Chiral Graphs: Reduced Representations of Ligand Scaffolds for Stereoselective Biomolecular Recognition, Drug Design, and Enhanced Exploration of Chemical Structure Space

Simoun Mikhael and Ravinder Abrol*

Department of Chemistry and Biochemistry, College of Science and Mathematics, California State University, Northridge, CA 91330, USA

*Corresponding author: abrol@csun.edu

Abstract: Rational structure-based drug design relies on a detailed atomic level understanding of the protein-ligand interactions. The chiral nature of drug binding sites in proteins has led to the discovery of predominantly chiral drugs. Mechanistic understanding of stereoselectivity (which governs how one stereoisomer of a drug might bind stronger than the others to a protein) depends on the topology of stereocenters in the chiral molecule. *Chiral graphs* and *reduced chiral graphs* are new topological representations of chiral ligands that are introduced here, utilizing graph theory, to facilitate a detailed understanding of chiral recognition of ligands/drugs by proteins. These representations are demonstrated by application to all ~14,000+ chiral ligands in the protein data bank (PDB), which will facilitate an understanding of protein-ligand stereoselectivity mechanisms. Ligand modifications during drug development can be easily incorporated into these chiral graphs. In addition, these chiral graphs present an efficient tool for a deep dive into the enormous chemical structure space to enable the sampling of unexplored structural scaffolds.

1. INTRODUCTION

Chirality of molecules, recognized very early on by Louis Pasteur[1], provides a unique geometrical and topological handle to deconstruct the symmetries and asymmetries of molecular structure. Its relevance extends beyond the chemistry of stereoselective synthesis and chiral separation methods into biology, where natural life (as we know it) depends on molecular workhorses called proteins that are the biggest chiral molecules known in nature. These bio-molecules are exclusively made from the L-amino acids, though D-amino acids do play specific biological roles in all domains of life[2].

An important aspect of protein function is binding to and recognizing small endogenous ligands (or substrates) to carry out specific signaling tasks. The chirality of proteins makes this molecular recognition process inherently stereoselective, because most protein binding sites are chiral. Many of the small endogenous ligands also exhibit chirality and the protein's chiral binding site typically recognizes only one stereoisomer of this chiral substrate over its other possible stereoisomer(s)[3].

Drug design and especially structure-based drug design is based on mimicking the mechanisms of biomolecular interactions found in nature. The stereoselective biomolecular interactions mentioned above are influenced by the ligand configuration (or topology), which is the subject of this work, and the physico-chemical properties including the thermodynamics of the interacting molecules[4]. It is not surprising then that stereoselectivity of these protein-drug interactions plays an important role in a drug's pharmacological action[5].

There is a lot of interest in enantioselectivity and diastereoselectivity of drug candidates as it affects most aspects of pharmacodynamics and pharmacokinetics[6]. Potential toxicity of the pharmacologically-inactive stereoisomer in a racemate drug, highlighted by the thalidomide tragedy[7], led the United States Food and Drug Administration (FDA) and similar agencies from the European Union, Canada, and Japan to issue specific guidelines for the development and use of chiral molecules as drugs.[8] Due to major advances in stereoselective synthesis[9] and the necessity to clinically test all stereoisomers of a potentially racemic or diastereomeric drug mixture, it is usually cheaper to develop pure stereoisomers as drugs. So, the number of chiral drugs especially with multiple stereocenters, whether approved or in the clinical pipeline, has been steadily increasing[10]. This has been matched with a steady increase in the number of studies around the stereochemical considerations in the pharmacodynamics and pharmacokinetics of chiral drugs[6], however, the full potential of the use of stereoselectivity in rational structure-based drug design has not yet been realized.

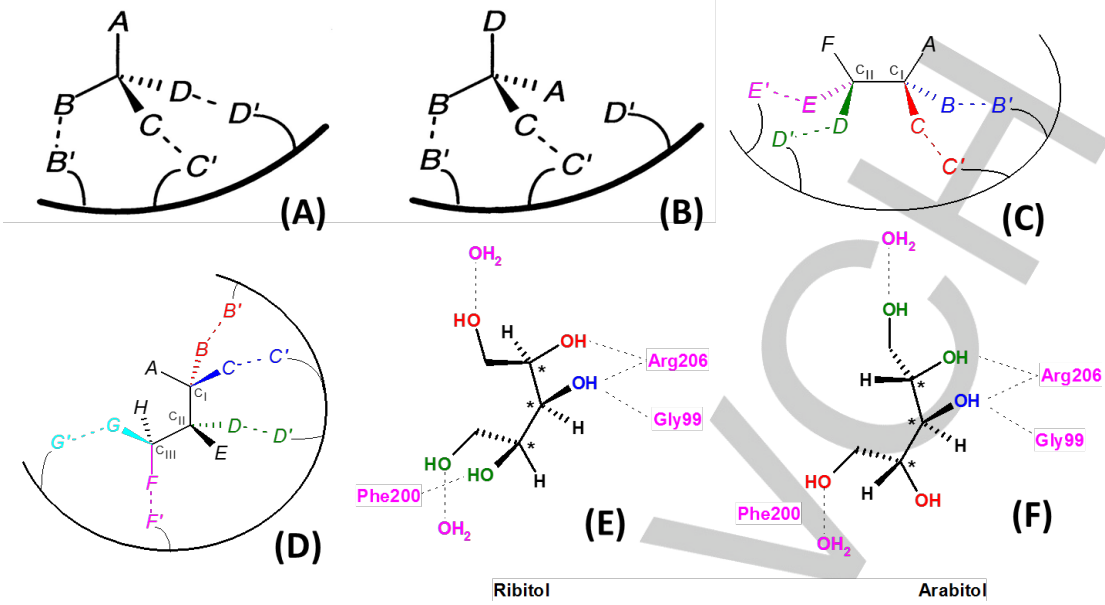


Figure 1: Panels (A) and (B) show the original 3-point attachment model of protein-ligand stereoselectivity, where for a ligand with one stereocenter, a stereoisomer shown in **(A)** needs three attachment points to be selectively bound to the protein (shown by a curve). The enantiomer shown in **(B)** can only interact with two attachment points. Panels **(C)** and **(D)** show the protein-ligand stereoselectivity requirements for a ligand with two and three stereocenters respectively, for one of the stereoisomers to be selectively bound to the protein. Panels **(E)** and **(F)** show stereoselectivity at work for a three-stereocenter ligand case (see text for details).

General stereoselectivity mechanisms underlying protein-drug interactions are not well understood, especially for ligands with multiple stereocenters that may be distributed in a molecule in any possible topology, given the gargantuan chemical structure space for small molecules. The structural basis of protein-ligand stereoselectivity[11] goes all the way back to 1933 through studies by Easson and Stedman[12] and later independently by Ogston[13] and Dalgliesh[14], where it was proposed that three attachment points between a protein and a chiral ligand are required for stereoselective recognition of only one stereoisomer of the ligand by the protein (**Figures 1A** and **1B**). This simple yet powerful model has been revised and updated to account for enantioselective situations with less than three attachment points[15, 16] or four interacting locations [17]. We have generalized these models to account for chiral ligands with linearly linked N stereocenters through the stereocenter-recognition (SR) model [18, 19] (**Figures 1C** and **1D**), where, for example, a ligand with two stereocenters needs at least four interaction points (**Figure 1C**) and a ligand with three stereocenters needs at least five interaction points (**Figure 1D**) with the protein, for stereoselective recognition by the protein of one stereoisomer over all other possible stereoisomers. This SR model has been used by others to explain the mechanisms behind observed protein-ligand stereoselectivities for acyclic ligands with multiple stereocenters. For example, ribitol (a ligand with three stereocenters) is transported more efficiently than its diastereoisomer arabitol by the membrane channel protein glycerol facilitator (GlpF) belonging to the aquaporin family.[20] It was shown that ribitol uses 5 interaction points with the GlpF protein compared to arabitol that can make at most 4 interaction points with GlpF, which enables ribitol to be transported faster and more efficiently than arabitol (**Figures 1E** and **1F**).[21] So, the SR model provides a simple and intuitive framework (**Figures 1C** and **1D**) to study stereoselectivity mechanisms for acyclic ligands with one or more asymmetric atoms.

The major challenge in stereoselectivity is the lack of a topological framework to handle the chemical complexity of general chiral ligands/drugs, where the asymmetric atoms may be topologically distributed in ring structures, branched structures, or mixed structures. The Protein Data Bank (PDB)[22] contains structures for thousands of protein-ligand complexes covering ~14000+ chiral ligands, where some ligands have 10 or more asymmetric atoms as will be presented below. A large number of these

structures include the same protein bound to different stereoisomers of the same ligand, which makes PDB a very valuable resource for studying protein-ligand stereoselectivity mechanisms and then incorporating this knowledge in the rational structure-based design of stereospecific drugs.

