can be a combination (e.g., weighted sum) of individual fitness scores. In nonaffinity property optimization experiments, a goal is to maintain predicted binding affinity while optimizing the targeted property by removing all child molecules from the population with a binding affinity score worse than a threshold value.

Force Field Optimization

[00140] Prior to docking, we perform force field optimization on all molecular structures using the MMFF94 force field in RDKit. When using the Universal Force Field, a lack of robustness in convergence was empirically observed; therefore, this method is not used here. Moreover, structures in the context of the receptor are not relaxed for at least the following reasons. First, docking algorithms are generally designed to take an energy-minimized ligand structure as input. For docking algorithms based on AutoDock, it is important to input chemically valid molecular coordinates. It is even recommended that ligands from crystal structures that are geometrically

distorted by the receptor have their structures optimized before docking. Second, as system size increases, convergence to an optimal minimum becomes less reliable. Including the surrounding receptor atoms in the calculation, even if fixed to their initial positions, may lead to a higher likelihood of convergence failure, false positives, or geometric distortions outside the distribution of structures on which the parameters of the Vina scoring function were fit. This step is important with de novo generated structures, which are likely to be far from an energy minimum. Using unconverged structures or skipping the force field optimization on structures before docking could lead to significantly exaggerated docking scores (Figures 11A-11C).

Evaluation Properties and Metrics

[00141] Binding affinity is estimated for each ligand with the Vina docking score calculated with QuickVina 2, a speed optimized version of AutoDock Vina. Δ Vina is defined as the difference between the Vina score of a generated ligand and the reference ligand (e.g., the ligand from the crystal structure). Molecular similarity is calculated as one minus the Tanimoto distance between the molecular fingerprints of a pair of molecules. Molecular diversity is defined on a set of molecules as the average Tanimoto distance between each pair in the set. The Quantitative Estimate of DrugLikeness (QED) score estimates the drug-likeness of a molecule by combining a set of molecular properties. The Synthetic Accessibility (SA) score estimates the ease of synthesis of a molecule.

Evaluation

[00142] Molecules generated by models provided for herein are evaluated against comparable approaches and a set of randomly sampled molecules from the ZINC20 drug-like subset by comparing the predicted binding affinities of the molecules. Like the present model, each compared model is a state-of-the-art structure-based deep generative model that conditions molecular generation on the target protein pocket structure. The three comparator models are LiGAN, a grid-based standard VAE; DiffSBDD, a graph-based equivariant diffusion model; and Pocket2Mol, a graph-based equivariant autoregressive model. 200 molecules were generated for each receptor in the test set for each method, and calculate *QED*, *SA*, and average diversity using RDKit. Force field optimization was performed on each molecule using the MMFF94 force field in RDKit. Each molecule was virtually docked to the corresponding receptor using QuickVina 2

(with default settings and a 20 Å-wide boundary box) to predict the binding affinity. All comparator models generate structures with explicit three-dimensional (3D) conformations, defined chirality, and protonation states. Therefore, only the structure generated by each model was docked against a rigid receptor and do not enumerate, for instance, tautomers or stereoisomers.

[00143] The *Vina* score is strongly correlated with the number of atoms (Figures 10A-10D) in the molecule, which prevents direct comparison between sets of molecules with different size distributions. To overcome this, bootstrap sampling was used to estimate the average *Vina* score given a uniform distribution of molecular sizes. First, the generated molecules were binned by size, n_a = number of heavy atoms. The largest common interval of molecular sizes over which to sample, such that all sets have at least 100 molecules in each bin, was chosen. This results in the interval $n_a \in [13, 25]$. 20 molecules from each group were randomly sampled, aggregated, and the average binding score, ρ_i was computed for each. This process was repeated $N = 200$ times and compute the cumulative average $\rho^{-}=1N\sum i\rho i\rho^{-}=1N\sum i\rho i$. This cumulative average for each method is reported as *Vina-n* in Figure 10D. In Figure 10B, the distributions of molecular size and the predicted affinity versus the number of heavy atoms for each generated set are plotted for each of the indicated models.

RESULTS

Improved De Novo *Drug Generation and Virtual Screening Efficiency*

[00144] The models provided for herein excel in generating new molecules with drug-like properties and a strong predicted binding affinity for their target receptors, outperforming diffusion and graph-based networks in the average predicted binding affinity and drug-likeness of generated molecules (see Figure 10-A-10D). The ability of the models provided for herein to generate new candidate ligands for a test set of 100 diverse receptors was evaluated and the results compared against other models (LiGAN, DiffSBDD, and Pocket2Mol) and a random sample of molecules from the ZINC drug-like subset (see Figure 10-A-10D). For each receptor, molecules were generated by randomly sampling from the latent prior and compute the estimated binding affinity and commonly reported properties of each molecule. The presently disclosed models perform well compared to other models on common benchmark metrics. Molecules generated by models provided for herein have unexpectedly improved predicted binding affinities and drug-likeness scores, on average, then current state-of-the-art models.

[00145] Although certain metrics are frequently reported in the literature as benchmarks of generative performance, there have been common oversights regarding how these metrics are calculated and interpreted. For instance, most studies in the field neglect to correct for the strong correlation of *Vina* score (the most commonly used metric for binding affinity) with molecule size. As a result, the best models in the literature may simply be the ones that generate the largest molecules. It is important in some embodiments to correct for this issue, given that many models, including point-cloud and graph-based approaches, require the size of each generated molecule to be specified. Moreover, studies commonly report the average diversity of generated molecules as a metric to be uniformly maximized, such that the best model is the one that generates the highest molecular diversity. This logic may hold for the generation of unbound molecules, but it breaks down for ligands conditioned on specific target receptors. However, given the specificity of drug-target interactions, one would not expect a good model to generate significantly higher molecular diversity than that of a random set of drug-like molecules from ZINC, for instance. In this case, it will be unclear which is best within a reasonable range of average diversity.

[00146] Approximately 8.7% of the molecules generated using the disclosed models are of equal or smaller size than the reference ligand has a predicted affinity as good as or better than the reference ligand. This is more than double the value (3.9%) of a random sample of the ZINC data set, which means that the virtual screening efficiency is significantly higher for DrugHIVE-generated molecules (see Figure 6E). To further demonstrate this result, 1000 molecules were generated for the Rho-associated protein kinase 1 (ROCK1) receptor (PDB ID 3TWJ), an important target with a documented role in cancer migration, invasion, and metastasis. Approximately 12% of generated ligands have a higher predicted binding affinity than the crystal ligand (RKI-1447), a type I kinase inhibitor with a reported IC_{50} of 14.5 nM.

[00147] In Figure 9, a comparison of the properties of reference ligands in the PDBbind data set and a random sample of molecules from the ZINC data set with those of ligands generated from the disclosed models is presented. Properties of generated molecules fall well within the distributions of the training data sets, indicating that the model effectively learns the training distributions. A significant distribution shift between molecules generated with and without protein receptors was observed, which indicates that the model is conditioning on receptor information during generation.

Multi-Objective Optimization Of Drug Properties Via Evolutionary Latent Search

[00148] The models provided for herein can generate new ligands with a high degree of control over similarity to a reference molecule using prior-posterior sampling (Figure 6F). Each ligand–receptor complex has a unique representation in the encoded latent space. Therefore, starting with a given reference ligand latent representation (posterior) and interpolating toward a randomly sampled point (prior), the resulting molecules will have a higher average diversity and lower molecular similarity to the reference as we interpolate toward the prior sampled point (Figure 6D). Taking advantage of this high level of control over variation, the present methods implement an evolutionary algorithm to search the latent space and optimize the properties of a reference molecule. Figures 6A-6C shows the optimization process for the ligand (PDB Chemical ID 57G) bound to the human bromodomain-containing protein 4 (BRD4) receptor (PDB ID 5D3H).

[00149] Figure 6A shows the successful optimization of binding affinity (*Vina*), drug-likeness (*QED*), synthesizability (*SA*), and hydrophobicity (*ALogP*) values. For *QED*, *SA*, and *ALogP*, we perform multiobjective optimization by including binding affinity as a secondary target in the fitness function. Significant improvement for each target property was observed (circled values), including a 51% improvement in the predicted binding affinity score for the best molecule, even though the ligand shows strong experimental affinity to the target receptor (K_d = 7.9 μM). *QED* was successfully increased from 0.49 to 0.94 and *SA* from 0.65 to 0.94. In both cases, the value is increased from a moderate to a high value while simultaneously improving binding affinity. In Figure 6C, *ALogP* is optimized in the upward direction, beginning with the slightly hydrophobic ligand (*ALogP* = 0.82) and ending with a very hydrophobic molecule (*ALogP* = 6.72). Next, the *ALogP* value was optimized in the downward direction, reversing the process to result in a hydrophilic molecule (*ALogP* = −1.28).

[00150] Figure 6G shows results from the generation and optimization of a set of ligands for selectivity toward human ataxia-telangiectasia mutated (ATM) kinase and against human vacuolar protein sorting (Vps34) kinase. It has recently been shown that ATM kinase is an important potential target for the treatment of Huntington's disease, and that it is important for ATM inhibitors to exhibit high selectivity over Vps34. Models disclosed here were used to generate an initial set of ligands for the ATM receptor (PDB ID 7NI5), which were then optimized for selectivity against the Vps34 receptor (PDB ID 4UWH). The results show considerable improvement of selectivity, with the best molecule having a significantly better binding affinity

score for ATM (−9.3 kcal/mol) compared to Vps34 (−6.1 kcal/mol). These experiments demonstrate that the latent encoding learned by the models provided for herein is a powerful representation of chemical space, enabling effective multiobjective optimization of molecular structures with a high degree of control.

Lead Optimization with Spatial Prior-posterior Interpolation

[00151] The hierarchical latent structure of models provided for herein preserves spatial context and allows molecules to be spatially modified by applying prior-posterior sampling on a spatial subset of the latent representation (as shown in Figure 6B). This method was applied to the common drug design problems of fragment optimization, linker design, and molecular pattern replacement (see Figure 6A).

Linker Design

[00152] Fragment-based screening is a common tool in the early drug discovery toolkit. The result of this type of screen is a set of relatively small molecular fragments with modest affinity for the target receptor. Given two or more such fragments, the challenge is to connect the individual fragments into a coherent, high-affinity drug-like molecule. Ordinarily, this is a manual and time-consuming process. Models as provided for herein can be used to automatically connect a set of individual fragments as oriented in the cocrystal structure. As an example (Figure 7A), it is shown that the models provided for herein successfully design a molecular linker to connect individual fragments (ligands 1DZ and 1XS) from a screening experiment for replication protein A (PDB ID 4LUV). All resulting ligands have a *Vina* score better than either of the individual fragments alone, with 21% of generated molecules having equal or better *Vina* scores, and 51% having equal or better *SA* scores, than the linked molecule (ligand 1XU from PDB ID 4LWC) designed by ref (experimental K_d = 20 μM).

Fragment Growing

[00153] Another common drug design task is to take a known fragment with modest binding affinity due to its size and grow the structure to improve its affinity for the target. Models as provided for herein can be used to automatically design new substructures to add to an existing molecular structure. To demonstrate this, it is shown that a model provided for herein successfully grows one fragment (ligand 1DZ) from the same fragment screening experiment, thereby

significantly increasing its predicted binding affinity for the target receptor (Figure 7D), with 85% of generated molecules having an improved *Vina* score.

Substructure Modification (Scaffold Hopping)

[00154] A common drug optimization technique is the systematic replacement of either chemical scaffolds or R-groups on a candidate molecule to improve its properties or grow a candidate set. Starting with the crystal ligand (3JU) bound to the von-Hippel Lindau tumor suppressor (PDB ID 4W9F), it was shown how automatic substructure optimization can be achieved using spatial prior-posterior sampling. BRICS decomposition was employed to break the ligand, an agonist with reported $K_d = 3.27$ μM, into molecular fragments. For each fragment, a set of molecules was generated for which only the fragment substructure was modified, leaving the rest of the molecule unchanged (Figure 7A). For each of the BRICS fragments, the distribution of properties for the generated molecules was analyzed to learn the contribution of each fragment to each property. Depending on which fragment is modified, there are significant differences on the effect on each property (Δ*Vina*, *ALogP*, *TPSA*, *QED*) (Figure 7C). For instance, drug-likeness scores are improved only for fragment 2. Modification of fragment 3 leads to the strongest decrease in hydrophobicity (*ALogP*) values, while having almost no effect on topological polar surface area (*TPSA*). Modification of fragment 1 leads to the strongest increase in *Vina* score, while fragment 3 has relatively little effect. This approach can generate valuable insights for lead optimization and structure–activity relationship analysis.

Pattern Replacement

[00155] Given a molecular SMILES or SMARTS pattern, it is possible to modify the molecular structure(s) corresponding to that pattern for a single molecule or a set of molecules, using models as provided for herein to generate a new set of molecules with the input pattern replaced. This functionality could be used to modify fragments with known toxic, promiscuous, or otherwise undesirable properties in existing or newly generated drug collections, mitigating the need to screen out candidates with otherwise promising qualities. One non-limiting example of application of this is in the replacement of PAINS patterns, which are a set of molecular substructures shown to be correlated with false positive results in HTS assays. These molecular patterns do not always lead to experimental interference; therefore, screening decisions should be made on a case-by-case

basis. To avoid risks of false-positive experimental results and false-negative screening decisions, all PAINS patterns in an existing collection can be replaced using generative spatial modification as provided for herein.

[00156] As an example, a set of filters was applied to identify PAINS molecules in PDB crystal ligands from the training data set and identify the ligand bound to phospholipase A2 (PDB ID 1Q7A) as a PAINS molecule. Figures 5A-5E show the results from generating 200 molecules with the PAINS pattern replaced with 28% of generated molecules exhibiting an improved binding affinity score. This pattern replacement capability is scalable to large collections of molecules, as the run time per molecule is on par with fast virtual docking algorithms.

Drug Design for Predicted Target Structures

[00157] A major limitation of structure-based drug design methods is their reliance on high-resolution crystal structures of the target receptor. Although the availability of protein crystal data is rapidly growing, with the number of structures deposited in the PDB averaging over 9000 per year in the past decade, still >80% of the human proteome remains unsolved. Recently, sequence-to-structure prediction models, such as AlphaFold2, have gained the ability to generate high confidence predictions for previously unknown proteins, filling in most of the missing structural data for foldable regions of the human proteome.

[00158] The present example shows that models as provided for herein are capable of generating high-affinity molecules for receptors using AlphaFold (AF)-predicted structures. To test the effectiveness of using predicted structures for generative drug design, ligands were generated for a test set of target receptors for which both the AF prediction and the PDB crystal structure are available (representing the predicted and actual structure (respectively)). The process for AF target curation is outlined in Figure 8A. Importantly, no crystal structure alignment information was used in selecting AF targets. Filtering is based solely on sequence alignment to ensure adequate sequence similarity and the AF confidence score *pLDDT* to ensure high-confidence predictions.

[00159] AF-generated ligands were docked into both the AF and crystallographically resolved receptors and the binding affinity predictions compared to ligands generated for the PDB receptor. Figure 8B shows the distribution of $\Delta Vina$ scores and Figures 13A-13B show selected properties for the AF- and PDB-generated ligands docked to corresponding PDB receptors. The range of

affinities is very similar between the two sets, with a slightly better median value for AF-generated ligands. The proportion of ligands with a better docking score than the crystal ligand is around 20% for both sets and is slightly better for AF-generated ligands. There is good correlation between *Vina* scores when docked to both the PDB and AF receptors for AF-generated ligands ($R = 0.8$) and PDB-generated ligands ($R = 0.6$). Furthermore, as shown in 8C, models as provided for herein allow optimization of the predicted affinity of ligands using the AF pocket. The optimization process begins with the PDB crystal ligand being placed in the AF-predicted structure corresponding to PDB ID 5D3H. The resulting molecule has a similarly high predicted binding affinity compared to the molecule obtained through the same optimization process with the PDB pocket (Figure 6A). All molecules generated during the optimization process were docked into the PDB pocket and the Pearson correlation coefficient was calculated for the two sets of affinity scores. We obtain a high correlation ($R = 0.96$, Figure 8C), with AF-docked scores consistently underestimating PDB-docked scores — a bias that is exaggerated for higher affinity ligands. Similar data is show in Figures 8E-8F, for ligand (S)-1-[N2-(1-carboxy-3-phenylpropyl)-L lysyl]-L-proline dihydrate bound to the human angiotensin-converting enzyme (ACE) receptor (PDB ID 1086). Despite this underestimation, the high correlation allows the DrugHIVE optimization process to significantly improve binding affinity scores of the initial ligand.

DISCUSSION

[00160] Provided for herein are computer implemented methods and systems to enhance the structure-based design of drug-like molecules based on a deep HVAE architecture. One motivation in designing the models provided for herein was the observation that, although molecular systems clearly possess structure and properties at varying spatial scales, previous encoder-decoder models have failed to adequately capture these relationships. In contrast, hierarchical prior provided for herein enables the disclosed models to better learn the distribution of intra- and intermolecular relationships and thereby generate high-affinity drug-like molecules with an unprecedented degree of spatial control. In particular, embodiments provided for herein display an impressive ability to optimize the properties of molecules over a wide range of values within the context of a protein receptor by performing local exploration of the latent space.

[00161] Beyond finding a more natural representation of chemical space, another motivation in designing the models provided for herein was improving practical applicability for drug-design

tasks. Many tasks in early drug design are time-consuming and rely heavily on the intuition and expertise of researchers. While such expert knowledge will always be a crucial part of the process, generative models such as those provided for herein speed up cycle times by automating the de novo design and modification of molecular structures. In this example, it was shown how models as provided for herein can be used to perform rapid and effective linker design, scaffold hopping, fragment growing, and high-throughput pattern replacement—all of which ordinarily require careful expert attention. While the models disclosed herein were only used to generate molecules that fit within the input density grid, and not larger molecules, larger molecules can be generated according to some embodiments. The grid width chosen (24 Å) fits most drug-like molecules, but extending this approach to larger structures may require additional computational requirements. As discussed herein, the dual-resolution implementation provided for mitigates this issue significantly by lowering the memory requirements. In some embodiments, wherein grids of about 48 Å or larger are used, multi-GPU training can synergistically work with the dual-resolution aspect of the model to generate new larger molecules.

[00162] The results of this example show that generative models provide an exciting new avenue for expanding development in the vast drug-like chemical space, most of which is currently unexplored and inaccessible. For the 100 diverse protein structures in the present test data set, 8.7% of molecules generated by the disclosed models have a predicted affinity higher than that of the crystal ligand. This is significantly higher than the 3.9% observed for a randomly chosen ZINC screening data set. In other words, performing hit identification by screening generated molecules has unexpectedly already surpasses the efficiency of current screening data sets. Additionally, virtually none of the generated molecules exist in the training data set, which means that disclosed models are able to access previously unexplored regions of chemical space. Figures 14A-14D illustrate the random molecules that are generated by the disclosed models under numerous sampling conditions. In the event that some fraction of these randomly generated examples is of poor quality and will need to be filtered, the nature and extent of filtering can be determined by the application and, for example, could include checks for chemical validity, reactivity, synthesizability, toxicity, or specific chemical properties.

[00163] While synthesizability of generated molecules does appear to be lower than that of current screening libraries, which is a significant factor in successful hit identification, this is partly due to the relatively low synthesizability of crystal ligands compared to that of drug-like molecule

collections. This is particularly true for those collections generated through chemical synthesis enumeration such as Enamine REAL. In additional embodiments, improvements to synthesizability of generated molecules, such as synthon-based generation, is believed to significantly improve their practical utility for hit identification. Additionally, embodiments can leverage the improvements to retrosynthesis planning using deep learning methods are leading to more accurate identification of synthesizable compounds and advances in chemical synthesis methods are expanding the accessible molecular space.

[00164] As demonstrated herein, the presently disclosed models are not limited to existing crystal structures. Molecular generation with the disclosed models using AF-predicted structures generalizes well to the PDB structures that were tested. Over 20% of ligands generated for a diverse set of AF-predicted structures have higher predicted binding affinities than the crystal ligand when docked to the PDB crystal structure, with a high correlation between docking scores ($R = 0.8$). Furthermore, optimization of binding affinity using the AF-predicted structure yields molecules with significantly improved predicted binding affinities for the PDB structure. In other words, in cases where a crystal structure is not available, docking to a high-confidence AF structure results in a reasonable estimate of the crystal structure docking score. In some embodiments, this may not, however, yield the correct binding pose. The success of generating molecules with enhanced characteristics is important given that AlphaFold2 currently provides high-confidence predictions for 58% of the human proteome, including perhaps all remaining foldable regions without crystal data. Thus, the presently disclosed models enable the reliable generation of new molecules well ahead of time as may otherwise be available given the time to generate the remaining crystal structures. Finally, application of the presently disclosed models is agnostic with regards to the choice of virtual docking algorithm. As virtual docking methods improve, the presently disclosed models will have more refined information to encode and decode, and enhanced drug design will result.

[00165] By developing a model with more fine-grained control over molecular generation, the presently disclosed models function to accelerate the impact of generative models on early drug discovery. The presently disclosed generative models enhance drug design by automating a wide array of drug design tasks. Further progress and widespread adoption of these methods has the potential to significantly accelerate and lower the cost of preclinical drug development. With an estimated cost of $0.04 to $0.20 per thousand generated molecules at current cloud computing

rates, on par with current high-throughput VLS methods, the presently disclosed models are a powerful and scalable tool for automated structure-based drug design.

[00166] Also provided for herein are the following non-limiting embodiments.

[00167] 1. A machine-learning method for generating a plurality of molecular structures having a predictable binding affinity for a target receptor, said method comprising:

[00168] (1) creating a molecular density grid comprising atomic occupancy and atom-type densities in the target receptor by processing data relating to a structure of said target receptor or relating to a structure of at least one ligand having a molecular structure capable of forming a ligand-receptor complex with said target receptor;

[00169] (2) fitting a plurality of atom positions and types to each atomic occupancy density and up-sampling atom type densities determined therefrom;

[00170] (3) assigning an atom type to each atomic occupancy density from the up-sampled atom type densities; and

[00171] (4) building molecular structures or molecular substructures by fitting chemical bonds between the assigned atom positions and types within the atomic occupancy densities, then optionally fitting additional chemical bonds or molecular linker structures between molecular substructures to generate the plurality of molecular structures.

[00172] 2. The method of embodiment 1, wherein data relating to a structure of said target receptor comprises crystal structure data.

[00173] 3. The method of embodiment 1, wherein the processing of data comprises creating a hierarchal latent space further comprising a set of tensors, each tensor spatially representing a reference ligand-receptor complex latent representation formed from each of said at least one existing ligand and the target receptor or representing a reference ligand-receptor complex formed from a molecular structure thus generated and the target receptor.

[00174] 4. The method of embodiment 3, wherein generating the plurality of molecular structures further comprises a prior-posterior sampling process further comprising interpolating toward a randomly sampled set of tensors in the latent space beginning from a reference ligand-receptor complex sampled from a latent distribution, said prior-posterior sampling configured to optimize a binding affinity of a molecular structure thus generated by said method and the target receptor.

[00175] 5. The method of embodiment 4, wherein said prior-posterior sampling configured to optimize a binding affinity further comprises a decoding of the interpolation into new molecular densities.

[00176] 6. The method of embodiment 1, wherein fitting a plurality of atom types to each atomic occupancy density further comprises application of an atom-fitting algorithm configured to iteratively place atomic coordinates for a chosen atom type within the molecular density grid to minimize a residual occupancy density.

[00177] 7. The method of embodiment 1, wherein the fitting of chemical bonds between the assigned atom types within the atomic occupancy densities comprises application of a bond fitting algorithm.

[00178] 8. The method of embodiment 1, wherein in creating the molecular density grid comprising atomic occupancy and atom-type densities within the target receptor by processing data relating to at least one ligand having a molecular structure capable of forming a ligand-receptor complex with said target receptor, each atom of the ligand is represented by a multi-channel pseudo-Gaussian density with channels, each channel representing a different atomic feature of each atom.

[00179] 9. The method of embodiment 8, wherein the atomic features comprise the elements C, N, O, F, P, S, Cl, Br, and I, and combinations thereof, and the features of each element being H-bond donor, H-bond acceptor, aromaticity, positive (+) charge, negative (-) charge, or neutral (0) charge.

[00180] 10. The method of embodiment 1, wherein the molecular density grid measures, for example, 24 x 24 x 24 Angstroms and the molecular density grid can be an input and/or output data structure.

[00181] 11. The method of embodiment 1, wherein the molecular density grid is a molecular graph or a point cloud or point set where the molecular graph is comprised of a set of atomic coordinates (vertices), bonds (edges), atom (vertex) types, and bond (edge) types and the point cloud or point set is a graph without edges or edge types.

[00182] 12. A system for generating a plurality of molecular structures having a predictable binding affinity for a target receptor, said system comprising:

[00183] a processor configured to perform operations comprising:

[00184] create a molecular density grid comprising atomic occupancy and atom-type densities in the target receptor by processing data relating to a structure of said target receptor or relating to a structure of at least one ligand having a molecular structure capable of forming a ligand-receptor complex with said target receptor;

[00185] fit a plurality of atom positions and types to each atomic occupancy density and up-sampling atom type densities determined therefrom;

[00186] assign an atom type to each atomic occupancy density from the up-sampled atom type densities; and

[00187] build molecular structures or molecular substructures by fitting chemical bonds between the assigned atom positions and types within the atomic occupancy densities, then optionally fitting additional chemical bonds or molecular linker structures between molecular substructures to generate the plurality of molecular structures.

[00188] 13. The system of embodiment 12, wherein data relating to a structure of said target receptor comprises crystal structure data.

[00189] 14. The system of embodiment 12, wherein the processing of data comprises creating a hierarchal latent space further comprising a set of tensors, each tensor spatially representing a reference ligand-receptor complex latent representation formed from each of said at least one existing ligand and the target receptor or representing a reference ligand-receptor complex formed from a molecular structure thus generated and the target receptor.

[00190] 15. The system of embodiment 14, wherein generate the plurality of molecular structures further comprises a prior-posterior sampling process further comprising interpolating toward a randomly sampled set of tensors in the latent space beginning from a reference ligand-receptor complex sampled from a latent distribution, said prior-posterior sampling configured to optimize a binding affinity of a molecular structure thus generated by said method and the target receptor.

[00191] 16. The system of embodiment 15, wherein said prior-posterior sampling configured to optimize a binding affinity further comprises a decoding of the interpolation into new molecular densities.

[00192] 17. The system of embodiment 12, wherein fit a plurality of atom positions and types to each molecular density further comprises application of an atom-fitting algorithm configured to

iteratively place atomic coordinates for a chosen atom type within the molecular density grid to minimize a residual occupancy density.

[00193] 18. The system of embodiment 12, wherein the fitting of chemical bonds between the assigned atom types within the atomic densities comprises application of a bond fitting algorithm.

[00194] 19. The system of embodiment 12, wherein in creating the molecular density grid comprising atomic occupancy and atom-type densities within the target receptor by processing data relating to at least one ligand having a molecular structure capable of forming a ligand-receptor complex with said target receptor, each atom of the ligand is represented by a multi-channel pseudo-Gaussian density with channels, each channel representing a different atomic feature of each atom.

[00195] 20. The system of embodiment 19, wherein the atomic features comprise the elements C, N, O, F, P, S, Cl, Br, and I, and combinations thereof, and the features of each element being H-bond donor, H-bond acceptor, aromaticity, positive (+) charge, negative (-) charge, or neutral (0) charge.

[00196] 21. The system of embodiment 12, wherein the molecular density grid measures, for example, 24 x 24 x 24 Angstroms and the molecular density grid can be an input and/or output data structure.

[00197] 22. The system of embodiment 12, wherein the molecular density grid is a molecular graph or a point cloud or point set where the molecular graph is comprised of a set of atomic coordinates (vertices), bonds (edges), atom (vertex) types, and bond (edge) types and the point cloud or point set is a graph without edges or edge types.

[00198] In the detailed description, references to “various embodiments”, "one embodiment", "an embodiment", "an example embodiment", etc., indicate that the embodiment described may include a particular feature, structure, or characteristic, but every embodiment may not necessarily include the particular feature, structure, or characteristic. Moreover, such phrases are not necessarily referring to the same embodiment. Further, when a particular feature, structure, or characteristic is described in connection with an embodiment, it is submitted that it is within the knowledge of one skilled in the art to affect such feature, structure, or characteristic in connection with other embodiments whether or not explicitly described. After reading the description, it will be apparent to one skilled in the relevant art(s) how to implement the disclosure in alternative embodiments.

[00199] Steps recited in any of the method or process descriptions may be executed in any order and are not necessarily limited to the order presented. Furthermore, any reference to singular includes plural embodiments, and any reference to more than one component or step may include a singular embodiment or step. Also, any reference to attached, fixed, connected, coupled or the like may include permanent (e.g., integral), removable, temporary, partial, full, and/or any other possible attachment option. Any of the components may be coupled to each other via friction, snap, sleeves, brackets, clips or other means now known in the art or hereinafter developed. Additionally, any reference to without contact (or similar phrases) may also include reduced contact or minimal contact.

[00200] With respect to ingredients that are grouped by function or characterized herein, unless otherwise expressly stated herein, it is not precluded that such an ingredient could be grouped with or characterized by another function. For example, an ingredient discussed herein as a solubilizer could also function as a surfactant and vice versa. The ranges disclosed herein also encompass any and all overlap, sub-ranges, and combinations thereof. Language such as "up to," "at least," "greater than," "less than," "between," and the like includes the number recited. Numbers preceded by a term such as "about" or "approximately" include the recited numbers. For example, "about 90%" includes "90%." Language such as "at least 95% (S)-nicotine in the form of a free base alkaloid includes 96%, 97%, 98%, 99%, and 100% (S)-nicotine in the form of a free base alkaloid. Any titles or subheadings used herein are for organization purposes and should not be used to limit the scope of embodiments disclosed herein.

[00201] Benefits, other advantages, and solutions to problems have been described herein with regard to specific embodiments. However, the benefits, advantages, solutions to problems, and any elements that may cause any benefit, advantage, or solution to occur or become more pronounced are not to be construed as critical, required, or essential features or elements of the disclosure. The scope of the disclosure is accordingly to be limited by nothing other than the appended claims, in which reference to an element in the singular is not intended to mean "one and only one" unless explicitly so stated, but rather "one or more." Moreover, where a phrase similar to 'at least one of A, B, and C' or 'at least one of A, B, or C' is used in the claims or specification, it is intended that the phrase be interpreted to mean that A alone may be present in an embodiment, B alone may be present in an embodiment, C alone may be present in an

embodiment, or that any combination of the elements A, B and C may be present in a single embodiment; for example, A and B, A and C, B and C, or A and B and C.

[00202] All structural, chemical, and functional equivalents to the elements of the above-described various embodiments that are known to those of ordinary skill in the art are expressly incorporated herein by reference and are intended to be encompassed by the present claims. Moreover, it is not necessary for an apparatus or component of an apparatus, or method in using an apparatus to address each and every problem sought to be solved by the present disclosure, for it to be encompassed by the present claims. Furthermore, no element, component, or method step in the present disclosure is intended to be dedicated to the public regardless of whether the element, component, or method step is explicitly recited in the claims. No claim element is intended to invoke 35 U.S.C. 112(f) unless the element is expressly recited using the phrase "means for." As used herein, the terms "comprises", "comprising", or any other variation thereof, are intended to cover a non-exclusive inclusion, such that a chemical, chemical composition, process, method, article, or apparatus that comprises a list of elements does not include only those elements but may include other elements not expressly listed or inherent to such chemical, chemical composition, process, method, article, or apparatus.

CLAIMS

What is claimed is:

1. A computer-implemented method comprising:

a) receiving a set of training data, the training data comprising one or more of:

(i) crystal structure data of a target receptor;

(ii) a predicted three-dimensional structure of a target receptor; and/or

(iii) a chemical structure of a known ligand for the target receptor;

b) training, using the set of training data, a model to generate, in response to input, output comprising a plurality of molecular structures having a predictable binding affinity for a target receptor, wherein the model comprises a ligand encoder, a protein encoder, and a ligand decoder;

c) generating, based on (i) and (ii), a prior molecular density grid representing prior information related to the atomic density of (i) and (ii), wherein the atomic density comprises atomic occupancy data and atomic features data, wherein atomic occupancy channels in the prior molecular density grid are sampled at full resolution and atomic feature channels in the prior molecular density grid are downsampled using a downsampling layer;

d) generating, based on (iii), a posterior molecular density grid representing posterior information related to the atomic density of (iii), wherein the atomic density comprises atomic occupancy data and atomic features data, wherein atomic occupancy channels in the prior molecular density grid are sampled at full resolution and atomic feature channels in the prior molecular density grid are downsampled using a downsampling layer;

e) generating, using the model, a plurality of candidate ligands by:

i) inputting a target protein structure into the protein encoder;

ii) inputting a randomly selected set of values from the prior molecular density grid into the ligand encoder; and

iii) generating, using the protein encoder and the ligand encoder, a first plurality of molecular density grids in a hierarchical latent space, representing one or more encodings of the target protein structure and one or more encodings of the randomly selected set of values; and

concurrently:

i) inputting the target protein structure into the protein encoder a second time;

ii) inputting a randomly selected set of values from the posterior molecular density grid into the ligand encoder; and

iii) generating, using the protein encoder and the ligand encoder, a second plurality of molecular density grids in the hierarchical latent space, representing one or more encodings of the target protein structure and one or more encodings of the randomly selected set of values from the previously encoded ligand; and, in combination

f) generating, based on the first and second plurality of molecular density grids, using the ligand decoder, an output molecular density grid representing atomic densities of the candidate ligand, wherein the atomic density comprises atomic occupancy data and atomic features data, wherein the atomic occupancy data from the first and second plurality of molecular density gradients is sampled at full resolution and atomic feature data from the first and second plurality of molecular density gradients is up-sampled to full resolution;

g) iteratively placing atomic coordinates within the output molecular density grid;

h) assigning atomic features to each atom using density from the upsampled atomic feature data; and

i) fitting chemical bonds between the assigned atom coordinates based on the geometry and features of the atomic density data, thereby generating a molecular structure having a predictable binding affinity for the target receptor.

2. The method of claim 1, wherein the hierarchal latent space further comprises a set of tensors, each tensor spatially representing a reference ligand-receptor complex latent representation formed from at least one existing ligand and the target receptor or representing a reference ligand-receptor complex formed from a molecular structure thus generated and the target receptor.

3. The method of claim 1, wherein generating the output molecular density grid further comprises a prior-posterior sampling process further comprising interpolating toward a randomly sampled set of tensors in the latent space beginning from a reference ligand-receptor complex sampled from a latent distribution, said prior-posterior sampling configured to optimize a binding affinity of a molecular structure thus generated by said method and the target receptor.

4. The method of claim 3, wherein said prior-posterior sampling configured to optimize the binding affinity further comprises a decoding of the interpolation into new molecular densities.

5. The method of claim 1, wherein iteratively placing atomic coordinates within the output molecular density grid further comprises application of an atom-fitting algorithm configured to iteratively place atomic coordinates for a chosen atom type within the output molecular density grid to minimize a residual occupancy density.

6. The method of claim 1, wherein the fitting of chemical bonds between the assigned atom types within the atomic occupancy densities comprises application of a bond fitting algorithm.

7. The method of claim 1, wherein in generating the prior molecular density grid, each atom of the ligand is represented by a multi-channel pseudo-Gaussian density with channels, each channel representing a different atomic feature of each atom.

8. The method of claim 7, wherein the atomic features comprise the elements C, N, O, F, P, S, Cl, Br, and I, and combinations thereof, and the features of each element comprise status as H-bond donor, H-bond acceptor, aromaticity, positive (+) charge, negative (-) charge, or neutral (0) charge.

9. The method of any one of claims 1 to 8, wherein at least one of the prior molecular density grid, the posterior molecular density grid, the first plurality of molecular density grids, second plurality of molecular density grids, and the output molecular density grid measures 24 x 24 x 24 Angstroms and the molecular density grid can be an input and/or output data structure.

10. The method of any one of claims 1 to 9, wherein at least one of the prior molecular density grid, the posterior molecular density grid, the first plurality of molecular density grids, second plurality of molecular density grids, and the output molecular density grid measures is a

molecular graph or a point cloud or point set where the molecular graph is comprised of a set of atomic coordinates (vertices), bonds (edges), atom (vertex) types, and bond (edge) types and the point cloud or point set is a graph without edges or edge types.

11. A system comprising:

one or more computing devices associated with a processor and a memory for executing computer-executable instructions configured to:

a) receive a set of training data, the training data comprising one or more of:

(i) crystal structure data of a target receptor;

(ii) a predicted three-dimensional structure of a target receptor; and/or

(iii) a chemical structure of a known ligand for the target receptor;

b) train, using the set of training data, a model to generate, in response to input, output comprising a plurality of molecular structures having a predictable binding affinity for a target receptor, wherein the model comprises a ligand encoder, a protein encoder, and a ligand decoder;

c) generate, based on (i) and (ii), a prior molecular density grid representing prior information related to the atomic density of (i) and (ii), wherein the atomic density comprises atomic occupancy data and atomic features data, wherein atomic occupancy channels in the prior molecular density grid are sampled at full resolution and atomic feature channels in the prior molecular density grid are downsampled using a downsampling layer;

d) generate, based on (iii), a posterior molecular density grid representing posterior information related to the atomic density of (iii), wherein the atomic density comprises atomic occupancy data and atomic features data, wherein atomic occupancy channels in the prior molecular density grid are sampled at full resolution and atomic feature channels in the prior molecular density grid are downsampled using a downsampling layer;

e) generate, using the model, a plurality of candidate ligands by:

i) inputting a target protein structure into the protein encoder;

ii) inputting a randomly selected set of values from the prior molecular density grid into the ligand encoder; and

iii) generating, using the protein encoder and the ligand encoder, a first plurality of molecular density grids in a hierarchical latent space, representing one or

more encodings of the target protein structure and one or more encodings of the randomly selected set of values; and

concurrently:

i) inputting the target protein structure into the protein encoder a second time;

ii) inputting a randomly selected set of values from the posterior molecular density grid into the ligand encoder; and

iii) generating, using the protein encoder and the ligand encoder, a second plurality of molecular density grids in the hierarchical latent space, representing one or more encodings of the target protein structure and one or more encodings of the randomly selected set of values from the previously encoded ligand; and, in combination

f) generate, based on the first and second plurality of molecular density grids, using the ligand decoder, an output molecular density grid representing atomic densities of the candidate ligand, wherein the atomic density comprises atomic occupancy data and atomic features data, wherein the atomic occupancy data from the first and second plurality of molecular density gradients is sampled at full resolution and atomic feature data from the first and second plurality of molecular density gradients is up-sampled to full resolution;

g) iteratively place atomic coordinates within the output molecular density grid;

h) assign atomic features to each atom using density from the upsampled atomic feature data; and

i) fit chemical bonds between the assigned atom coordinates based on the geometry and features of the atomic density data, thereby generating a molecular structure having a predictable binding affinity for the target receptor.