Before the PDB can be surveyed for stereoselective protein-ligand interactions, we need a general but reduced representation of chiral ligands that will enable the probe of their stereoselectivity mechanisms, especially for those ligands whose asymmetric atom(s) are topologically distributed in cyclic or branched structures to capture a large portion of the chemical diversity in structure space. For this study, we will focus on ligands exhibiting chirality due to the presence of one or more asymmetric atoms. Axial/planar chirality and cis/trans isomerism are not considered at this time. In organic and biological molecules, carbon atoms are the most common chiral centers. Occasionally, other heavy atoms such as phosphorus or sulphur or positively charged quaternary nitrogen with different groups, or a heteroatom that connects three chiral centers can have an influence on the ligand's conformation[23], which needs to be taken into account in any reduced representation of chiral ligands.

The in silico representations of ligands have proven to be a boon for virtual ligand screening (VLS)[24], which can be both time and cost effective in drug discovery campaigns. Using these virtual representations of ligands, the computer simulations can use either the protein binding sites to find strongly binding ligands or the ligand-structure based pharmacophores to find similar ligands with potentially better pharmacodynamics and ADMET properties.

Chemical graphs have a long history in cheminformatics, drug discovery, and quantitative structure-activity relationship (QSAR) studies.[25] In addition, reduced graph representations of ligands have found use in QSAR, virtual screening, and cheminformatics applications based on similarity searches.[26] Graph representations have also been used to classify all the ligands in the PDB. [27]

We needed a reduced representation of ligands that could provide a topological framework to study the mechanisms of protein-ligand stereoselectivity. So, with this primary goal in mind and the need to recognize interactions around asymmetric centers in ligands, we present here two reduced representations of chiral ligands based on graph theory. These representations are called *chiral graphs* as the asymmetric centers act as nodes in these graphs. These *chiral graph* representations can handle the presence of multiple asymmetric atoms in any structural topology and can also enable a seamless incorporation of new chiral centers in a molecule that could happen during drug development and optimization. These reduced representations also have the potential to enable explorations of novel scaffold space in the vast chemical space of potential drugs due to their coarse nature, as will be shown later.

To demonstrate these *chiral graph* representations, we have applied them to all ~14,000+ chiral ligands out of a total of ~26000+ ligands in the PDB database (as of June 2018),[22] The PDB has been updated since 2007 to include chirality information (determined by CACTVS[28] and OEChemTk[29]) and other rich chemical content for ligands found in the protein/DNA structures.[30] The PDBeChem server provides a user-friendly searchable interface to this rich chemical content of PDB ligands,[31] which is utilized in the construction of these graphs.

Next, the methods behind these reduced representations (*chiral graphs* and *reduced chiral graphs*) are described, followed by their first application to all ~14,000+ chiral ligands mentioned above. These *graphs* are catalogued in a searchable publically accessible webserver http://chiralig.abrollab.org for all the chiral ligands. The methods are followed by a discussion of how these representations can be utilized for probing protein-ligand stereoselectivity mechanisms, drug structure optimization, and exploration of chemical structure space for novel scaffolds.

2. METHODS

The size of the chemical structure space is enormous due to the complex topologies of ligand structures. In addition, there are no major restrictions on how multiple stereocenters can be distributed in the complex topologies of chiral ligands. Keeping these considerations in mind along with the primary goal to provide a topological framework for studying protein-ligand stereoselectivity, we adopt two graph-based representations for ligands: *chiral graphs* and *reduced chiral graphs*. The asymmetric atoms of

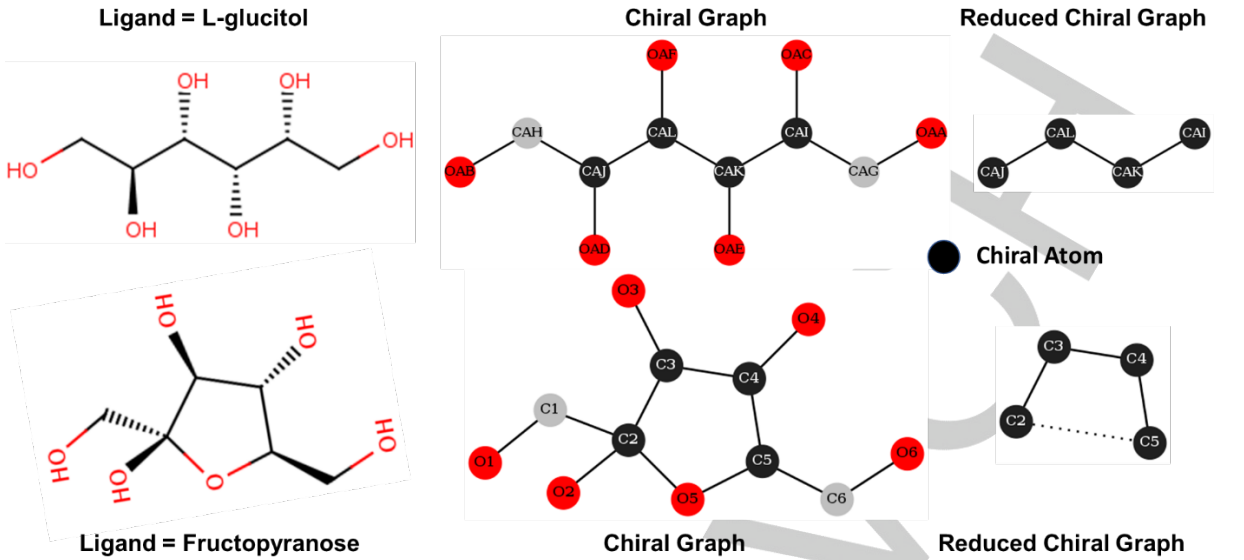


Figure 2: Chiral representations for the ligands L-glucitol (Ligand Code: 62W, shown in top leftmost panel) and Fructopyranose (Ligand Code: FRU, shown in bottom leftmost panel). *Chiral Graphs* (center panels) show the chiral atoms in black, achiral carbon atoms in grey, and oxygen atoms in red. *Reduced Chiral Graphs* (rightmost) show the reduced representation of *chiral graphs*. Atom labels correspond to atom names found in PDB files.

ligands act as nodes in both representations. We utilize the multigraph type of graphs, which can handle all complex chemical topologies as they allow for multiple edges and self-loops in these graph-based representations. Multiple edges between two nodes are possible in chemical structures if only two chiral centers are present in a ring structure. Self-loops are also possible if a chiral center is present in a ring structure.

Chiral Graphs

A *chiral graph* (CG) for a ligand represents how all the heavy atoms (non-hydrogen atoms) in the ligand are connected to each other. The hydrogen atoms are ignored in this representation. The PDBeChem server[31] is used to get the coordinate information for all atoms, the connectivity information about all bonds, and the identity of all the chiral atoms in the ligand. Each heavy atom is displayed as a node and each bond (single or double or triple) between the directly connected heavy atoms is displayed as an edge. A *chiral graph* is essentially similar to the graph representation of a molecule's line drawing. The main difference is that the chiral atoms are shown as black nodes in the *chiral graphs*. **Figure 2** shows the *chiral graphs* for an acyclic ligand (L-glucitol) and a cyclic ligand (fructopyranose) without any branched chiral topologies. **Figure 3** shows the *chiral graphs* for an acyclic ligand and a cyclic ligand with branched chiral topologies. The unbranched and branched chiral topologies combined with acyclic and cyclic structures covers all potential local structural complexities that will be encountered in potential ligands. A *chiral graph*, which highlights the chiral atoms as black nodes, provides a tool to probe stereoselectivity mechanisms when overlaid with protein-ligand interactions as will be shown in the results and discussion section on the applications of *chiral graphs*.

Reduced Chiral Graphs

Different reduction approaches[32] have been used in the past to reduce the chemical structures into simpler graph representations, where each node represents a group of atoms rather than one atom. However, we take a different approach for a reduced representation of the ligand structure, where the *reduced chiral graph* (RCG) takes the *chiral graph* introduced above and reduces it to only focus on the chiral atoms and how they are connected to each other. So, the *reduced chiral graph* captures the core structural framework of a ligand, which can enable novel scaffold searches for ligands with same/similar structural core, as will be shown in the last application in the results and discussion section. In this *reduced chiral graph*, all the chiral atoms are displayed as nodes and most other atoms are ignored. The nodes are

connected to each other through direct edges (displayed as solid lines) for chiral atoms directly bonded to each other or through long edges (displayed as dashed lines) for chiral atoms indirectly connected to each other through achiral atoms. Only other atoms that may be displayed are achiral atoms that bridge three or more chiral centers (nodes).