12. The system of claim 11, wherein the hierarchal latent space further comprises a set of tensors, each tensor spatially representing a reference ligand-receptor complex latent representation formed from at least one existing ligand and the target receptor or representing a reference ligand-receptor complex formed from a molecular structure thus generated and the target receptor.

13. The system of claim 11 or 12, wherein generating the output molecular density grid

further comprises a prior-posterior sampling process further comprising interpolating toward a randomly sampled set of tensors in the latent space beginning from a reference ligand-receptor complex sampled from a latent distribution, said prior-posterior sampling configured to optimize a binding affinity of a molecular structure thus generated and the target receptor.

14. The system of claim 11, 12, or 13, wherein said prior-posterior sampling configured to optimize the binding affinity further comprises a decoding of the interpolation into new molecular densities.

15. The system of any one of claims 11-14, wherein to iteratively place atomic coordinates within the output molecular density grid further comprises application of an atom-fitting algorithm configured to iteratively place atomic coordinates for a chosen atom type within the molecular density grid to minimize a residual occupancy density.

16. The system of any one of claims 11-15, wherein to fit chemical bonds between the assigned atom types within the atomic occupancy densities comprises application of a bond fitting algorithm.

17. The system of any one of claims 11-16, wherein in generating the prior molecular density grid, each atom of the ligand is represented by a multi-channel pseudo-Gaussian density with channels, each channel representing a different atomic feature of each atom.

18. The system of any one of claims 11-17, wherein the atomic features comprise the elements C, N, O, F, P, S, Cl, Br, and I, and combinations thereof, and the features of each element being H-bond donor, H-bond acceptor, aromaticity, positive (+) charge, negative (-) charge, or neutral (0) charge.

19. The system of any one of claims 11-18, wherein at least one of the prior molecular density grid, the posterior molecular density grid, the first plurality of molecular density grids, second plurality of molecular density grids, and the output molecular density grid is a molecular

graph or a point cloud or point set where the molecular graph is comprised of a set of atomic coordinates (vertices), bonds (edges), atom (vertex) types, and bond (edge) types and the point cloud or point set is a graph without edges or edge types.

20. A non-transitory computer-readable medium having instructions stored thereon that, when executed by one or more processors, cause the one or more processors to perform operations comprising:

a) receiving a set of training data, the training data comprising one or more of:

(i) crystal structure data of a target receptor;

(ii) a predicted three-dimensional structure of a target receptor; and/or

(iii) a chemical structure of a known ligand for the target receptor;

b) training, using the set of training data, a model to generate, in response to input, output comprising a plurality of molecular structures having a predictable binding affinity for a target receptor, wherein the model comprises a ligand encoder, a protein encoder, and a ligand decoder;

c) generating, based on (i) and (ii), a prior molecular density grid representing prior information related to the atomic density of (i) and (ii), wherein the atomic density comprises atomic occupancy data and atomic features data, wherein atomic occupancy channels in the prior molecular density grid are sampled at full resolution and atomic feature channels in the prior molecular density grid are downsampled using a downsampling layer;

d) generating, based on (iii), a posterior molecular density grid representing posterior information related to the atomic density of (iii), wherein the atomic density comprises atomic occupancy data and atomic features data, wherein atomic occupancy channels in the prior molecular density grid are sampled at full resolution and atomic feature channels in the prior molecular density grid are downsampled using a downsampling layer;

e) generating, using the model, a plurality of candidate ligands by:

i) inputting a target protein structure into the protein encoder;

ii) inputting a randomly selected set of values from the prior molecular density grid into the ligand encoder; and

iii) generating, using the protein encoder and the ligand encoder, a first plurality of molecular density grids in a hierarchical latent space, representing one or

more encodings of the target protein structure and one or more encodings of the randomly selected set of values; and

concurrently:

i) inputting the target protein structure into the protein encoder a second time;

ii) inputting a randomly selected set of values from the posterior molecular density grid into the ligand encoder; and

iii) generating, using the protein encoder and the ligand encoder, a second plurality of molecular density grids in the hierarchical latent space, representing one or more encodings of the target protein structure and one or more encodings of the randomly selected set of values from the previously encoded ligand; and, in combination

f) generating, based on the first and second plurality of molecular density grids, using the ligand decoder, an output molecular density grid representing atomic densities of the candidate ligand, wherein the atomic density comprises atomic occupancy data and atomic features data, wherein the atomic occupancy data from the first and second plurality of molecular density gradients is sampled at full resolution and atomic feature data from the first and second plurality of molecular density gradients is up-sampled to full resolution;

g) iteratively placing atomic coordinates within the output molecular density grid;

h) assigning atomic features to each atom using density from the upsampled atomic feature data; and

i) fitting chemical bonds between the assigned atom coordinates based on the geometry and features of the atomic density data, thereby generating a molecular structure having a predictable binding affinity for the target receptor.

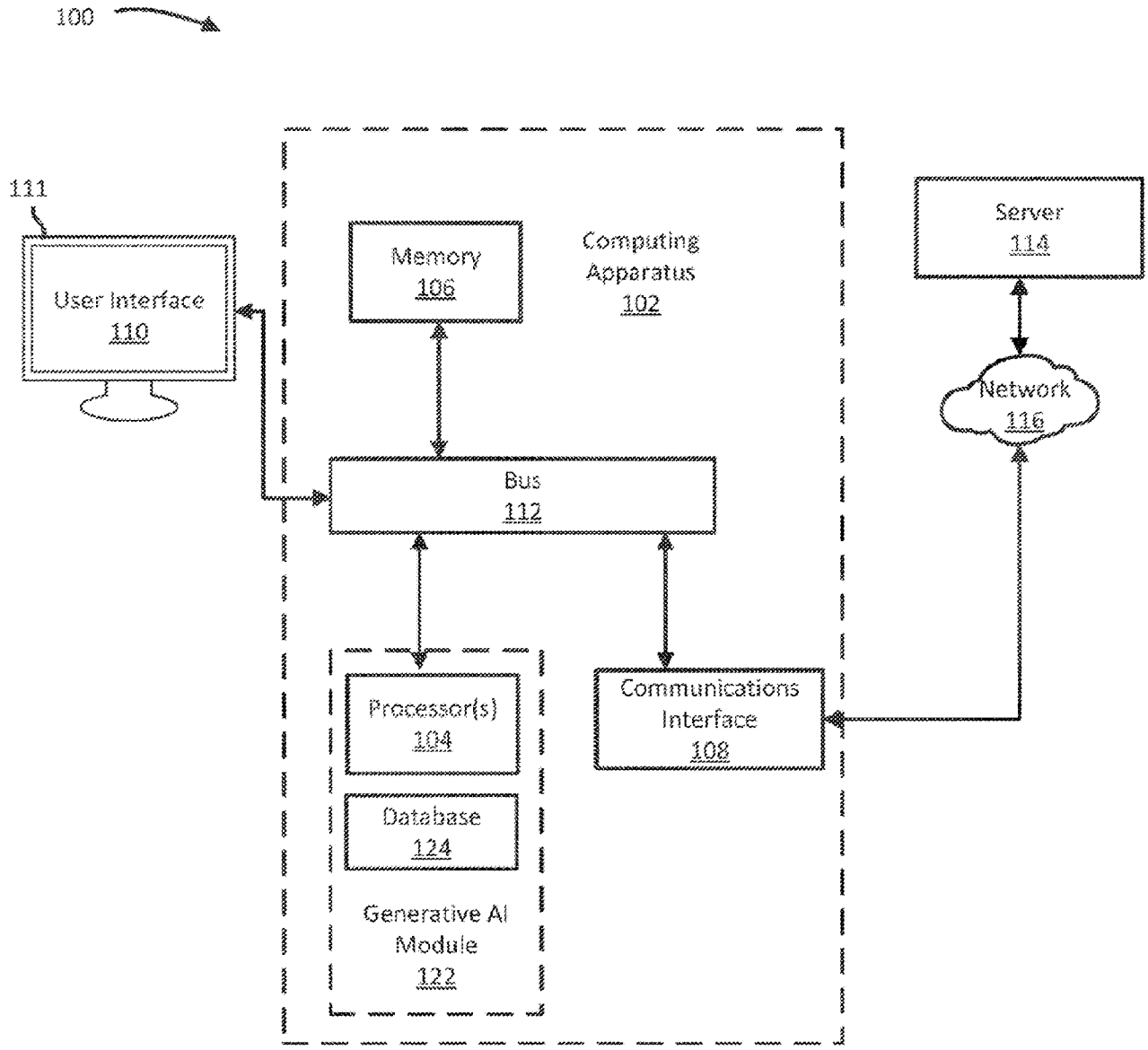
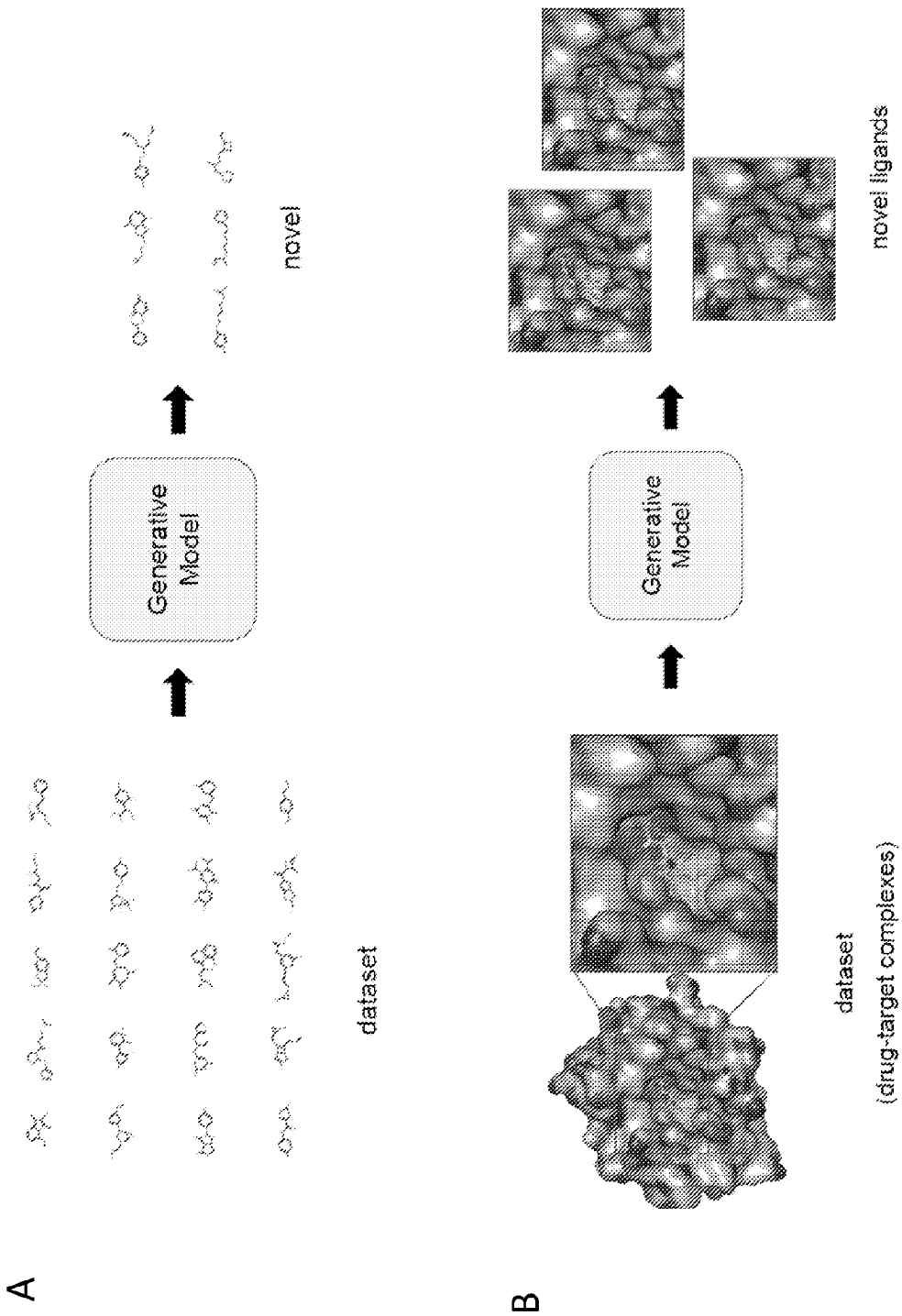
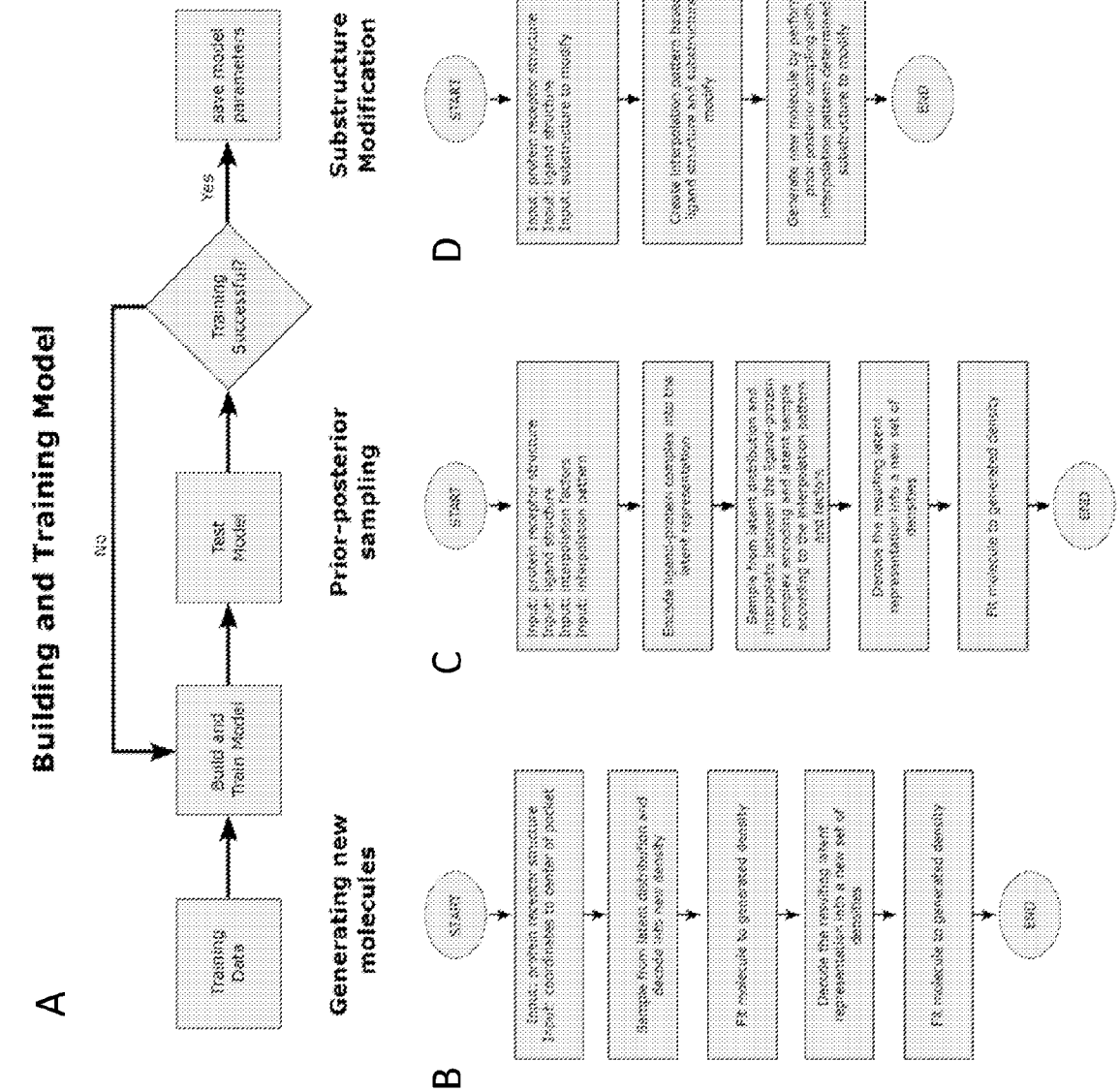

100
111
User Interface
110
Memory
106
Computing
Apparatus
102
Server
114
Network
116
Bus
112
Processor(s)
104
Database
124
Generative AI
Module
122
Communications
Interface
108