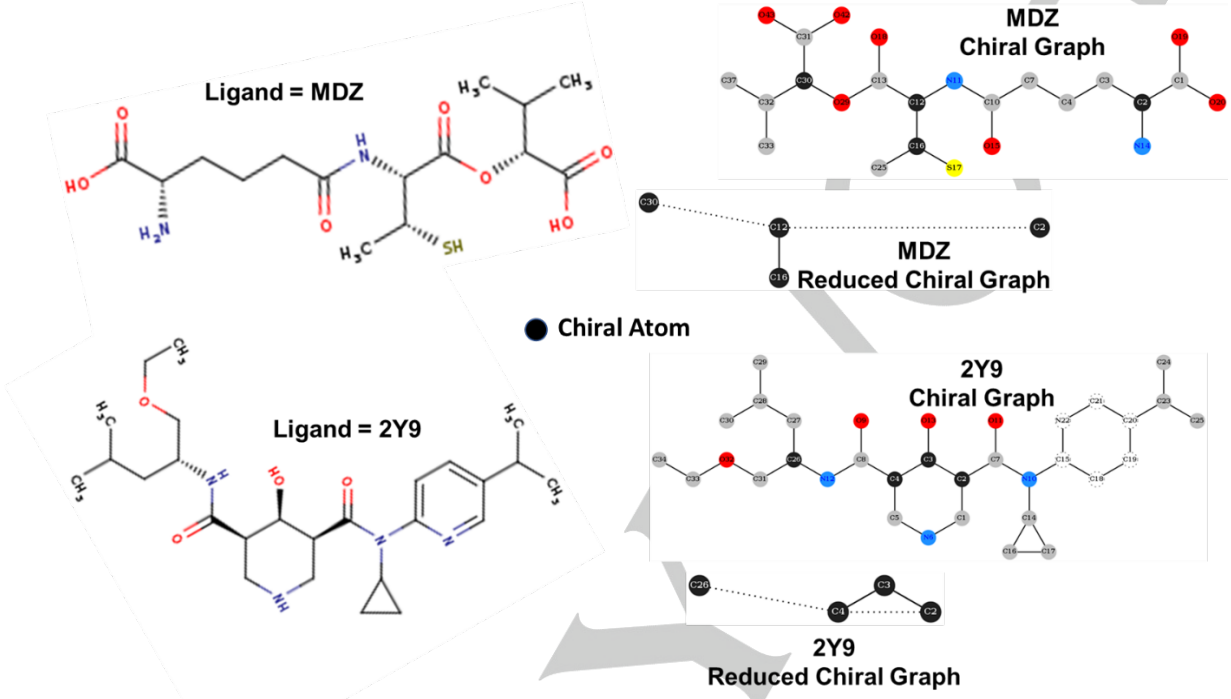


Figure 3: Chiral representations for the branched ligands: Top panel (Ligand Code MDZ) = N~6~-methyl-6-oxo-L-lysine - 2-[(3-mercaptobutanoyl) oxy]-3-methylbutanoic Acid) and bottom panel (Ligand Code 2Y9) = (3S,4R,5R)-N-cyclopropyl-N'-[(2R)-1-ethoxy-4-methylpentan-2-yl]-4-hydroxy-N-[5-(propan-2-yl) pyridin-2-yl] piperidine-3,5-dicarboxamide. Atom labels correspond to atom names found in PDB files.

The reduction algorithm works as follows: The PDBeChem server provides the atom connectivity, coordinates, and chiral/achiral identity for each atom in the ligand. The algorithm goes through the list of all chiral atoms one by one in the default order obtained from PDBeChem. From each chiral atom, we explore all 4 paths based on the atom connectivity information. If along a path, another chiral atom is found, exploration along that path stops immediately and an edge is created between the two chiral centers (nodes). The exploration continues along each remaining path, until the atoms are exhausted or another chiral atom is encountered. If an achiral atom is found to be linked to three or more chiral centers, it is treated as a pseudonode. Starting from a chiral center, if another chiral center is encountered twice (along two different paths), it implies that there are two chiral centers in a ring. Once all edges and any pseudonodes are obtained, that information is combined with the coordinates to generate a *reduced chiral graph*.

The right panels in **Figure 2** show two simple cases for *reduced chiral graph*s where all chiral atoms are distributed either linearly (acyclic unbranched case) or in a ring (cyclic unbranched case). **Figure 3** shows *reduced chiral graph*s for the corresponding acyclic and cyclic branched cases. These *reduced chiral graphs* capture the major topological features of a ligand, which can be utilized in several cheminformatics applications like: a) scaffold and similarity searches[33] for ligands with different structures but with similar topology of the chiral centers; b) biochemical applications like studying biomolecular interactions[34] between proteins and the stereoisomers of ligands; and c) examining the outcome of hit or lead drug candidates' structural modifications[35] in terms of the effect on protein-ligand interactions. Some of these applications will be highlighted below in the results and discussion section.

Application to all chiral ligands in the PDB

Table 1. Classification of the 26,349 ligands in the PDB database by chirality and number of ligand stereoisomers in the PDB. In each of the 916 ligand stereoisomer groups (corresponding to a total of 1,955 ligands), the ligand and its stereoisomer(s) are bound to a protein in the PDB.

Ligands		Ligand + 1 stereoisomer	+ 2	+ 3	+ 4	+ 5	> 5	Total Ligands
Chiral	916 Ligand Groups	792	85	19	8	5	7	
	Stereoisomers	1507	243	72	40	30	63	1,955*
	Chiral ligands with no other stereoisomers in the PDB							12,181
Achiral ligands in the PDB								12,115
Cis/Trans Ligands								98
Total number of ligands in the PDB								26,349

*The 1,955 stereoisomeric ligands are found in 60,829 PDB protein chains.

The *chiral graph* and the corresponding *reduced chiral graph* representations can be applied to a large number of ligands, for example, those ligands found in: the PubChem [36] database (~96 million compounds), the ZINC [37] database (~35 million purchasable molecules), the DrugBank [38] database (~12,000 approved / experimental drugs), the ChEMBL [39] database (~2 million compounds), the PDB [22, 40] (~26,000 ligands), the Chemical Abstract Services (CAS) [41] database (~140 million unique organic and inorganic chemical substances), and the GDB-17 [42] database (~166 billion organic small molecules).

For the purpose of this study and to facilitate the study of protein-ligand stereoselectivity mechanisms, the *chiral graphs* and *reduced chiral graphs* were created for all ~14,000+ chiral ligands out of a total of ~26,000+ ligands found in the PDB. The detailed method is described in Supplementary Information **Section 1** and **Figures S1** through **S4**. We began by data mining the PDBeChem database[31] using the three-letter code for each ligand as an input to obtain information about the structure of those ligands. They were then categorized into three groups: a) chiral ligands that have stereoisomers in the PDB, b) other chiral ligands that don't have stereoisomers in the PDB, and c) achiral ligands. First, the *chiral graphs* were created based on the method described earlier, using the atom connections in the CONECT records of the PDB files found in the ligand dictionary[31] of all ligands in the PDBeChem server. These atom connections in the PDB files are consistent with connectivity records in the mmCIF data files for the cases we looked at manually. Then, the *reduced chiral graphs* were created by reducing those atom connections in the PDB files into only connections across all the chiral atoms, based on the algorithm described earlier. These graph representations for all chiral ligands are available and searchable at the ChiraLig database: http://chiralig.abrollab.org.

In the next section, we will present the results of the application of these chiral representations to all the chiral ligands in the PDB. This will be followed by brief discussions on how these representations can be utilized to: probe protein-ligand stereoselectivity, study structural modifications during drug development, and discover novel scaffolds.

3. RESULTS AND DISCUSSION

Classification of Ligands and Generation of Chiral Graphs

Based on the information obtained from the PDB database[22] and the ligand dictionary in PDBeChem[31], we generated *chiral graphs* and *reduced chiral graphs* for 14,136 chiral ligands in the PDB, which included ligands with other stereoisomers in the PDB and ligands without any other stereoisomers in the PDB. As shown in **Table 1**, the ligands were classified into three major categories: chiral ligands represented by two or more stereoisomers in the PDB (1,955 ligands), chiral ligands with only one stereoisomer in the PDB (12,181 ligands), and achiral ligands in the PDB (12,115 ligands). The first ligand category was further divided into subgroups (see **Table 1**) based on the number of stereoisomers that each ligand had in the PDB. For instance, some ligands had only one other stereoisomer per ligand, so they made ligand

pairs and there were 792 such ligand pairs or groups. Some other ligands had 2 additional stereoisomers per ligand so they made ligand groups of 3 stereoisomers and there were 85 such ligand groups. Each group of stereoisomers found in the PDB per ligand is called a *Ligand Stereoisomer Group*, which is independent of the number of stereocenters in the ligands.

Table 2. Enumeration of the chiral ligands with more than one stereoisomer in the PDB in terms of number of stereocenters and presence of ring structures.

Number of Stereocenters	Number of Ligands	Ligands with Rings
1	645	425
2	303	196
3	217	164
4	293	238
5	217	187
6+	280	251
Total	**1,955**	**1,461**

The mining search over the entire PDB resulted in 916 *Ligand Stereoisomeric Groups* with the total of 1,955 individual ligands. Finally, 88 ligand codes were removed from the list of 26,437 because they were obsoleted, reducing the number to 26,349 total ligands (See Supplementary **Section 1** and **Figures S1** through **S4** for further details).

The 1,955 chiral ligands in the PDB with one or more other stereoisomers in the PDB, were also analyzed as a function of the number of stereocenters. **Table 2** presents this data in terms of the distribution of the 1,955 chiral ligands as a function of the number of stereocenters and the subset of those ligands that contain rings. The data presented in **Table 1** with 916 ligand isomeric groups suggests that the structural data in the PDB is quite rich to probe the stereoselectivity mechanisms as there are thousands of structural comparisons that will be possible with those ligand groups. **Table 1** data combined with **Table 2** also suggests the rich structural complexity of PDB ligands as this set contains large number of stereocenters and structures with rings dominate each group of ligands with the same number of stereocenters.

The reduced representation algorithms were designed to handle the huge variety of complex ligand topologies. Our purpose was to create a software tool that generates *chiral graphs* and *reduced chiral graphs* for all possible ligand structures. **Figure 2** shows the graphs for linear and cyclic structures with unbranched topologies. **Figure 3** shows the graphs for linear and cyclic structures with branched topologies.

Other examples with additional structural variations are shown in Supplementary **Figure S5**, **Figure 4**, and **Figure 5**.

In **Supplementary Figure S5**, four symmetrical groups are attached to a central *prochiral* atom, which is represented as a *pseudonode* in that figure.

Figure 4 shows the ligand QDN with adjacent rings, and the representation of N1 nitrogen atom as a *pseudonode,* which connects the three chiral centers; C11, C13, and C17.

Figure 5A shows an example of a ligand with a *loop* representation, which happens when only one stereocenter is present in a ring topology. **Figure 5B** shows the simplest *reduced chiral graph (single node)*, for a ligand containing only one chiral center, which is relevant in the limit of the three-point models of protein-ligand stereoselectivity proposed by Easson and Stedman[12] and Ogston[13]. For ligands with only one chiral center, the three-point models of

stereoselectivity work reasonably well and the chiral graphs or reduced chiral graphs do not provide major additional insight. As discussed in the next section, these three-point models of stereoselective protein-ligand interactions are inadequate for ligands with multiple stereocenters.

Examples shown in **Figure 2** through **Figure 5** demonstrate that chiral graphs and reduced chiral graphs can handle the structural complexities of the chemical structure space for all chiral ligands based on asymmetric atoms

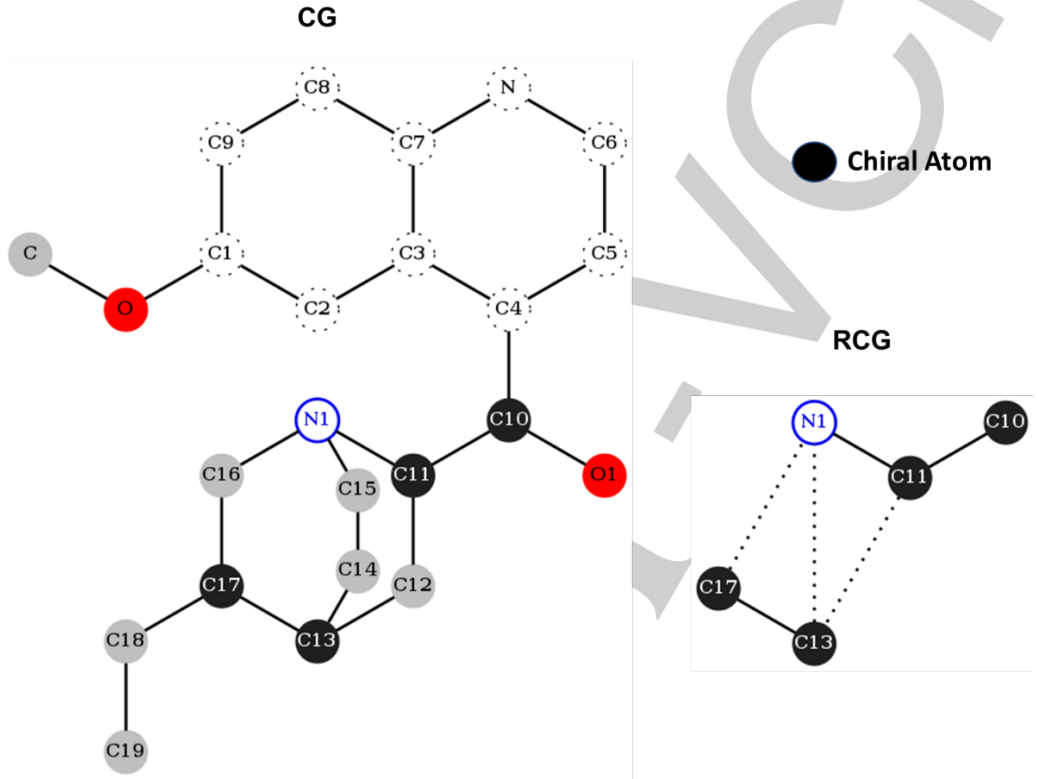


Figure 4: *Graph representations* for the ligand Quinidine with adjacent rings and pseudonode (N1). *Ligand Code*: QDN. *Left panel: Chiral Graph (CG)*, aromatic rings (atoms with dashed circles), pseudonode N1 (empty solid circle), and chiral centers (filled solid circle). *Right panel: Reduced Chiral Graph (RCG).* Atom labels are from PDB files.

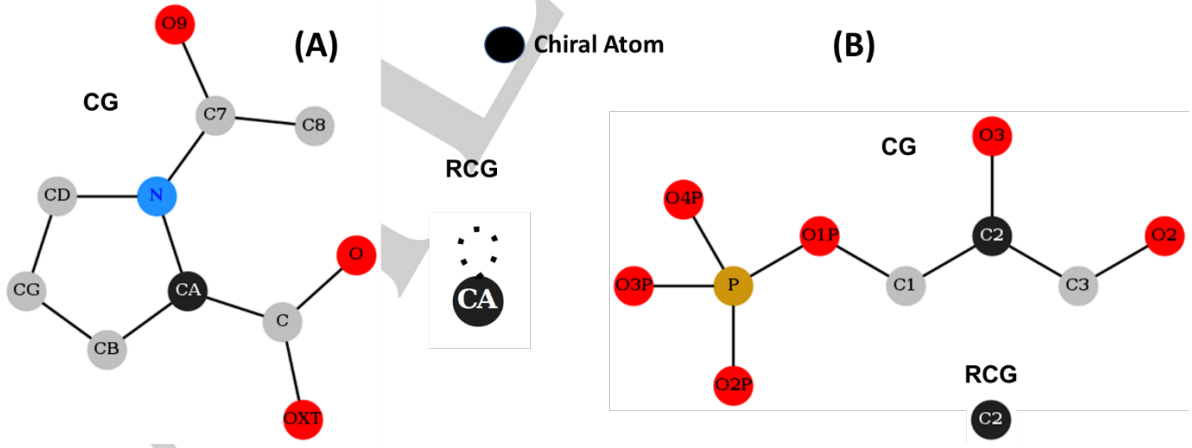


Figure 5. (A) *Graph representations* for the ligand N-ACETYL-D-PROLINE. *Ligand Code*: N8P. *Left:* Chiral Graph (*CG*). *Right: Reduced Chiral Graph (RCG)* showing the loop graph corresponding to the chiral atom in the ring. **(B)** *Graph representations* for the ligand SN-GLYCEROL-1-PHOSPHATE. *Ligand Code*: 1GP. *Top:* Chiral Graph (*CG*). *Bottom: Reduced Chiral Graph (RCG)* showing the simplest RCG possible. Atom labels are from PDB files.

Application in the Study of Protein-ligand Interactions and Stereoselectivity

The *chiral graphs* (CGs) and *reduced chiral graphs* (RCGs) can serve as effective tools to look at the stereoselectivity of protein-ligand interactions. Stereoselectivity usually manifests itself in terms of binding affinity differences in how two or stereoisomers bind to the same protein. The *chiral graphs* presented here provide minimal representations of the chiral ligands and when combined with known interactions with protein residues in a binding site, they can shed light on specific ligand-residue interaction(s) that might explain observed binding affinity differences behind stereoselectivity. If stereoselectivity manifests itself in terms of enzymatic activity differences between two or more substrate stereoisomers, the corresponding *chiral graphs* combined with how the substrate positions itself in the active binding site will show different modes of interactions for the substrate stereoisomers

The ligand topology approach has been used to construct chiral recognition models such as the three-point attachment or interaction models[12],[13],[14] (**Figures 1A** and **1B**) or the four location model[17]. We have generalized these through the introduction of stereocenter recognition (SR) model (**Figures 1C** and **1D**) to account for chiral ligands with linearly linked *N* stereocenters[18, 19] and through the general SR model (GSR model) for chiral ligands with *N* stereocenters distributed in any topology (unpublished).

In the SR model[18, 19], a molecule with one chiral center can interact with a protein or receptor through either binding interactions or repulsive (steric) interactions to achieve stereoselectivity. The interacting locations on the ligand have been previously defined as functional groups or any extended groups attached to the chiral center [19]; therefore, a molecule with one chiral center has four potential interaction locations with the protein and a molecule with two chiral centers has at least six interaction locations with the protein (3 on each chiral center and additional interactions through any bridging functional groups between the two chiral centers). A single interaction location on the ligand can interact with one or more residues on the protein through binding or steric interactions[15]. The SR/GSR models provide a protein-ligand interaction framework for simple chiral structures as well as more complex ones such as ligands with multiple stereocenters that contain rings and branches, and currently there is no other model that provides such a framework[18]. So, the use of *chiral graphs* and *reduced chiral graphs* in combination with the SR/GSR models provide a generalized structural framework to analyze the stereoselectivity in protein-ligand interactions for any chiral ligand because these graphs retain full topological information of all stereocenters and their protein interactions. They can be also combined with tools like IChem[43] and PLIP[44] that are used for analyzing protein-ligand interactions.