Figure 1

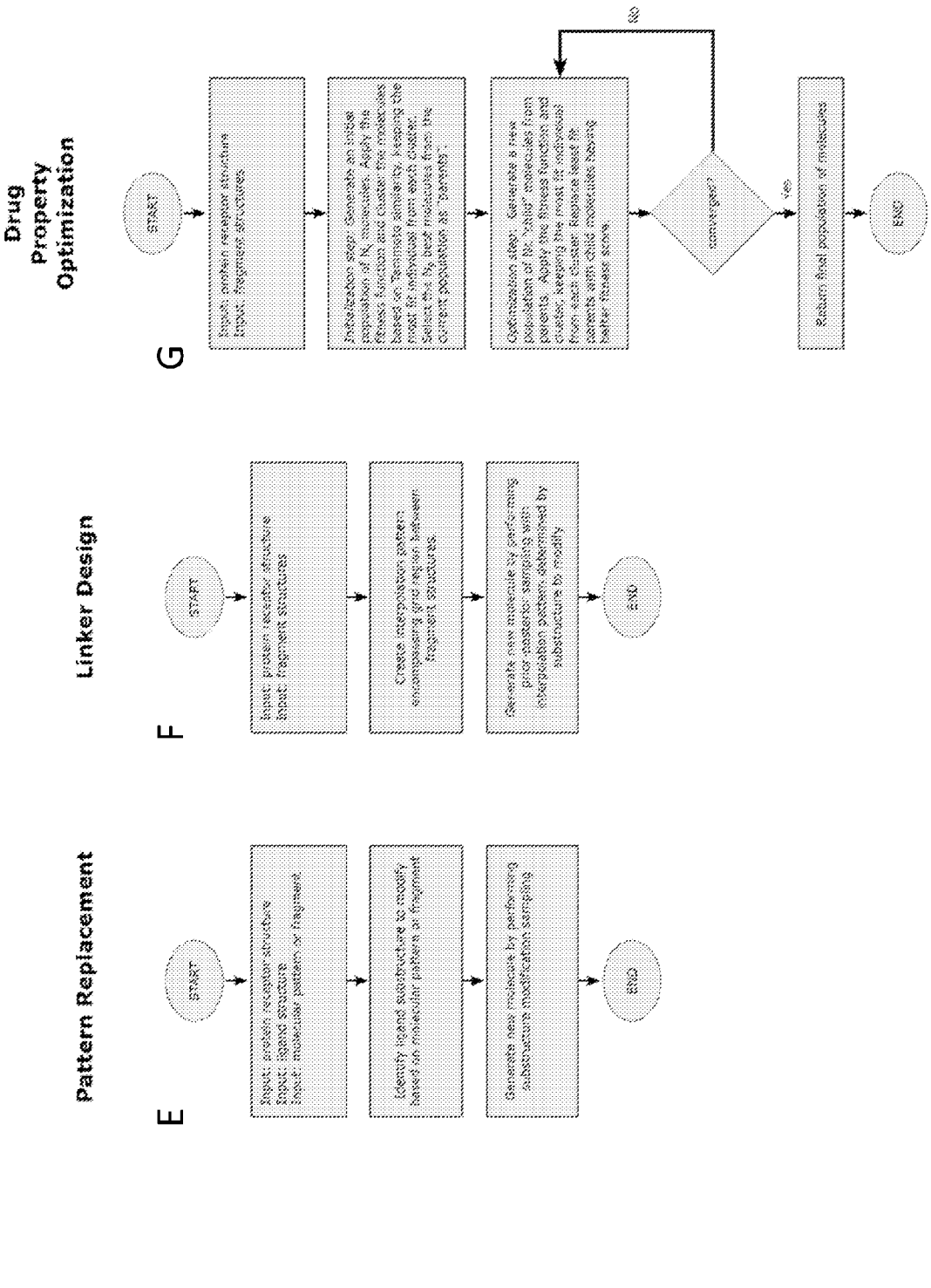
Figures 2A-2B



Figures 3A-3D



Figures 3E-3G



Figures 4A-4G

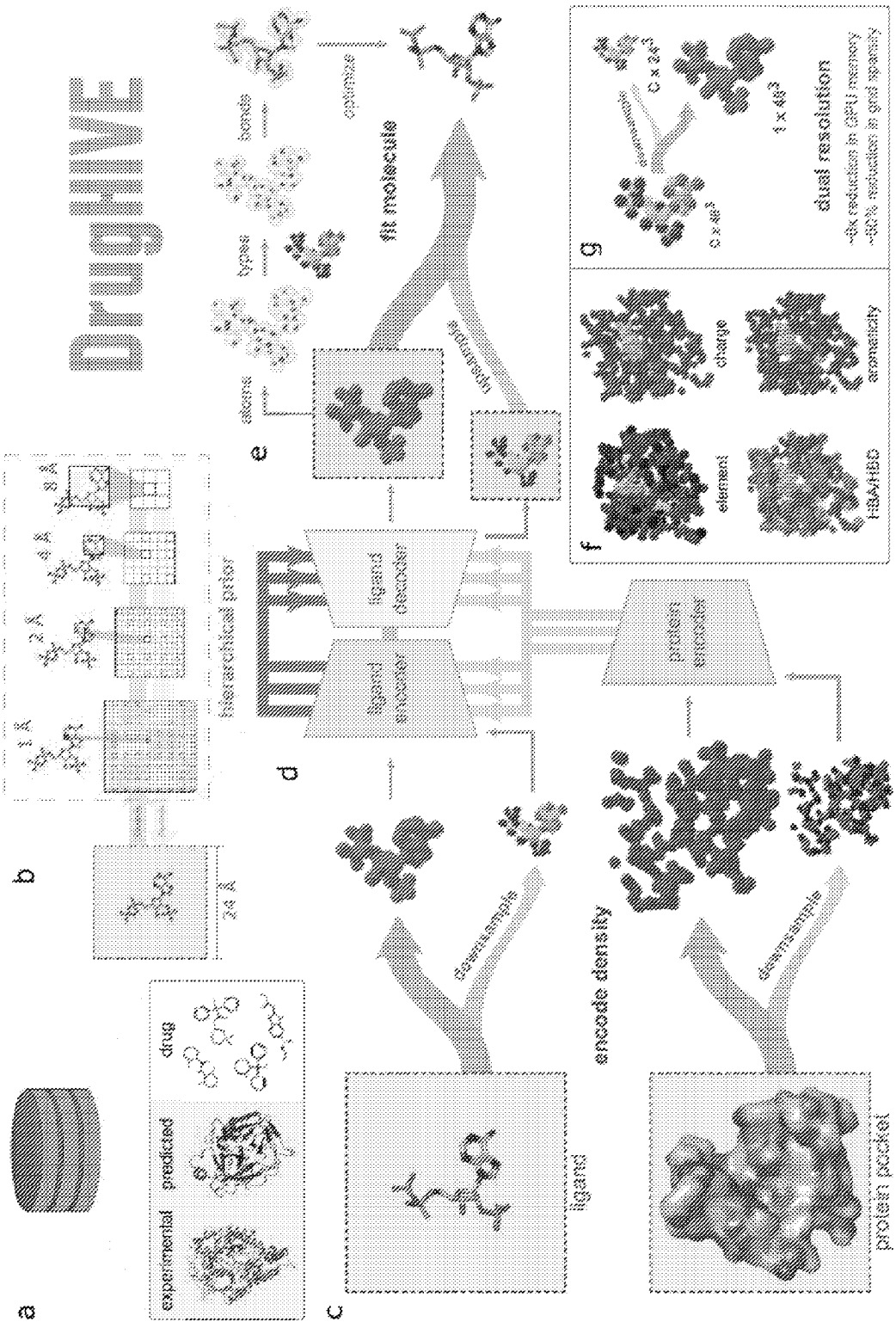
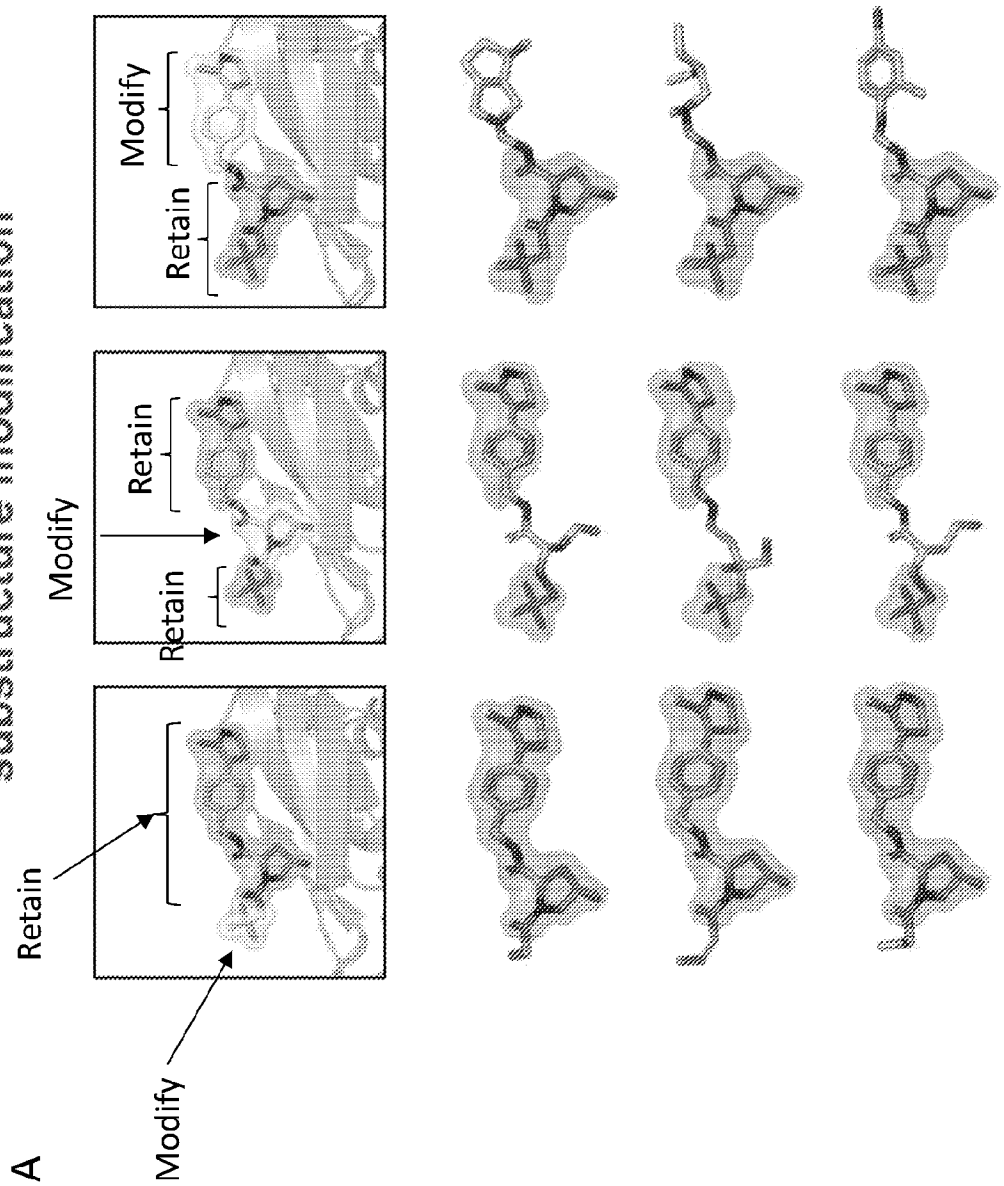
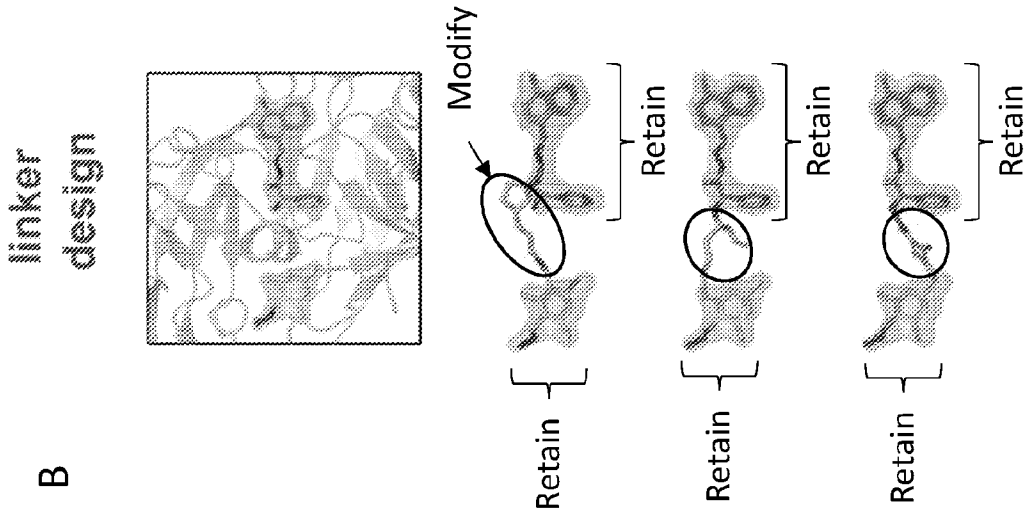


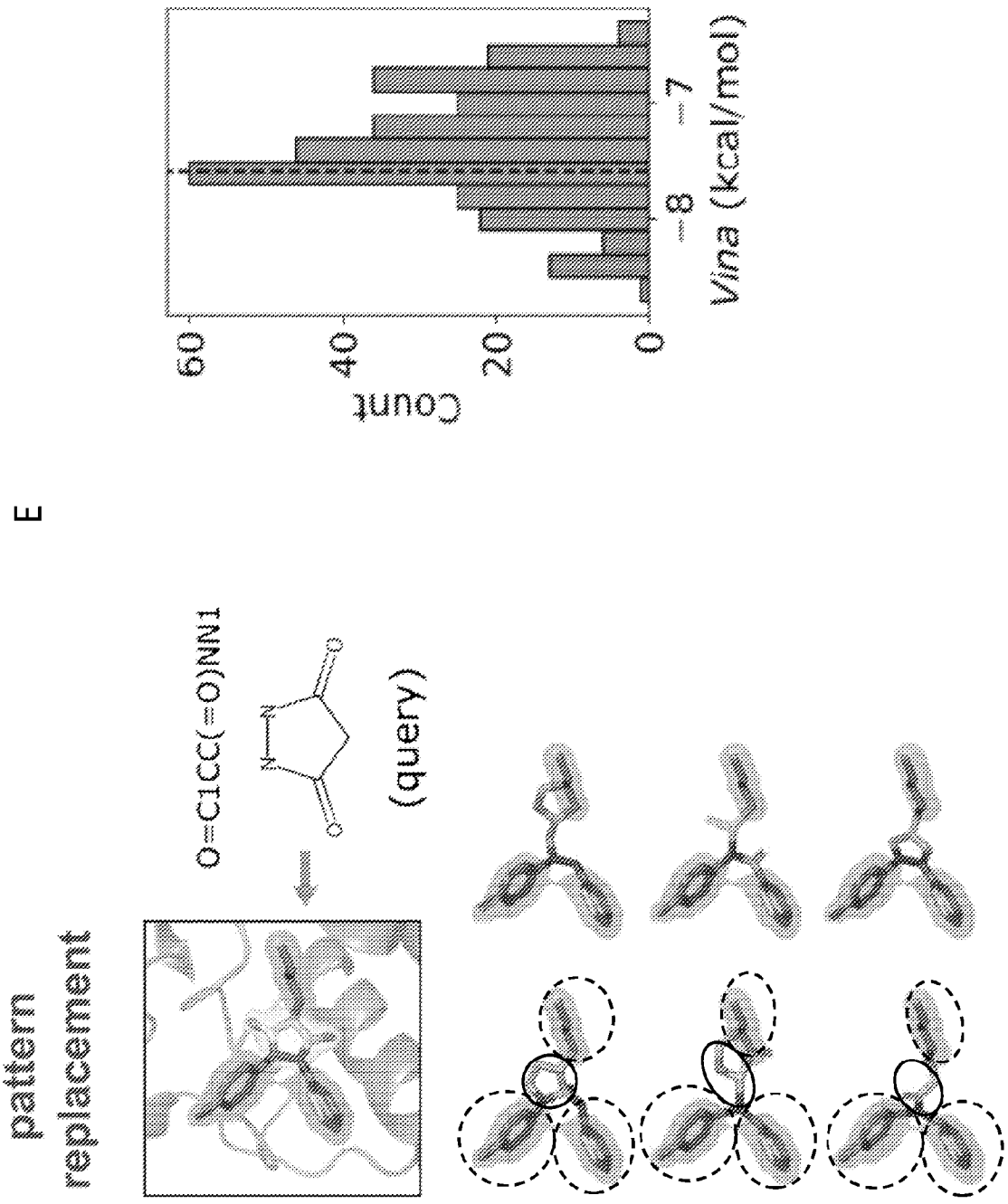
Figure 5A


A
substructure modification
Retain
Modify
Modify
Retain
Retain
Modify
Retain

Figures 5B-5C



Figures 5D-5E



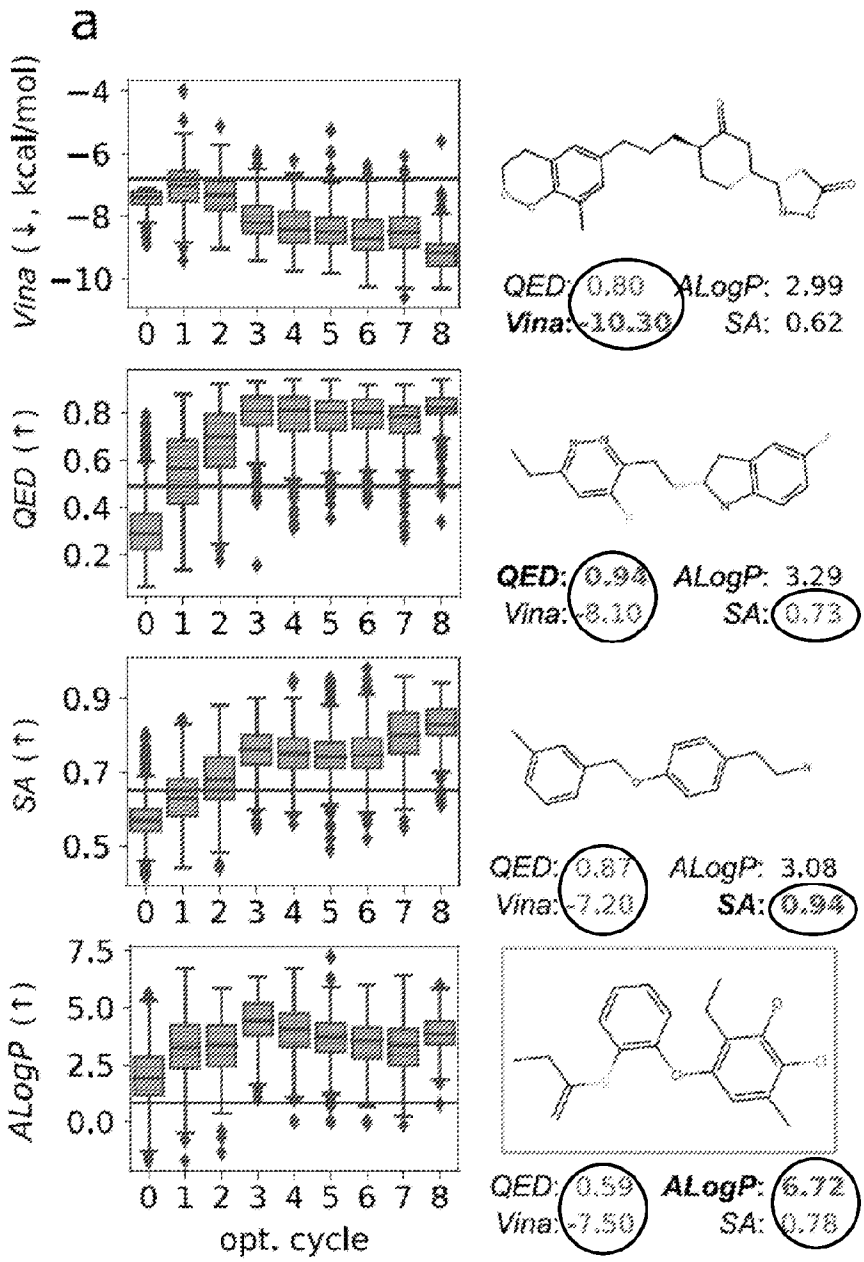
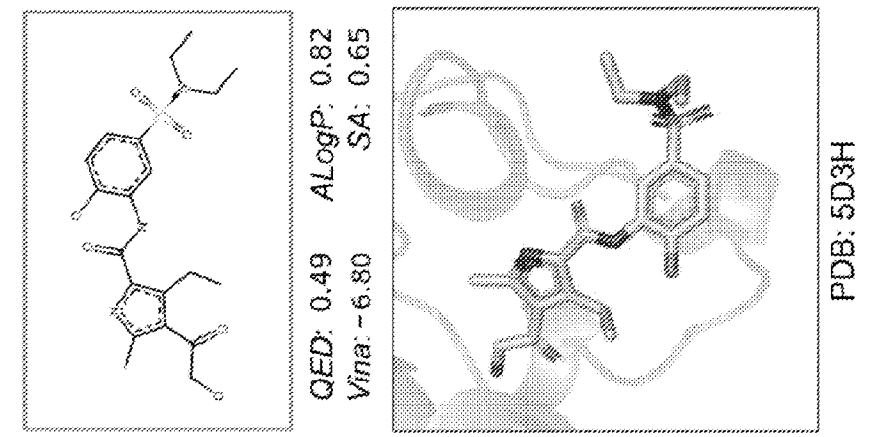
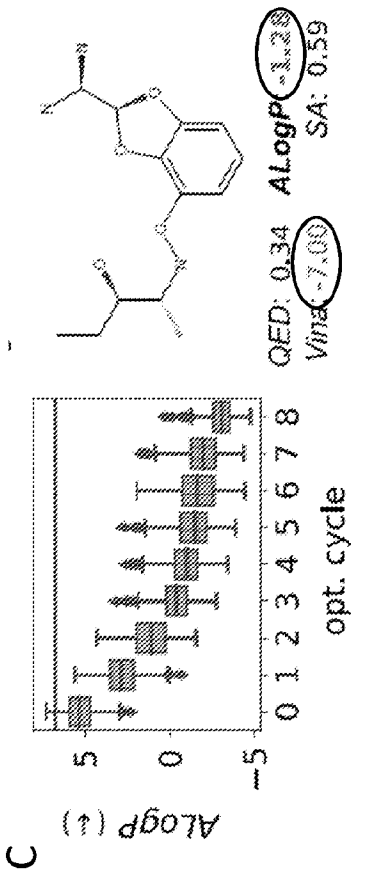
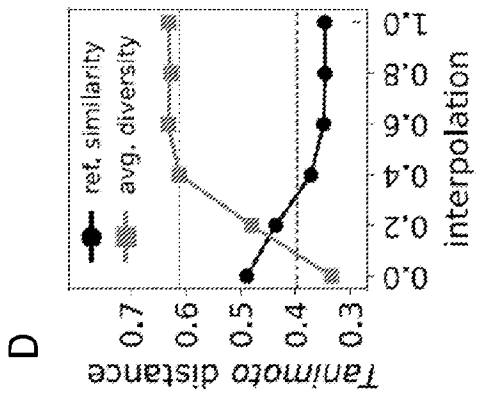

a
Vina (↓, kcal/mol)
−4
−6
−8
−10
0 1 2 3 4 5 6 7 8
QED (↑)
0.8
0.6
0.4
0.2
SA (↑)
0.9
0.7
0.5
ALogP (↑)
7.5
5.0
2.5
0.0
opt. cycle
QED: 0.80 ALogP: 2.99
Vina: -10.30 SA: 0.62
QED: 0.94 ALogP: 3.29
Vina: -8.10 SA: 0.73
QED: 0.87 ALogP: 3.08
Vina: -7.20 SA: 0.94
QED: 0.59 ALogP: 6.72
Vina: -7.50 SA: 0.78