To highlight the use of *chiral graphs,* we used these graphs to visualize protein-ligand interactions provided from the protein-ligand interaction profiler (PLIP) server[44]. **Figure 6** shows the interaction maps of two chiral ligands (**6A:** quinidine-QDN and **6B:** quinine-QI8) interacting with the human cytochrome P450 protein[45], in terms of *chiral graphs* (left panels) and *reduced chiral graphs* (right panels). The *chiral graphs* capture detailed interactions between the protein and the ligands, whereas the *reduced chiral graphs* allow for an assignment of those interactions to the nearest stereocenter to enable mechanistic insight into stereoselectivity. Interacting residues can be easily identified using a distance cutoff from the ligand using automated tools or scripts. If the same residue is interacting with two stereocenters, it is constraining both centers, so that interacting residue should be assigned to both stereocenters. In **Figure 6**, both of the two stereoisomers (quinidine and quinine) interact with protein residues Phe120, Glu216, & Ser304, but quinidine selectively interacts with residues Glu244, Phe247, & Phe483, and quinine selectively interacts with residues Leu121, Val370, & Val374. These differences in interacting residues and modes of interaction highlighted by striking differences between the interaction maps shown by r*educed chiral graphs* in **Figure 6** provide valuable structural insights into potential mechanisms of stereoselectivity, as it was shown in **Figures 1E** and **1F** for ribitol vs arabitol transport by GlpF.[20, 21]

As mentioned earlier, we have found 1,955 ligands in the PDB that have one or more stereoisomers in the PDB bound to some protein. Based on these ligands, we have identified ~28,000 protein-ligand complex pairs in the PDB where a ligand and its stereoisomer are bound to the same protein. We are utilizing *chiral graphs* to represent their interactions to understand the stereochemical interactions for those complex pairs in the PDB in terms of hydrogen bonding

interactions, hydrophobic interactions, water-bridges, etc. This will be presented in a follow-up publication as it is beyond the scope of this study.

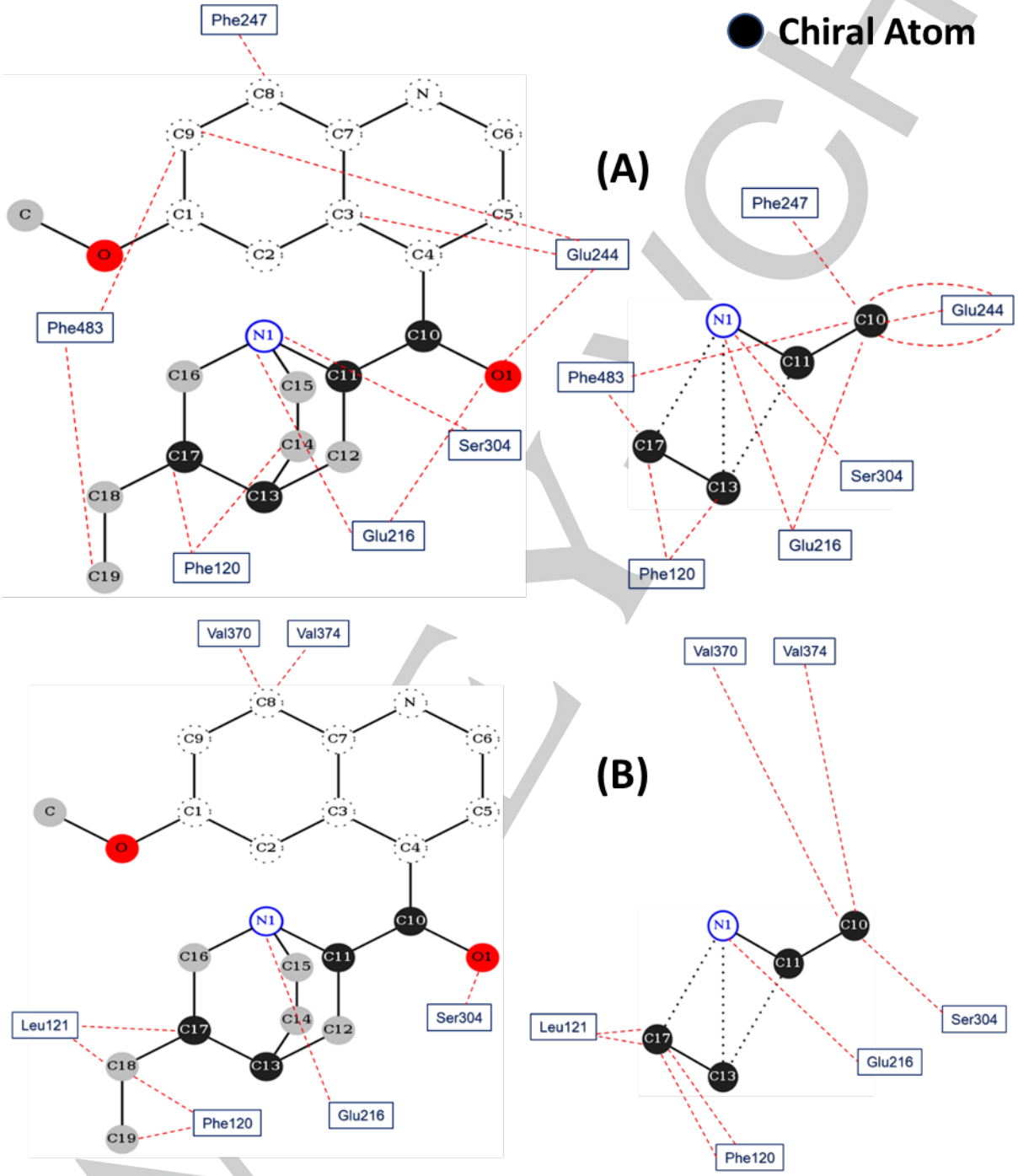
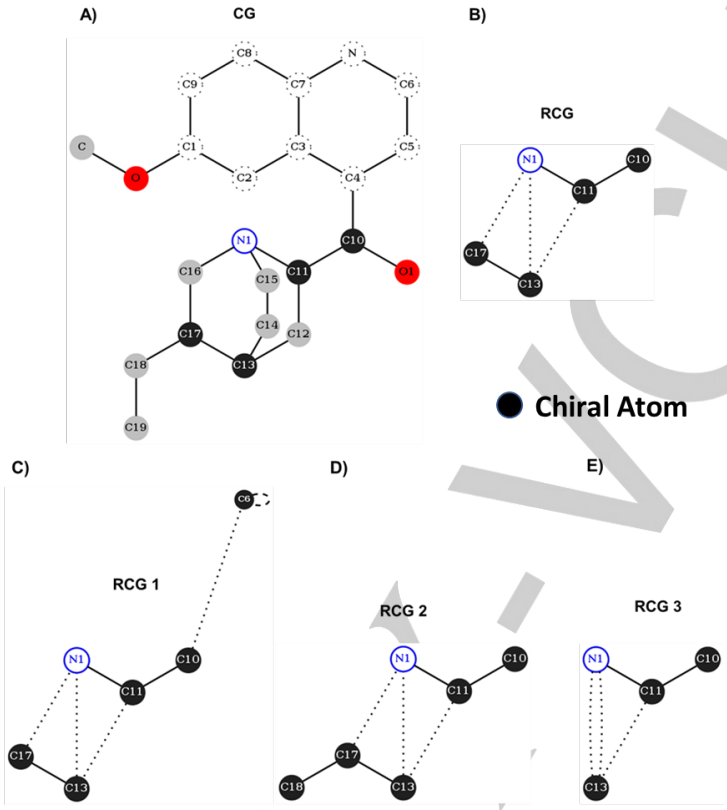


Figure 6: (A) Protein-ligand interaction information retrieved from PLIP server, and comparison between the ligand QDN and its stereoisomer QI8; *Left:* The complete QDN ligand *Chiral Graph (CG)*, previously shown in **Figure 3,** after adding the active site interacting residues based on **PDBid 4wnu_A**. *Right: Reduced chiral graph (RCG)* representation of the same protein-ligand interactions. **(B)** *Left:* The complete QI9 ligand *Chiral Graph (CG)* after adding the active site interacting residues based on **PDBid 4wnv_A**. *Right: RCG* representation of the same protein-ligand interactions. Atom labels correspond to atom names from the PDB files.

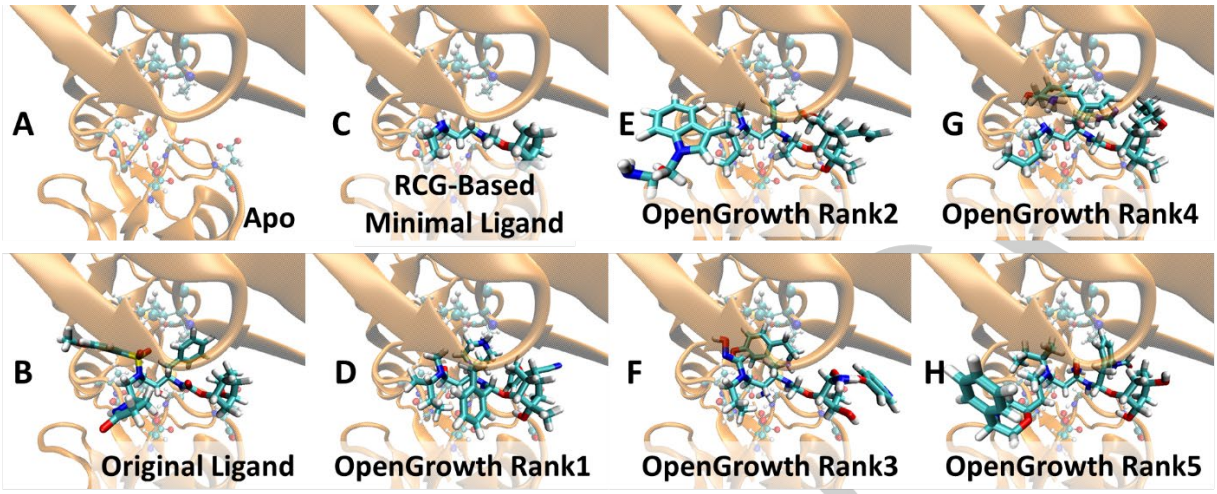


Figures 7: *RCG* topology after different structure modifications for the ligand QDN; **A)** *Chiral Graph (CG)*; **B)** *Reduced Chiral Graph (RCG)*; **C)** New RCG after adding C6 to the chiral atoms list; **D)** New RCG after adding C18 to the chiral atoms list; **E)** New RCG after removing C17 from the chiral atoms list. Atom labels are from PDB files.

Application in Drug Design and Development

The reduced representations presented in this study have some practical applications in structure-guided drug discovery. During drug development, initial hits are structurally modified to create many lead molecules with desirable pharmacodynamics and pharmacokinetic parameters. A large amount of structure-activity relationship data is generated around many series of promising molecules. During these structural modifications, sometimes a new chiral center is added to the molecules and at other times an existing chiral center might be lost. These structural optimizations can be facilitated by utilizing *reduced chiral graphs* in general and more so when protein structure and binding site residues are guiding those optimizations.

Figure 7 presents some *reduced chiral graphs* that will result from structural modifications of the previously presented ligand quinidine (QDN). **Figure 7A** shows the *chiral graph* for QDN and **Figure 7B** shows the corresponding *reduced chiral graph*. The chiral centers are shown as solid black nodes. The atom C6 is prochiral, but if a functional group was added to it to make it chiral, then the corresponding *reduced chiral graph* would look like **Figure 7C**. This modified ligand would have a slightly different interaction map (like one shown in **Figure 6**) with the protein if modification on atom C6 engages with the protein, which can be captured by the *reduced chiral graph*. Similarly, if a new chiral center is introduced at the other end of the molecule at carbon atom C18, then the corresponding *reduced chiral graph* would look like one shown in **Figure 7D**. If one of the chiral atoms C17 is either lost during structural modifications or becomes achiral, that structure is captured by the *reduced chiral graph* shown in **Figure 7E**. Such structural modifications are usually done during the hit to lead stage of drug development or during lead optimization to enhance the parent ligand's binding potency for the protein.



Figures 8. The use of *reduced chiral graph* in exploration of novel scaffolds determined by the binding site. (A) Apo Multidrug Resistant HIV-1 Protease Clinical isolate PR20 (PDB: 4j55); (B) Original ligand bound to the HIV-1 Protease; (C) *Reduced chiral graph* based minimal ligand; (D)-(H) Top 5 ranked ligand derivatives grown from the RCG based minimal ligand using the OpenGrowth program (see text for details).

As an illustration, if the protein binding site for parent ligand shown in **Figure 7A** is known, then a protein-ligand interaction map for this ligand, similar to that shown in **Figure 6**, can be used to assess the creation of the new chiral center at C6 (using stereoselectivity models like one shown in **Figure 1**) to get the ligand in **Figure 7C**. Using in silico docking, both possible stereoisomers due to C6 chirality should be docked to the binding site while optimizing the protein sidechains within interaction distance of the ligand. The new binding sites and interactions for the two possible stereoisomers should be overlaid on the reduced chiral graph of **Figure 7C**, to evaluate potential stereoselectivity between the two isomers, so that a decision can be made about which stereoisomer to synthesize. Such modifications can be done very efficiently using in silico methods. The *reduced chiral graph*s can facilitate and guide these structural modifications, as the graphs are flexible enough to allow the addition or removal of chiral centers at will. This can help prioritize potential new potent and stereoselective derivatives to synthesize. *Reduced chiral graphs* can also assist with the fragment based drug discovery campaigns, where chiral fragments can be combined together in the binding site to assess the viability of a multi-fragment assembled ligand before an expensive synthesis. Advances in stereoselective synthesis combined with the regulatory burden of testing individual stereoisomers of a drug, are aided by tools like *reduced chiral graph*s in rational structure-based addition of chiral centers during drug development, so that pure and efficacious stereoisomeric drugs can be produced.

Application in Structure Exploration and Scaffold Search

Reduced chiral graphs were introduced in this study as a structural framework to probe the protein-ligand stereoselectivity. However, these graphs also represent one of the coarsest reduced representations of a ligand's topology, where functional groups, achiral groups, achiral extensions, and achiral linkers can be treated as variable decorations around the chiral framework of the ligand represented in the *reduced chiral graph*. The enormity of the chemical structure space, even for small drug like molecules with molecular weight under 1000 Daltons, cannot be overstated. The *chiral graphs* and *reduced chiral graphs* represent a range of coarseness that can be exploited to explore the de novo chemical structure space and to perform similarity searches at different coarseness levels.

In order to demonstrate the potential of reduced chiral graphs in exploration of novel structural scaffolds[46], we selected one of the ligands from our database with the code "031". Based on PDBeChem server data for this ligand, we selected the PDB entry 4j55[47], which is an extreme multidrug resistant HIV-1 protease bound to this ligand. The protein is shown in **Figure 8A** without the ligand and in **Figure 8B**

with the original ligand. Based on the reduced chiral graph of this ligand, a chemically accurate minimal ligand was constructed shown in **Figure 8C**. This minimal ligand retains all the chiral atoms and any achiral atoms from the original ligand that are necessary to connect all the chiral atoms. This results in a chemically accurate ligand that still fits in the binding site, but is now ready for small and large chemical group decorations to explore novel scaffold space around the original ligand. The RCG influences the diversity of the generated scaffolds because the backbone of the minimal ligand (based on the RCG) is not altered during scaffold exploration, and only relaxed (energy minimized) in response to the flexibility of the ligand and the protein binding site. This minimal ligand was used as a seed for *de novo* growth into different ligands/scaffolds by utilizing a fragment library in the protein's binding site using the OpenGrowth program[48]. This algorithm was chosen as it provided the most agnostic way to explore the scaffold space around the RCG backbone. It grows ligands inside a protein pocket, minimizes the ligand geometries along with that of residue sidechains, and then scores the new grown ligand. The top 5 ranked ligands that were grown from the RCG-based seed ligand are shown in Figures 8D-8H. The top 20 ranked ligands out of a total of ~900 are shown in the Supplementary **Figure S12**. The scaffolds show a lot of diversity as expected, so this *de novo* growth has the potential to generate novel scaffolds which can result in patentable hit and lead series.

Some of the biggest collections of small molecules and drugs can be found in databases, like the Chemical Abstract Services (CAS) [41] database with ~140 million unique compounds or the GDB-17 database[42] with ~166 billion organic small molecules. These databases are still scratching the surface of the total number of small molecules that can be utilized as ligands for proteins or as drugs. The scaffold exploration starting from the minimal ligand seed that is based on the *reduced chiral graph* can also be carried out in the spirit of the generation of the GDB-17 database[42], which has strong potential and is planned as one of the future directions. The *reduced chiral graphs* can also serve as building blocks for scaffold hopping[46] over which the chemical structure space can be explored very efficiently since this representation captures the core topology of ligands very well.

CONCLUSIONS

Chirality and chiral recognition is one of the important features in the interactions between biomolecules and the biochemistry of biological processes. Here we presented the *Chiral Graphs* and *Reduced Chiral Graphs* as simple and reduced representations of a chiral ligand's structure that focuses on the core topology of the chiral centers in the ligand structure with the primary motivation to probe protein-ligand stereoselectivity. Current stereoselectivity frameworks can handle multiple chiral centers in acyclic molecules only. Our chiral graph representations can take into account the topology of any complex chiral ligand based on the presence of one or more asymmetric atom centers and provides a good structural framework to study protein-ligand stereoselectivity. The current limitations of these representations are that they cannot handle axial/planar chirality containing stereoisomers or cis/trans isomers. As a first application of these graphs, they were applied to all such chiral ligands in the PDB (~14,000+ ligands), as one of the motivations for these tools is their utilization to understand the mechanisms of protein-ligand stereoselective interactions. These reduced representations can be combined with stereoselectivity models like the three-point interaction models or our general stereocenter-recognition model to provide a framework to understand the structural basis of biological stereoselectivity. The compactness of these representations is also valuable in enabling their use during structure-based drug design campaigns where chiral atoms might be added to or removed from hit or lead molecules to optimize their pharmacodynamics and pharmacokinetic profiles. These compact graph representations can also be used in similarity searches and efficient exploration of the vast chemical structure space of small molecules, which will enhance new scaffold discovery. So, the *chiral graphs* represent a general cheminformatics tool with a range of potential applications centered on ligand topologies.