Figure 6A

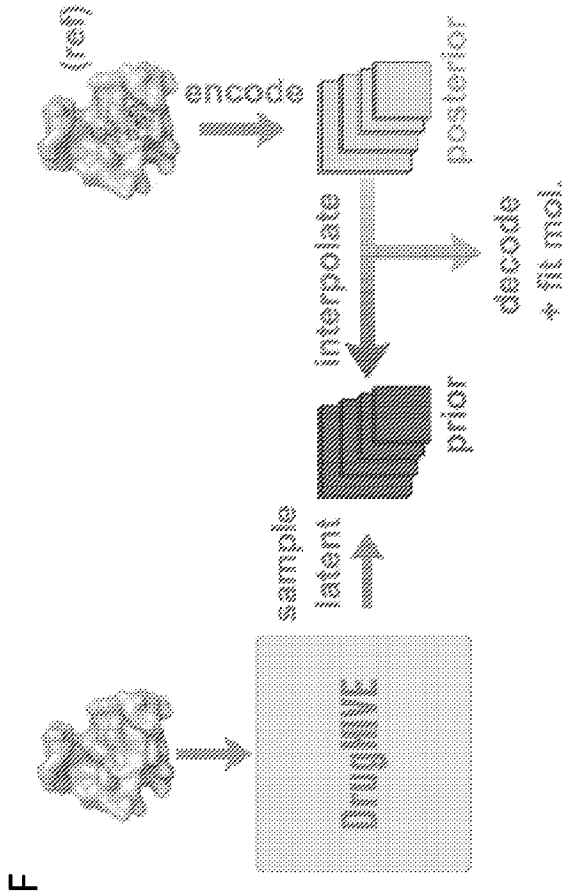
Figures 6B-6D



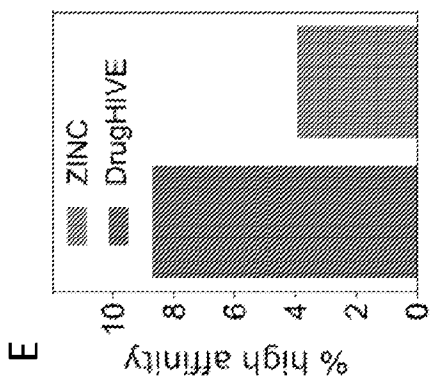




Figures 6E-6F



F



E

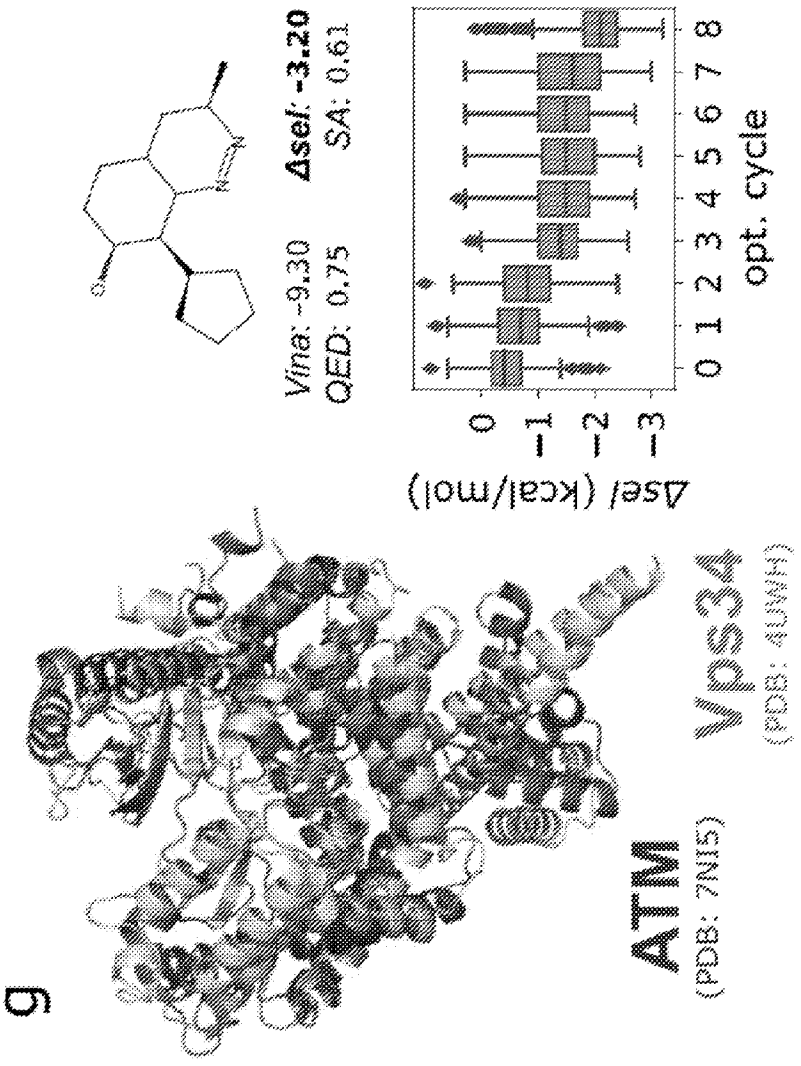

9
ATM
(PDB: 7NIS)
Vps34
(PDB: 4UWH)
Δsel (kcal/mol)
0
−1
−2
−3
0 1 2 3 4 5 6 7 8
opt. cycle
Vina: −9.30
QED: 0.75
Δsel: −3.20
SA: 0.61

Figure 6G

H

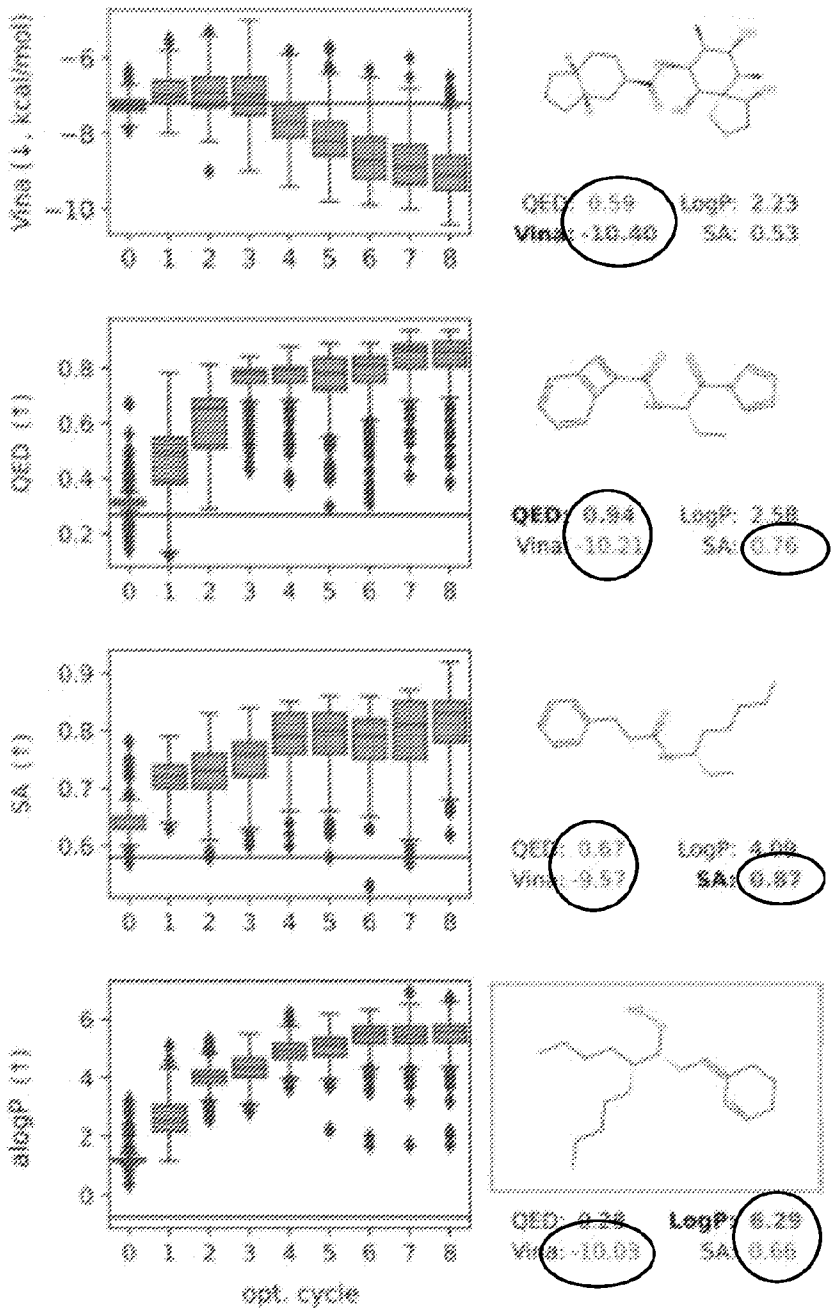
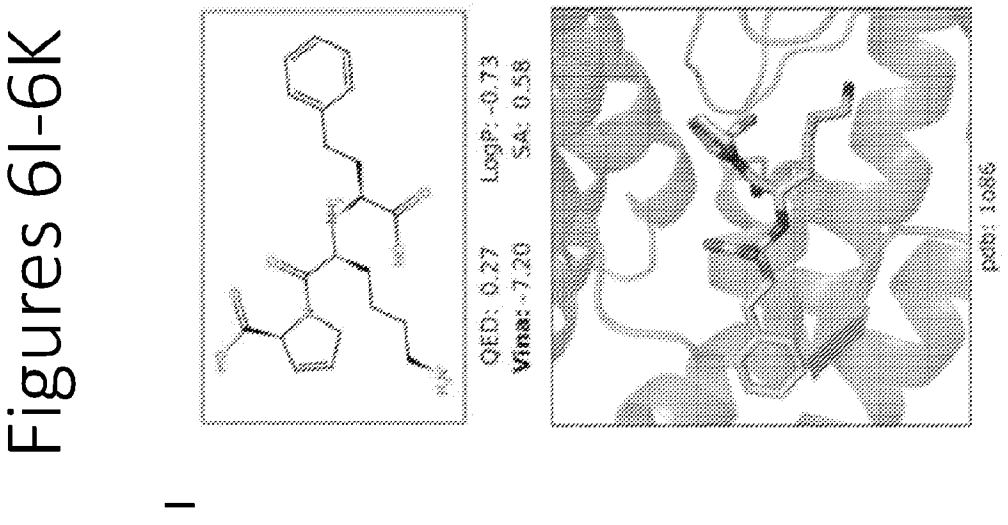
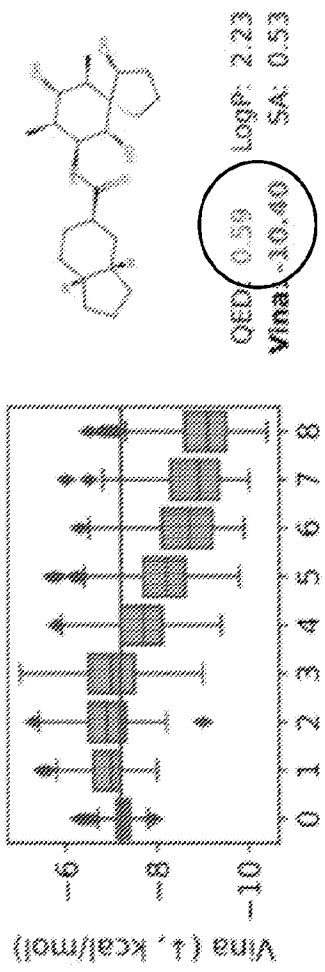

Vina (↓, kcal/mol)
QED (↑)
SA (↑)
algP (↑)
opt. cycle
QED: 0.59
LogP: 2.23
Vina: -10.40
SA: 0.53
QED: 0.94
LogP: 2.58
Vina: -10.21
SA: 0.76
QED: 0.61
Vina: -9.57
SA: 0.87
QED: 0.36
LogP: 6.29
Vina: -10.03
SA: 0.66