Acknowledgements

This research was supported in part by the startup funds provided to RA by the Department of Chemistry and Biochemistry at California State University, Northridge. RA would like to thank Vidyasankar Sundaresan for helpful discussions on protein-ligand stereoselectivity. We would also

like to thank the reviewers for their helpful comments that improved this manuscript.

Keywords: Stereochemistry • Chirality • Stereoselectivity • Protein-Ligand Interactions • Graph Theory • Drug Discovery • Chemical Structure Space.

References:

[1] L. Pasteur, *Comptes rendus de l'Académie des sciences* **1848**, *26*, 535-538; H. Flack, *Acta Crystallographica Section a* **2009**, *65*, 371-389.

[2] G. Genchi, *Amino Acids* **2017**, *49*, 1521-1533.

[3] R. Bentley, *Molecular asymmetry in biology*, Academic Press, New York,, **1969**; W. L. Alworth, *Stereochemistry and its application in biochemistry; the relation between substrate symmetry and biological stereospecificity*, Wiley-Interscience, New York,, **1972**.

[4] J. A. McCammon, *Curr Opin Struct Biol* **1998**, *8*, 245-249.

[5] H. Y. Aboul-Enein, I. W. Wainer, *The impact of stereochemistry on drug development and use*, Wiley, New York, **1997**; I. W. Wainer, *Drug stereochemistry : analytical methods and pharmacology*, 2nd ed., M. Dekker, New York, **1993**; R. Crossley, *Chirality and the biological activity of drugs*, CRC Press, Boca Raton, **1995**.

[6] E. ARIENS, *European Journal of Drug Metabolism and Pharmacokinetics* **1988**, *13*, 307-308; B. Testa, G. Vistoli, A. Pedretti, J. Caldwell, *Helvetica Chimica Acta* **2013**, *96*, 747-798; V. Campo, L. Bernardes, I. Carvalho, *Current Drug Metabolism* **2009**, *10*, 188-205; H. Lu, *Expert Opinion on Drug Metabolism & Toxicology* **2007**, *3*, 149-158; Z. Shen, C. Lv, S. Zeng, *J Pharm Anal* **2016**, *6*, 1-10; A. Zask, G. A. Ellestad, *Chirality* **2015**, *27*, 589-597.

[7] W. Lenz, *Teratology* **1988**, *38*, 203-215.

[8] FDA, *Chirality* **1992**, *4*, 338-340; R. L. Zeid, in *Chiral Separation Methods for Pharmaceutical and Biotechnological Products* (Ed.: S. Ahuja), Wiley, New York, **2010**, pp. 9-34.

[9] E. M. Carreira, L. Kvaerno, *Classics in stereoselective synthesis*, Wiley-VCH, Weinheim Germany, **2009**.

[10] H. Caner, E. Groner, L. Levy, I. Agranat, *Drug Discov Today* **2004**, *9*, 105-110.

[11] R. BENTLEY, *Nature* **1978**, *276*, 673-676; R. BENTLEY, *Chemistry in Britain* **1994**, *30*, 191-&.

[12] L. Easson, E. Stedman, *Journal of the Chemical Society* **1933**, 1094-1098; L. Easson, E. Stedman, *Biochemical Journal* **1933**, *27*, 1257-1266.

[13] A. OGSTON, *Nature* **1948**, *162*, 963-963; A. OGSTON, *Nature* **1958**, *181*, 1462-1462.

[14] C. DALGLIESH, *Journal of the Chemical Society* **1952**, 3940-3942.

[15] V. Davankov, *Chirality* **1997**, *9*, 99-102.

[16] V. SOKOLOV, N. ZEFIROV, *Doklady Akademii Nauk Sssr* **1991**, *319*, 1382-1383.

[17] A. D. Mesecar, D. E. Koshland, *Nature* **2000**, *403*, 614-615.

[18] V. Sundaresan, R. Abrol, *Protein Sci* **2002**, *11*, 1330-1339.

[19] V. Sundaresan, R. Abrol, *Chirality* **2005**, *17 Suppl*, S30-39.

[20] D. Fu, A. Libson, L. J. Miercke, C. Weitzman, P. Nollert, J. Krucinski, R. M. Stroud, *Science* **2000**, *290*, 481-486.

[21] P. Grayson, E. Tajkhorshid, K. Schulten, *Biophys J* **2003**, *85*, 36-48.

[22] H. Berman, *Faseb Journal* **2013**, *27*.

[23] B. M. Trost, *Pure & Appl. Chem* **1996**, *68*, 779-784.

[24] S. Ekins, J. Mestres, B. Testa, *Br J Pharmacol* **2007**, *152*, 9-20; B. O. Villoutreix, R. Eudes, M. A. Miteva, *Comb Chem High Throughput Screen* **2009**, *12*, 1000-1016; G. Schneider, K.-H. Baringhaus, *Molecular design : concepts and applications*, Wiley-VCH, Weinheim,

2008; E. Glaab, *Brief Bioinform* **2016**, *17*, 352-366.

[25] O. Ivanciuc, *Curr Comput Aided Drug Des* **2013**, *9*, 153-163.

[26] R. Carrasco-Velar, J. O. Prieto-Entenza, A. Antelo-Collado, J. A. Padrón-García, G. Cerruela-García, Á. Maceo-Pixa, R. Alcolea-Núñez, L. G. Silva-Rojas, *SAR QSAR Environ Res* **2013**, *24*, 201-214; K. Birchall, V. J. Gillet, G. Harper, S. D. Pickett, *J Chem Inf Model* **2008**, *48*, 1543-1557; J. R. Fischer, M. Rarey, *J Chem Inf Model* **2007**, *47*, 1341-1353; G. Harper, G. S. Bravi, S. D. Pickett, J. Hussain, D. V. Green, *J Chem Inf Comput Sci* **2004**, *44*, 2145-2156; Y. TAKAHASHI, M. SUKEKAWA, S. SASAKI, *Journal of Chemical Information and Computer Sciences* **1992**, *32*, 639-643.

[27] M. Saito, N. Takemura, T. Shirai, *J Mol Biol* **2012**, *424*, 379-390; C. Shionyu-Mitsuyama, A. Hijikata, T. Tsuji, T. Shirai, *J Struct Funct Genomics* **2016**, *17*, 135-146.

[28] W. IHLENFELDT, Y. TAKAHASHI, H. ABE, S. SASAKI, *Journal of Chemical Information and Computer Sciences* **1994**, *34*, 109-116.

[29] M. Stahl, H. Mauser, *Journal of Chemical Information and Modeling* **2005**, *45*, 542-548.

[30] K. Henrick, Z. Feng, W. F. Bluhm, D. Dimitropoulos, J. F. Doreleijers, S. Dutta, J. L. Flippen-Anderson, J. Ionides, C. Kamada, E. Krissinel, C. L. Lawson, J. L. Markley, H. Nakamura, R. Newman, Y. Shimizu, J. Swaminathan, S. Velankar, J. Ory, E. L. Ulrich, W. Vranken, J. Westbrook, R. Yamashita, H. Yang, J. Young, M. Yousufuddin, H. M. Berman, *Nucleic Acids Res* **2008**, *36*, D426-433; J. D. Westbrook, C. Shao, Z. Feng, M. Zhuravleva, S. Velankar, J. Young, *Bioinformatics* **2015**, *31*, 1274-1278.

[31] D. Dimitropoulos, J. Ionides, K. Henrick, *Curr Protoc Bioinformatics* **2006**, *15*, Unit14.13.

[32] K. Birchall, V. J. Gillet, P. Willett, P. Ducrot, C. Luttmann, *J Chem Inf Model* **2009**, *49*, 1330-1346.

[33] M. Rarey, M. Stahl, *J Comput Aided Mol Des* **2001**, *15*, 497-520.

[34] E. A. Kennewell, P. Willett, P. Ducrot, C. Luttmann, *J Comput Aided Mol Des* **2006**, *20*, 385-394.

[35] R. P. Sheridan, P. Hunt, J. C. Culberson, *J Chem Inf Model* **2006**, *46*, 180-192.

[36] E. Bolton, Y. Wang, P. Thiessen, S. Bryant, R. Wheeler, D. Spellmeyer, *Annual Reports in Computational Chemistry, Vol 4* **2010**, *4*, 217-241; Q. Li, T. Chen, Y. Wang, S. Bryant, *Drug Discovery Today* **2010**, *15*, 1052-1057.

[37] J. Irwin, *Abstracts of Papers of the American Chemical Society* **2016**, *251*; T. Sterling, J. Irwin, *Journal of Chemical Information and Modeling* **2015**, *55*, 2324-2337.

[38] D. Wishart, C. Knox, A. Guo, S. Shrivastava, M. Hassanali, P. Stothard, Z. Chang, J. Woolsey, *Nucleic Acids Research* **2006**, *34*, D668-D672; D. Wishart, Y. Feunang, A. Guo, E. Lo, A. Marcu, J. Grant, T. Sajed, D. Johnson, C. Li, Z. Sayeeda, N. Assempour, I. Iynkkaran, Y. Liu, A. Maciejewski, N. Gale, A. Wilson, L. Chin, R. Cummings, D. Le, A. Pon, C. Knox, M. Wilson, *Nucleic Acids Research* **2018**, *46*, D1074-D1082.