Figure 6H

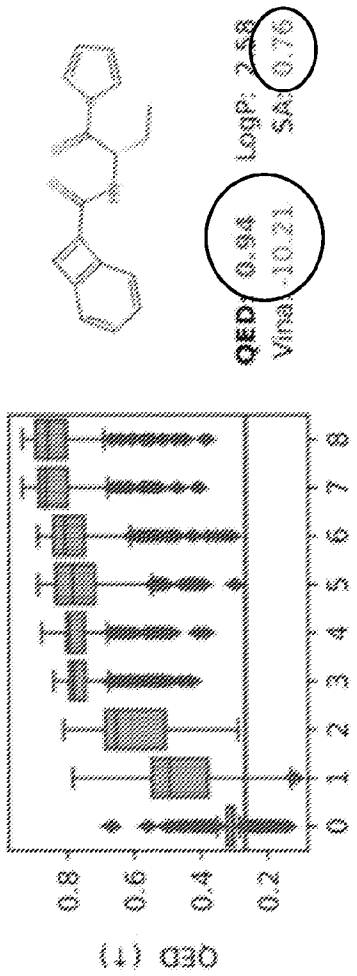
Figures 6I-6K



I



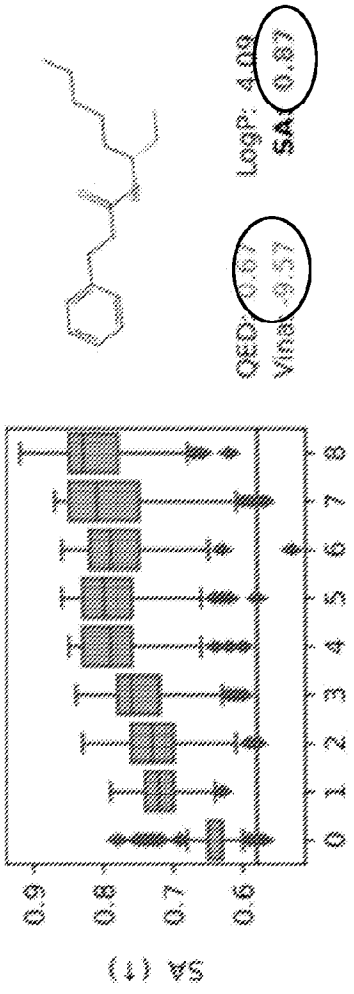
J



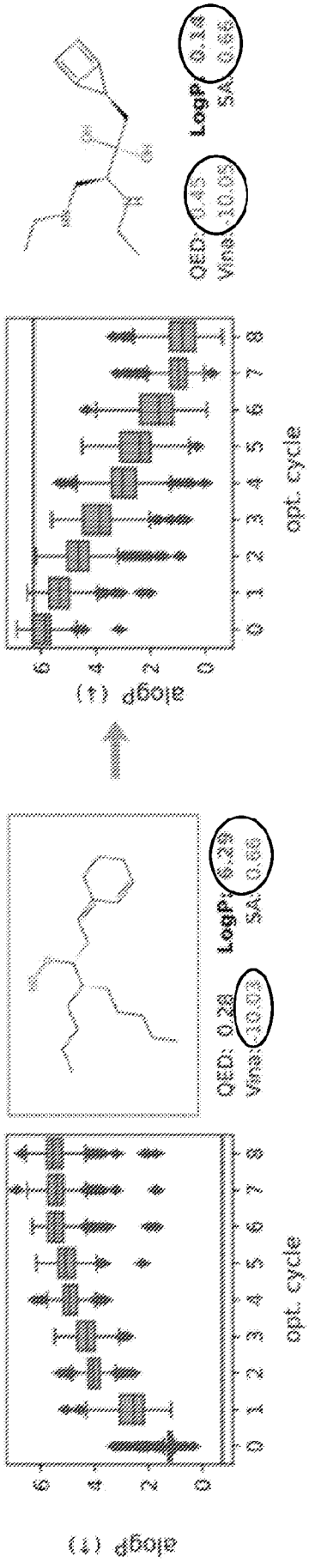
K

Figures 6L-6M

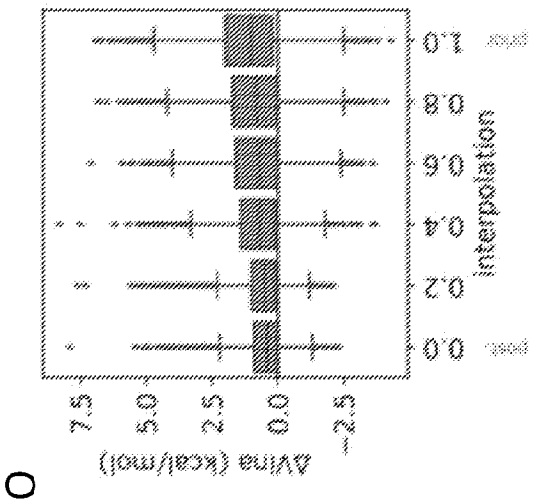
L



M



Figures 6N-6O



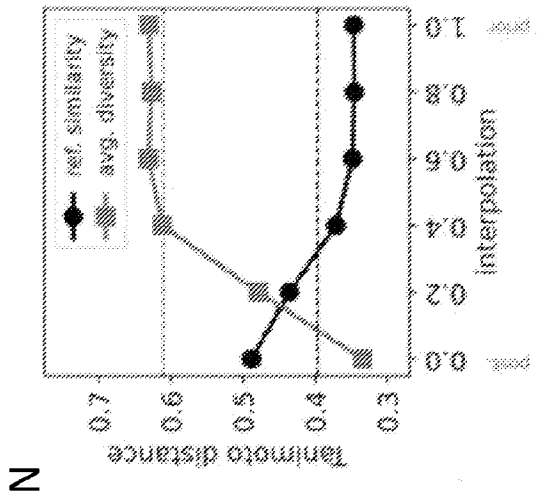
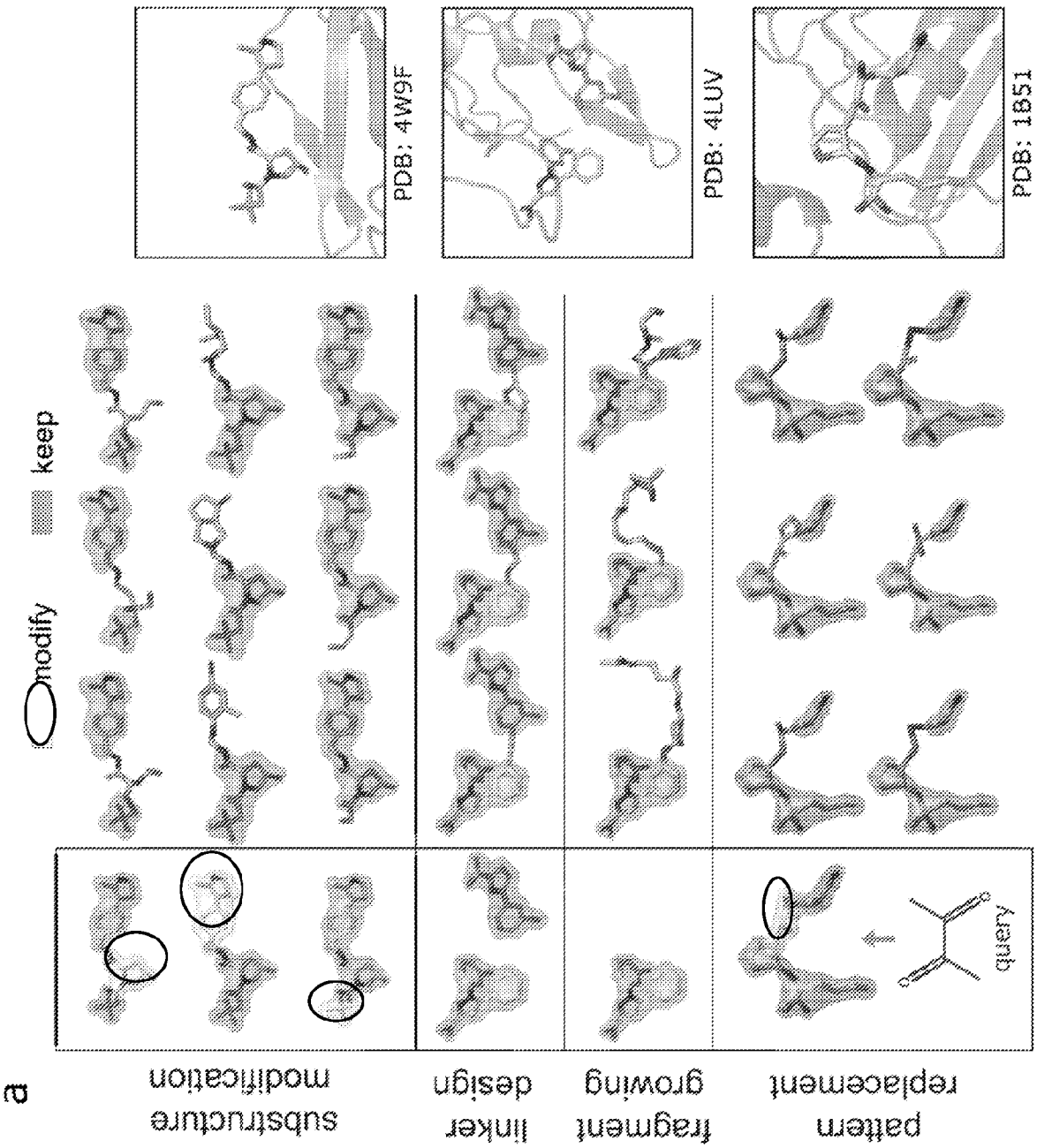


Figure 7A


a
modify
keep
query
pattern replacement
fragment growing
linker design
substructure modification
PDB: 1B51
PDB: 4LUV
PDB: 4W9F

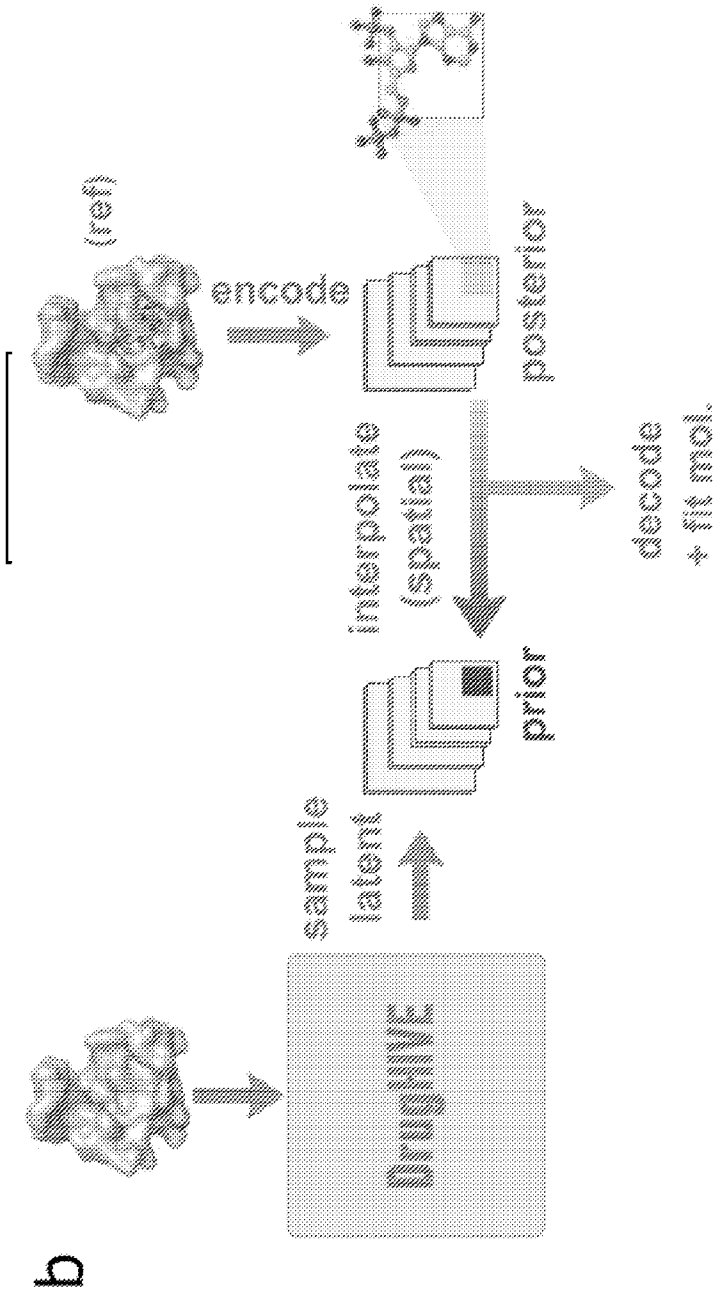
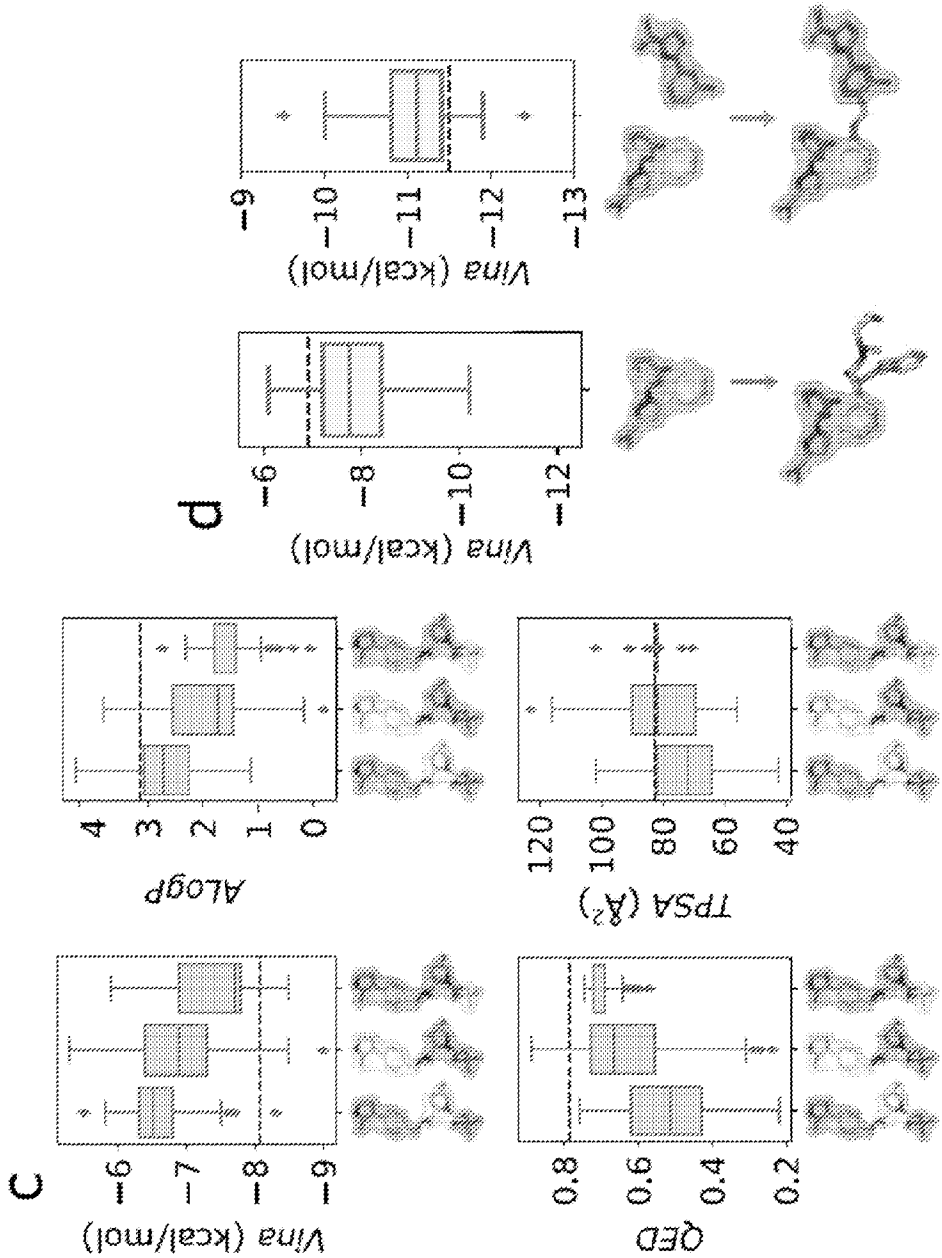
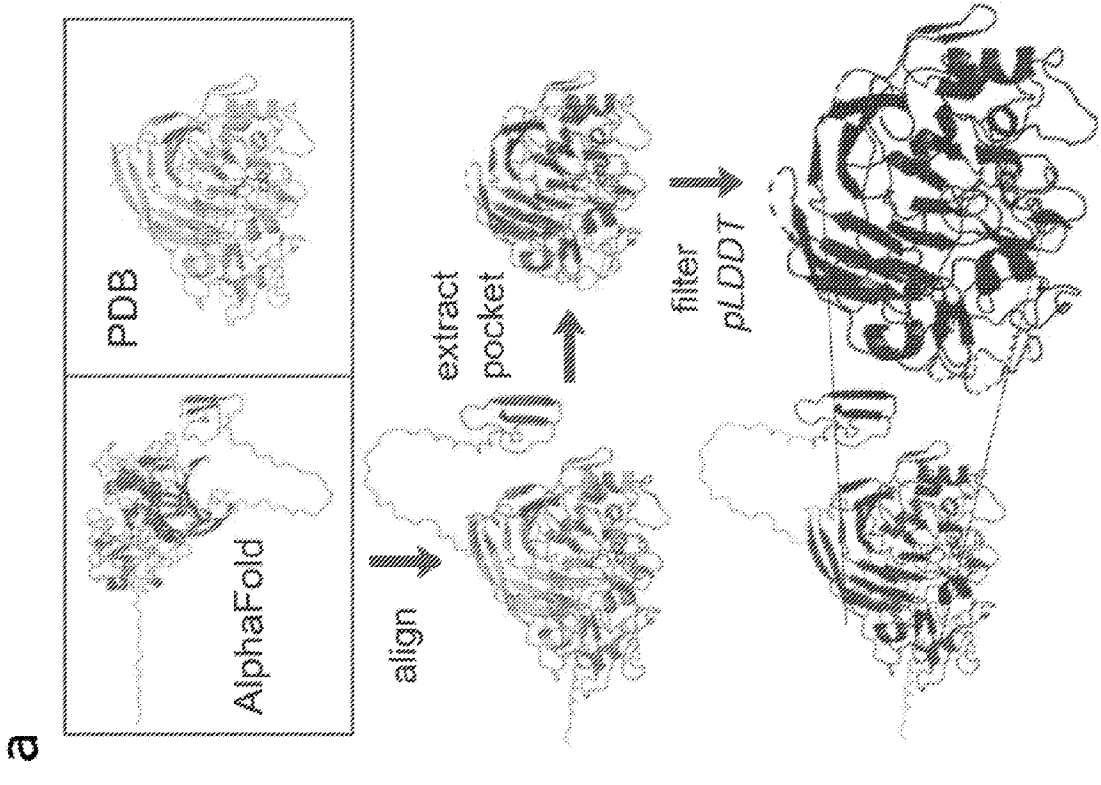

b
drugMNE
sample
latent
prior
(interpolate
(spatial)
posterior
encode
(ref)
decode
+ fit mol.

Figure 7B

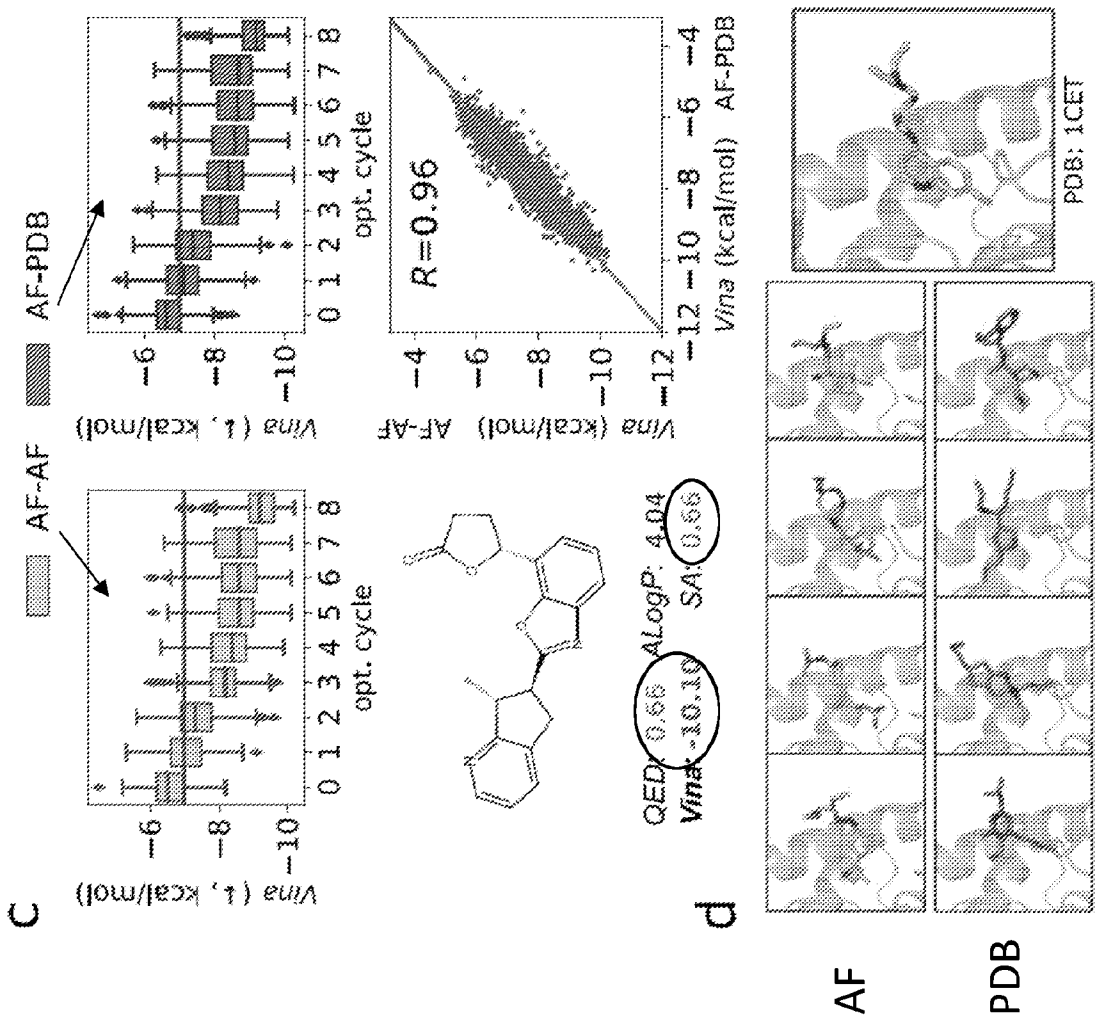
Figure 7C-7D


c
Vina (kcal/mol)
−6
−7
−8
−9
QED
0.8
0.6
0.4
0.2
ALogP
4
3
2
1
0
TPSA (Å²)
120
100
80
60
40
d
Vina (kcal/mol)
−6
−8
−10
−12
Vina (kcal/mol)
−9
−10
−11
−12
−13

Figures 8A-8B



Figures 8C-8D



Figures 8E-8F

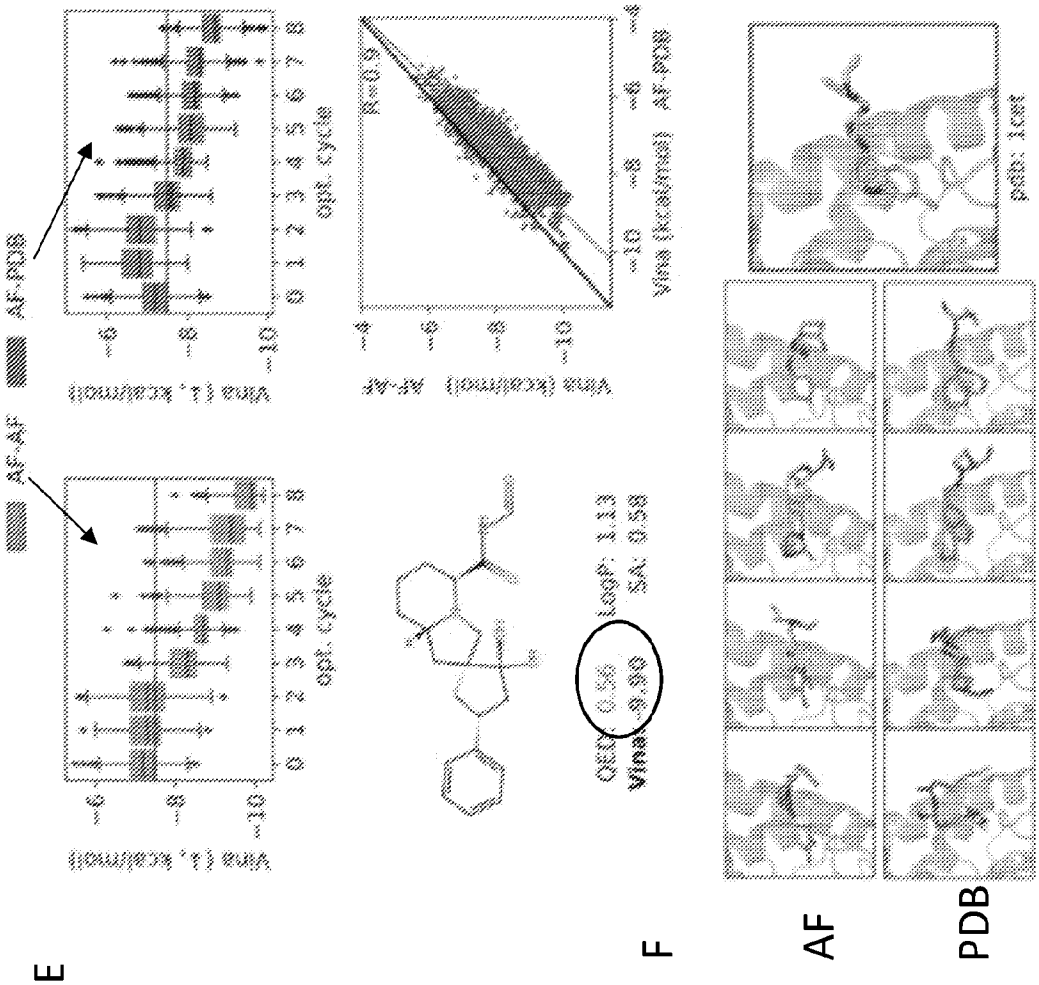
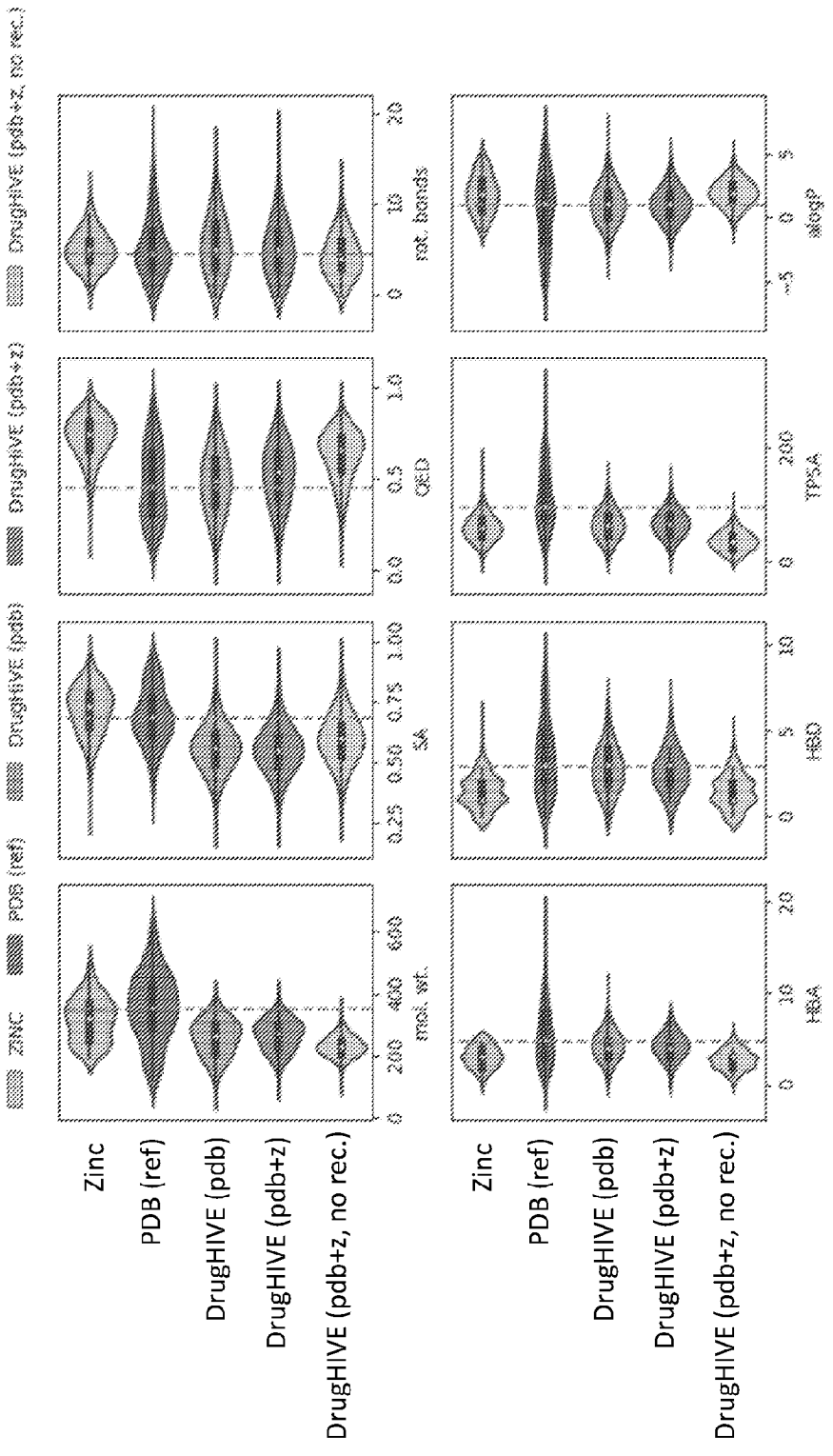
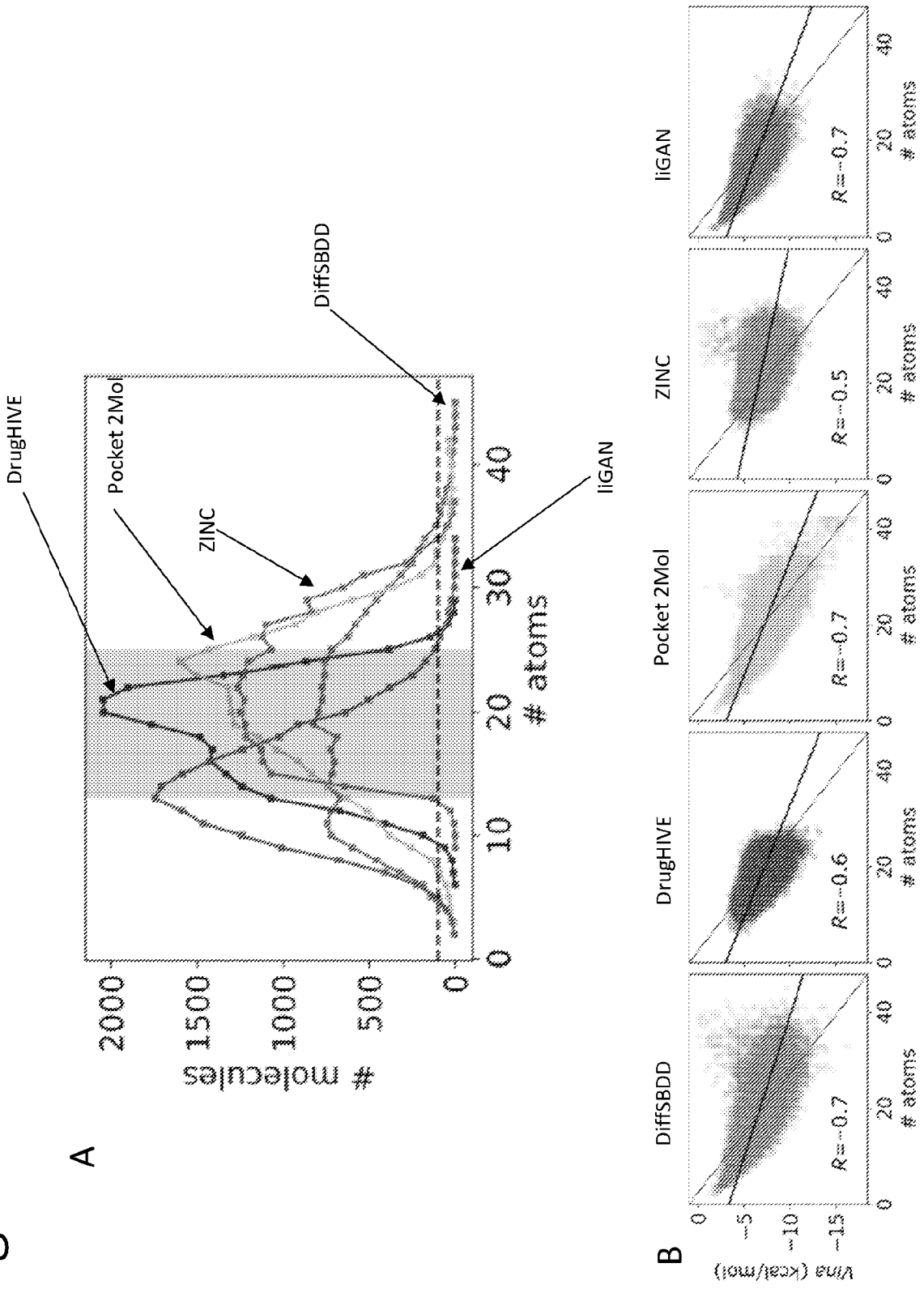


Figure 9

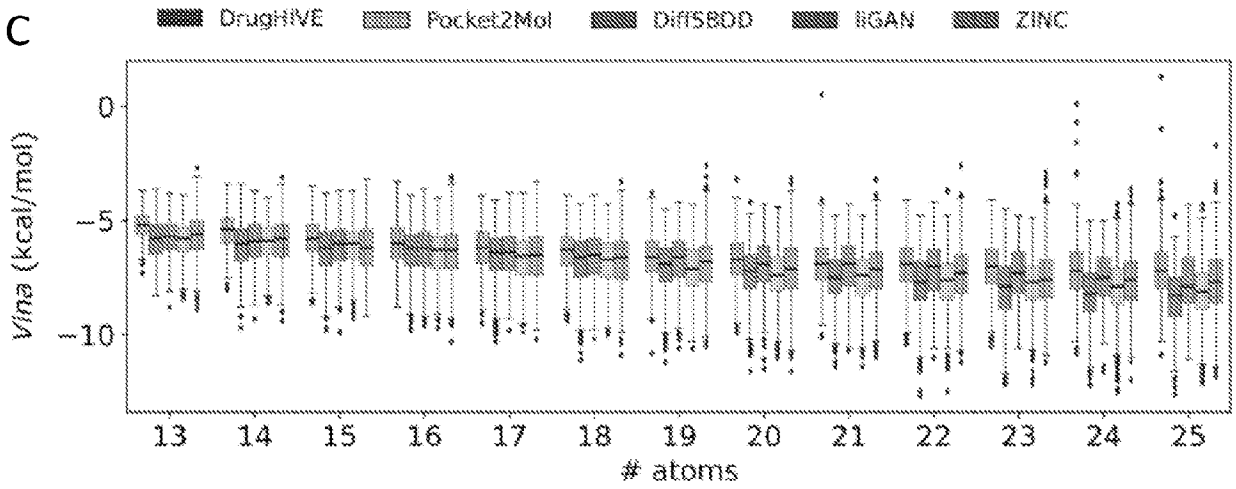

ZINC
PDB (ref)
DrugHIVE (pdb)
DrugHIVE (pdb+z)
DrugHIVE (pdb+z, no rec.)
Zinc
PDB (ref)
DrugHIVE (pdb)
DrugHIVE (pdb+z)
DrugHIVE (pdb+z, no rec.)
mol. wt.
SA
QED
rot. bonds
HBA
HBD
TPSA
xlogP

Figures 10A-10B



Figures 10C-10D

C

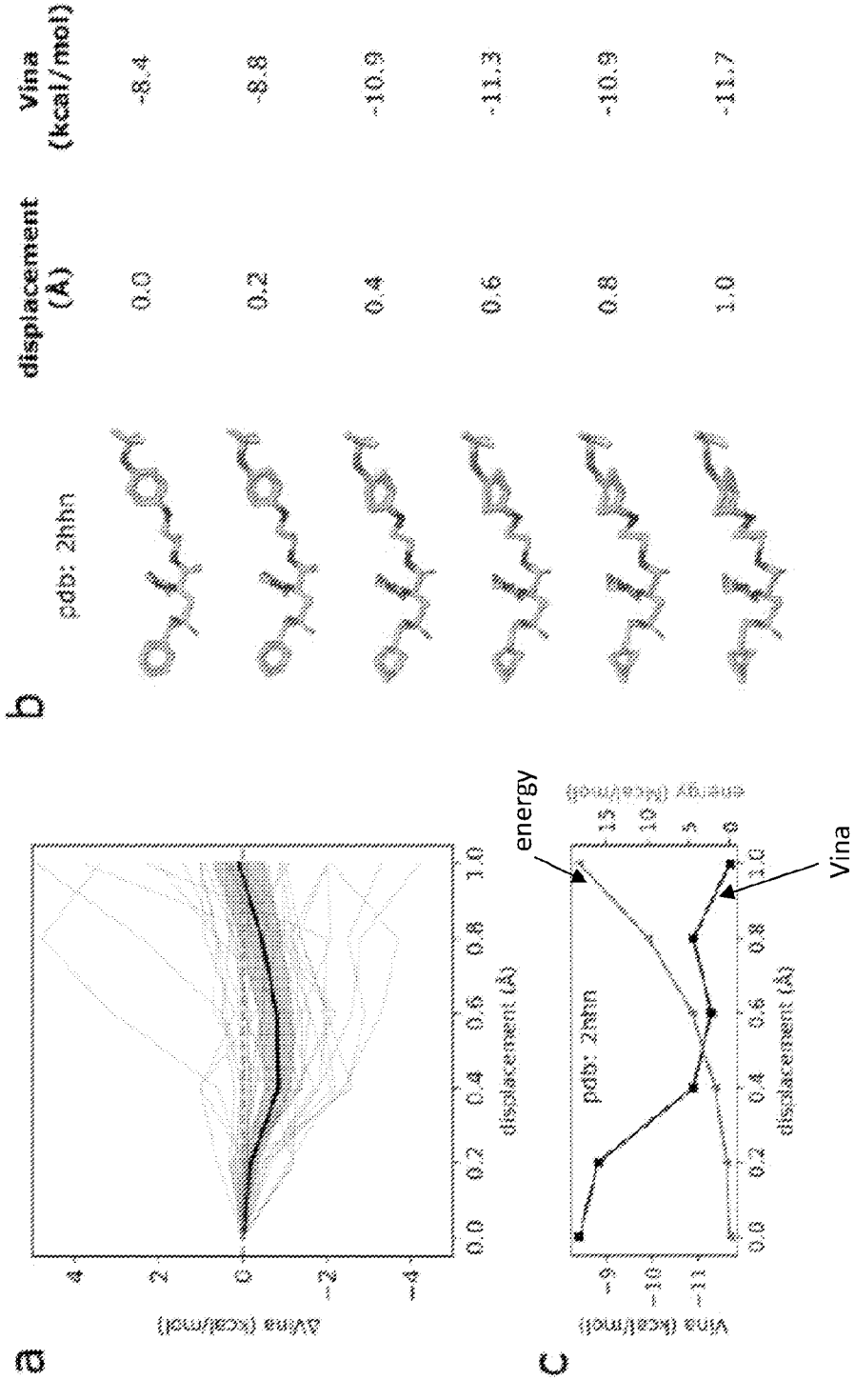


D

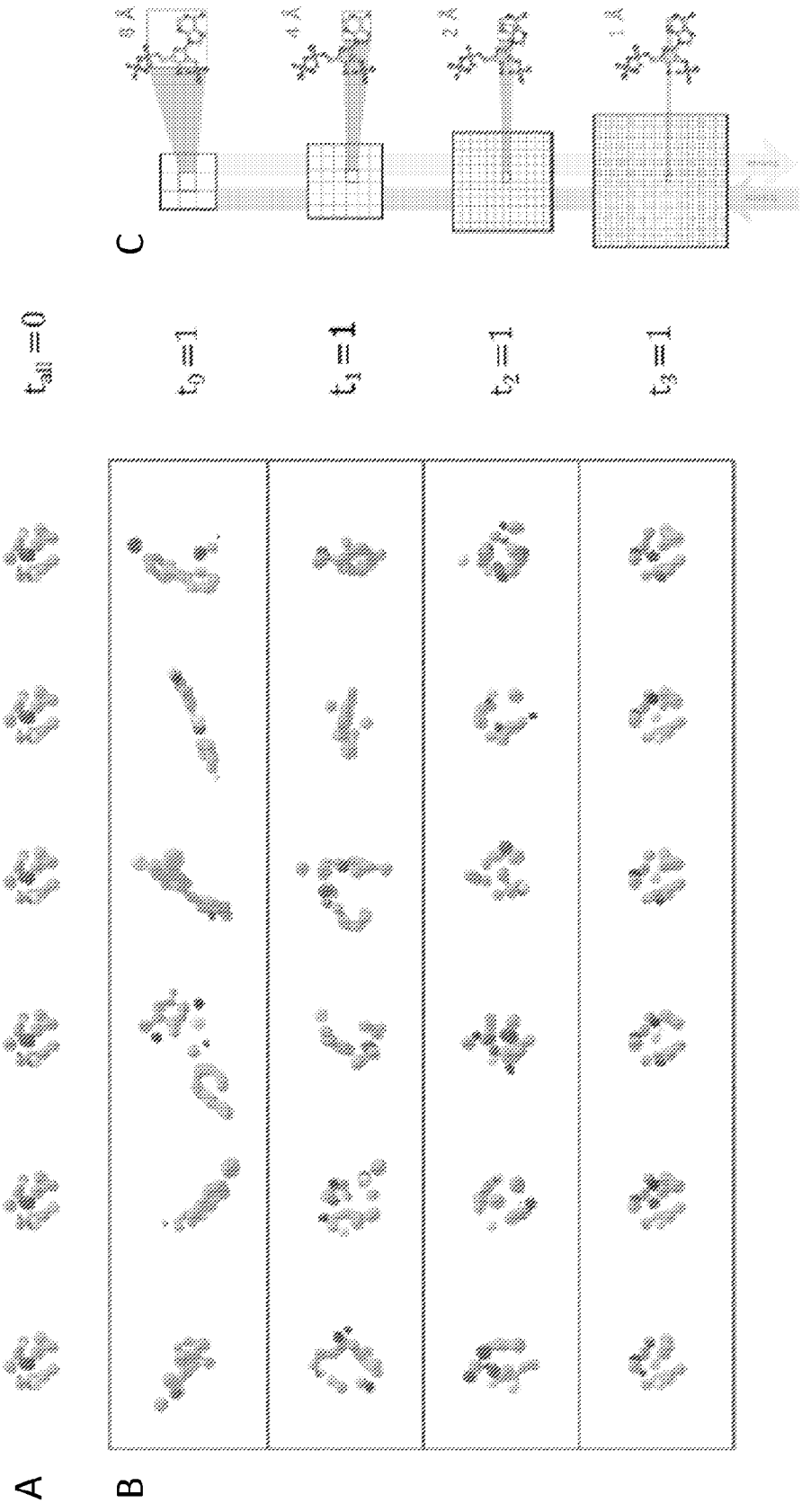
	Vina-n (↓)	*QED* (↑)	*SA* (↑)	Diversity
ref	−7.97 ± 1.98[b]	0.46 ± 0.19	0.60 ± 0.21	--
ZINC	−6.46 ± 0.06	0.72 ± 0.12	0.71 ± 0.11	0.66 ± 0.02
LiGAN (25)	−6.79 ± 0.06	0.46 ± 0.16	0.58 ± 0.12	0.67 ± 0.02
DiffSBDD (25)[c]	−6.83 ± 0.07	0.56 ± 0.17	0.68 ± 0.13	0.69 ± 0.03
Pocket2Mol (28)	−7.04 ± 0.07	0.58 ± 0.19	0.71 ± 0.13	0.63 ± 0.04
DrugHiVE	−7.08 ± 0.07	0.59 ± 0.15	0.56 ± 0.11	0.62 ± 0.02

[a]Summary values (mean ± std. dev./std. err.) are reported for comparable models, a random sample from the ZINC drug-like subset, and the test set crystal ligands (ref). *Vina-n* is the normalized *Vina* score. Diversity is reported as the average per-target molecular diversity. No diversity score is reported for the reference set (ref) as there is only one molecule per target.
[b]For crystal ligands (ref), *Vina* mean and std. dev. are reported instead of *Vina-n* mean and std. err.
[c]DiffSBDD-joint model trained on Binding MOAD data set is used for comparison.

Figures 11A-11C



Figures 12A-12C



Figures 13A-13B

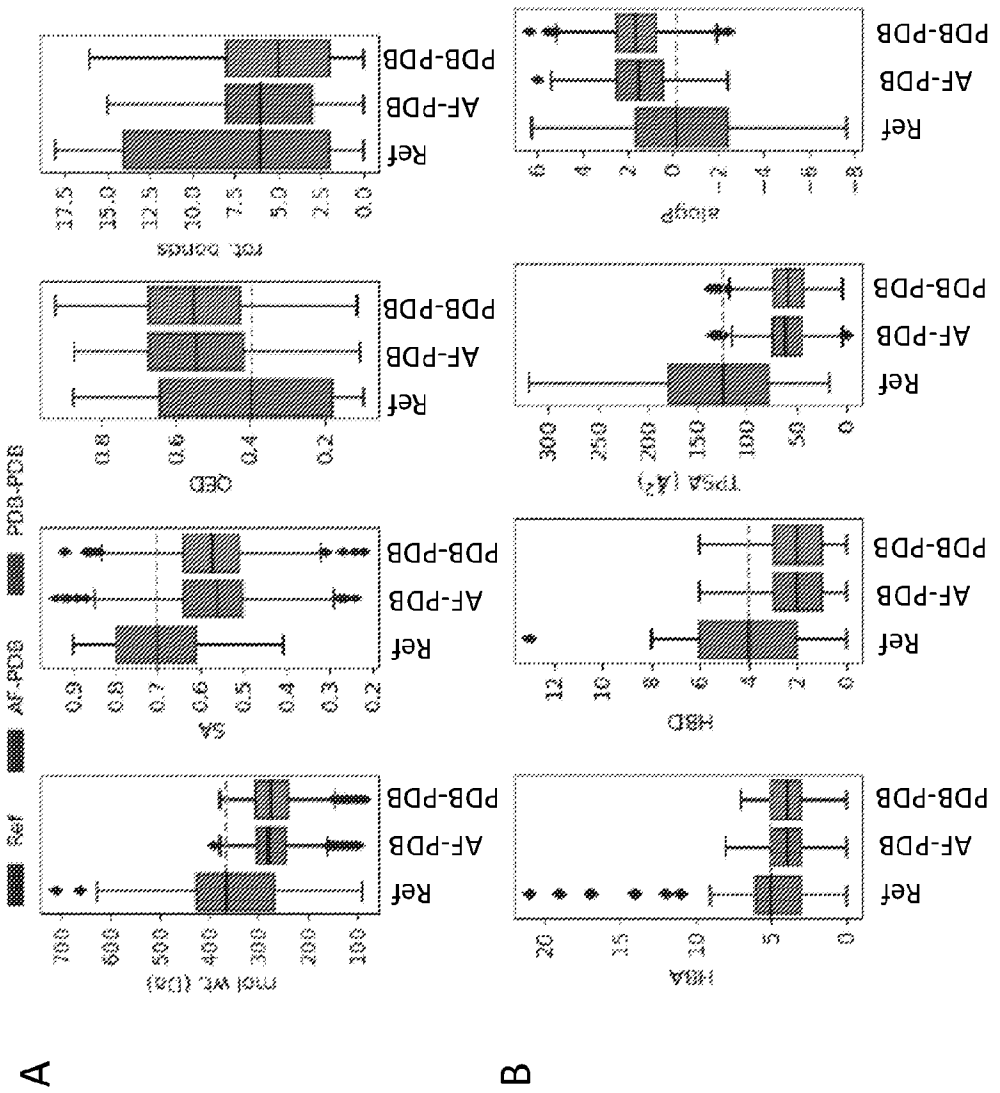


Figure 14A



Figure 14B



Figure 14C

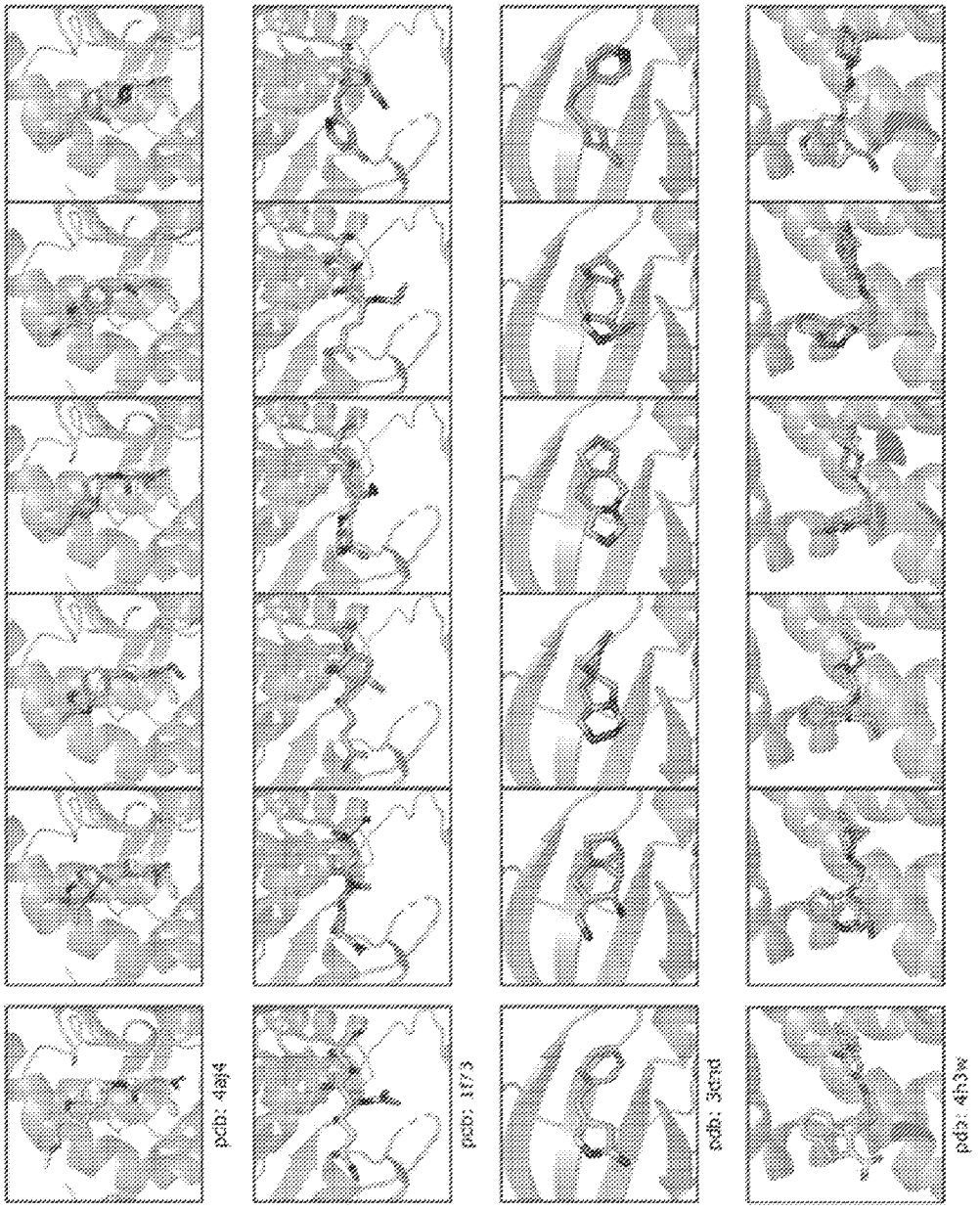


Figure 14D

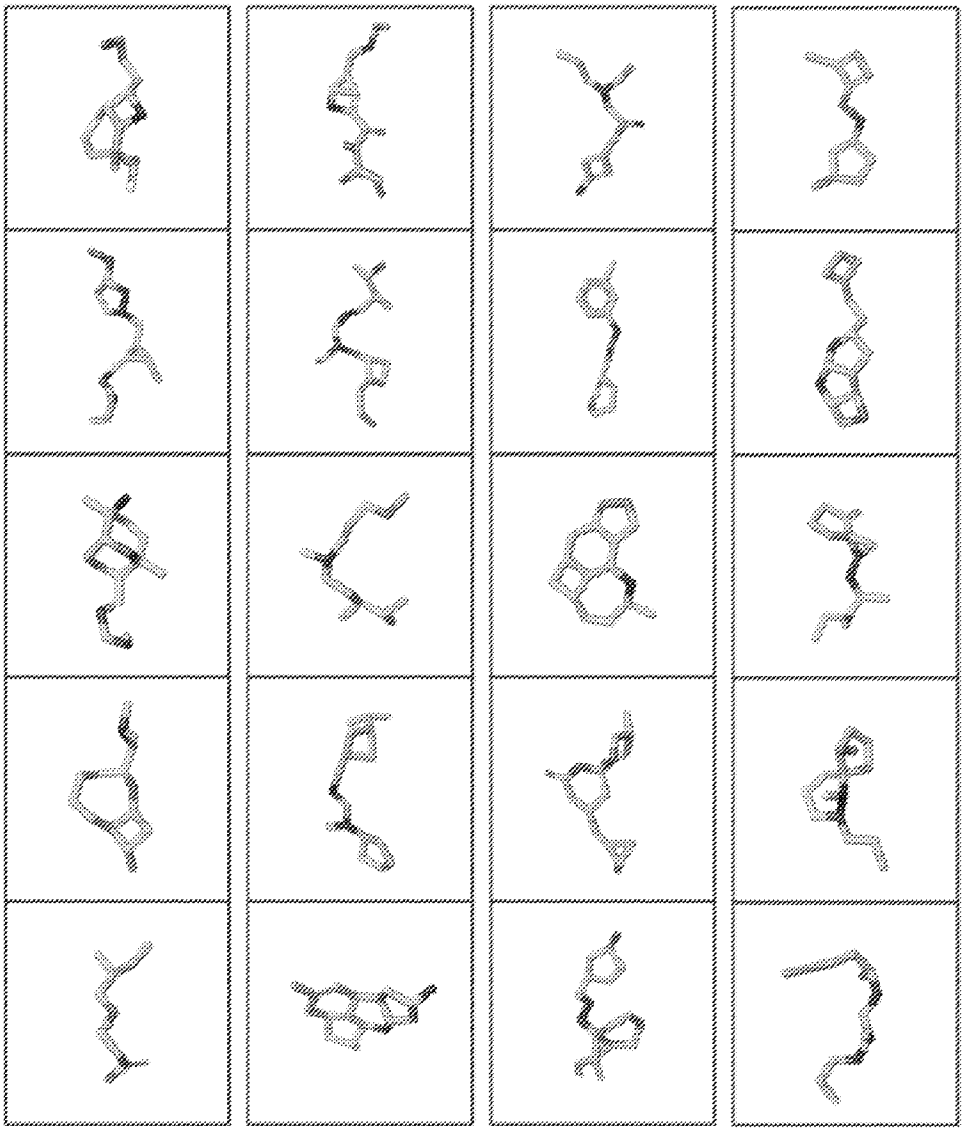
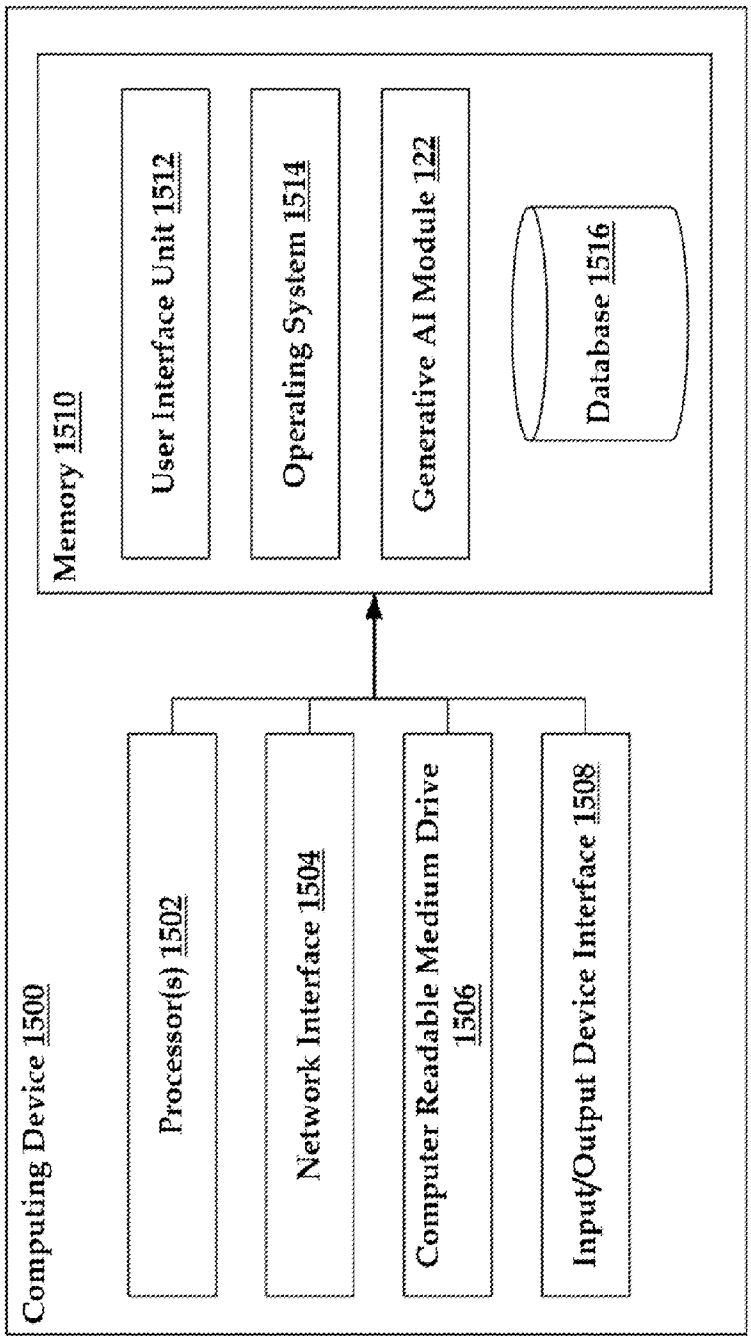


Figure 15


Computing Device 1500
Processor(s) 1502
Network Interface 1504
Computer Readable Medium Drive 1506
Input/Output Device Interface 1508
Memory 1510
User Interface Unit 1512
Operating System 1514
Generative AI Module 122
Database 1516

INTERNATIONAL SEARCH REPORT

International application No.
PCT/US2024/054511

A. CLASSIFICATION OF SUBJECT MATTER

IPC: ***G06N 3/08*** (2024.01); ***G16B 40/00*** (2024.01); *G16B 15/30* (2024.01); *G16B 15/00* (2024.01); *G06N 3/0455* (2024.01); *G16C 20/50* (2024.01)

CPC: ***G06N 3/08; G16B 40/00;*** *G16B 15/30; G16B 15/00; G06N 3/0455; G16C 20/50*

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.
A	US 2022/0358373 A1 (Ro5 Inc.) 10 November 2022 (10.11.2022) entire document (especially para [0108], [0248], [0260]-[0262])	1-9, 11-13, and 20
A	US 2023/0281443 A1 (InSilico Medicine IP Ltd)07 September 2023 (07.09.2023) entire document (especially [0073], [0097], [0118])	1-9, 11-13, and 20
A	US 2022/0406403 A1 (Innoplexus AG) 22 December 2022 (22.12.2022) entire document (especially [0035], [0038], [0064], [0069])	1-9, 11-13, and 20
A	US 2021/0057050 A1 (INSILICO MEDICINE IP LIMITED) 25 February 2021 (25.02.2021) entire document	1-9, 11-13, and 20
A	US 2023/0290435 A1 (Wipro Ltd) 14 September 2023 (14.09.2023) entire document	1-9, 11-13, and 20

☐ 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	Date of mailing of the international search report
30 December 2024 (30.12.2024)	**16 January 2025 (16.01.2025)**

Name and mailing address of the ISA/US	Authorized officer
COMMISSIONER FOR PATENTS **MAIL STOP PCT, ATTN: ISA/US** **P.O. Box 1450** **Alexandria, VA 22313-1450** **UNITED STATES OF AMERICA**	**KARI RODRIQUEZ**
Facsimile No. **571-273-8300**	Telephone No. **PCT Help Desk: 571-272-4300**

INTERNATIONAL SEARCH REPORT

International application No.
PCT/US2024/054511

Box No. II Observations where certain claims were found unsearchable (Continuation of item 2 of first sheet)

This international search report has not been established in respect of certain claims under Article 17(2)(a) for the following reasons:

1. ☐ Claims Nos.:
because they relate to subject matter not required to be searched by this Authority, namely:

2. ☐ Claims Nos.:
because they relate to parts of the international application that do not comply with the prescribed requirements to such an extent that no meaningful international search can be carried out, specifically:

3. ☑ Claims Nos.: **10, 14-19**
because they are dependent claims and are not drafted in accordance with the second and third sentences of Rule 6.4(a).