[39] A. Gaulton, L. Bellis, A. Bento, J. Chambers, M. Davies, A. Hersey, Y. Light, S. McGlinchey, D. Michalovich, B. Al-Lazikani, J. Overington, *Nucleic Acids Research* **2012**, *40*, D1100-D1107; A. Gaulton, A. Hersey, M. Nowotka, A. Bento, J. Chambers, D. Mendez, P. Mutowo, F. Atkinson, L. Bellis, E. Cibrian-Uhalte, M. Davies, N. Dedman, A. Karlsson, M. Magarinos, J. Overington, G. Papadatos, I. Smit, A. Leach, *Nucleic Acids Research* **2017**, *45*, D945-D954.

[40] H. Berman, K. Henrick, H. Nakamura, *Nat Struct Biol* **2003**, *10*, 980; A. Kouranov, L. Xie, J. De la Cruz, L. Chen, J. Westbrook, P. Bourne, H. Berman, *Nucleic Acids*

Research **2006**, *34*, D302-D305; P. W. Rose, A. Prlić, C. Bi, W. F. Bluhm, C. H. Christie, S. Dutta, R. K. Green, D. S. Goodsell, J. D. Westbrook, J. Woo, J. Young, C. Zardecki, H. M. Berman, P. E. Bourne, S. K. Burley, *Nucleic Acids Res.* **2015**, *43*, D345-356.

[41] K. A. Hamill, R. D. Nelson, G. G. Vander Stouw, R. E. Stobaugh, *J Chem Inf Comput Sci* **1988**, *28*, 175-179; M. A. Huffenberger, R. L. Wigington, *J Chem Inf Comput Sci* **1975**, *15*, 43-47.

[42] L. Ruddigkeit, R. van Deursen, L. C. Blum, J. L. Reymond, *J Chem Inf Model* **2012**, *52*, 2864-2875.

[43] F. Da Silva, J. Desaphy, D. Rognan, *ChemMedChem* **2018**, *13*, 507-510.

[44] S. Salentin, S. Schreiber, V. J. Haupt, M. F. Adasme, M. Schroeder, *Nucleic Acids Res* **2015**, *43*, W443-447.

[45] A. Wang, C. D. Stout, Q. Zhang, E. F. Johnson, *J Biol Chem* **2015**, *290*, 5092-5104.

[46] Y. Hu, D. Stumpfe, J. Bajorath, *J Med Chem* **2017**, *60*, 1238-1246.

[47] J. Agniswamy, C. H. Shen, Y. F. Wang, A. K. Ghosh, K. V. Rao, C. X. Xu, J. M. Sayer, J. M. Louis, I. T. Weber, *J Med Chem* **2013**, *56*, 4017-4027.

[48] N. Chéron, N. Jasty, E. I. Shakhnovich, *J Med Chem* **2016**, *59*, 4171-4188.

Supplementary Section 1: Materials and Methods

The 26,000+ ligands in the protein databank (PDB) were used as input for mining the data about the ligand structure and obtaining their PDB files to parse the information about their atom connections and coordinates. **Figure S1** shows the work flow of obtaining these data. **Figures S2** through **S4** show further algorithmic details of the method**.** The construction of the *RCGs* involved the reduction of the ligand's topology by first converting the adjacency lists found in the PDB files into dictionaries to remove all hydrogens and non-chiral heavy atoms while keeping track of the atoms' connections as well as the locations around each chiral center, then reconnecting the chiral atoms and placing them back into reduced adjacency lists (see Supplementary **Figure S9** later).

Additionally, Unicon servers [64] were used to convert the 3D coordinates in the PDB files into 2D coordinates for chemically realistic mapping. The Unicon server can also take sdf or SMILES as alternative input formats to convert between different file formats. Here we used the PDB format as an input because it contains the atom names and is convenient when specifying the chiral centers. Also, the file conversion only requires the one-line command:

```
unicon -i lig.pdb -o lig2D.mol2 -g 2
```

("lig" refers to the three-letter code of the ligand name, i.e. QDN for Quinidine).

The 2D coordinates facilitated the utilization of Graph Theory [65] in making the *RCGs* where a chiral center was represented by a *node*, two *nodes* were connected by an *edge*, and a *node* connected to itself by a *self-loop*. This *self-loop* is most commonly found for ring structures with one chiral center. Since nitrogen influences the ligand's conformation when connecting three chiral centers, it was considered as a *pseudonode*.

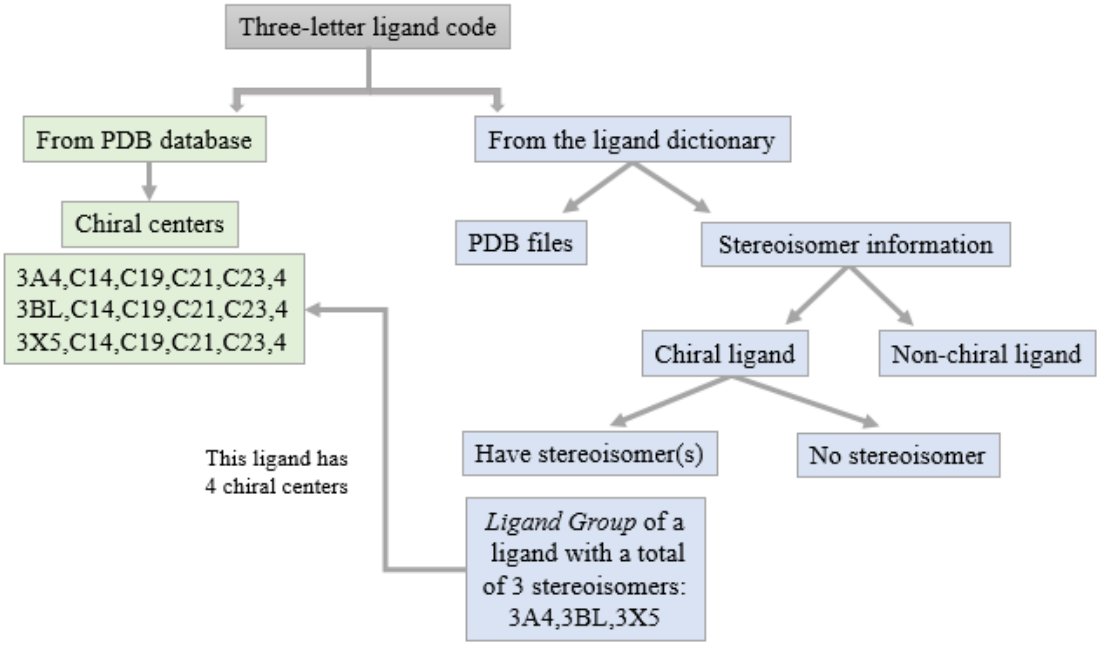


Figure S1. Data mining workflow with an illustrated example of the ligand (2S)-2-(([(3S,4aS,8aR)-2-(biphenyl-4-ylcarbonyl) decahydroisoquinolin-3-yl] methyl) amino)-3-(1H-imidazol-5-yl) propanal. Three-letter ligand code: 3A4.

Using the conventional molecular visualization tools wasn't a viable option since the *RCGs* contain *loops and multi-edges (2 edges* between 2 *nodes*) that are unconventional in chemical structures and can't be displayed clearly. Therefore, the reduced lists were used as inputs for making dot files using the DOT language **(**Supplementary **Figure S10)**, an input format for GraphViz software program [66]. It was initially created by developers to visualize webs and networks as directed graphs, but it also has the capability to display *loops* and *multi-edges.* Here, we needed to visualize our molecular *RCGs* as undirected graphs, so we used the Neato utility in GraphViz [66]. It is used to create the 2D layout of each *RCG* with the coordinates for the *nodes* form the lig.mol2 files using the command line:

```
neato -n2 -Tpdf lig_rcg.dot -o lig_rcg.pdf
```

instead of `dot -Tps lig_rcg.dot -o lig_rcg.pdf,` but with the same dot file as an input.

The *RCG* output can be displayed graphically in GIF, PNG, TIF, SVG, PDF, or PostScript. The original molecules were also generated by GraphViz software as non-reduced *Chiral Graphs (CG)* to compare the difference in structure between the reduced and non-reduced structures.

Data mining: Data mining was done to obtain and examine all structures of the ~ 26,000 ligands in the protein data bank (PDB). The format of each output was a comma-separated CSV file listing each ligand and its stereoisomers parsed form the ligand dictionary in PDBeChem. The three-letter code of each ligand was used as an input to generate and classify different outputs that included information about stereoisomers of each ligand, if any **(Figures S2** through **S4).** The information about the chiral centers and their counts were parsed from the PDB RCSB website and were also placed in CSV files. The PDB files of ligands containing the atoms coordinates and connections were parsed from PDBeChem. All the required data were extracted from RCSB and PDBeChem using python scripts and html parsers **(**Supplementary **Section 2).** An xml parser with a python script was created for obtaining information about protein-ligand contacts from PLIP server **(**Supplementary **Section 3)**.

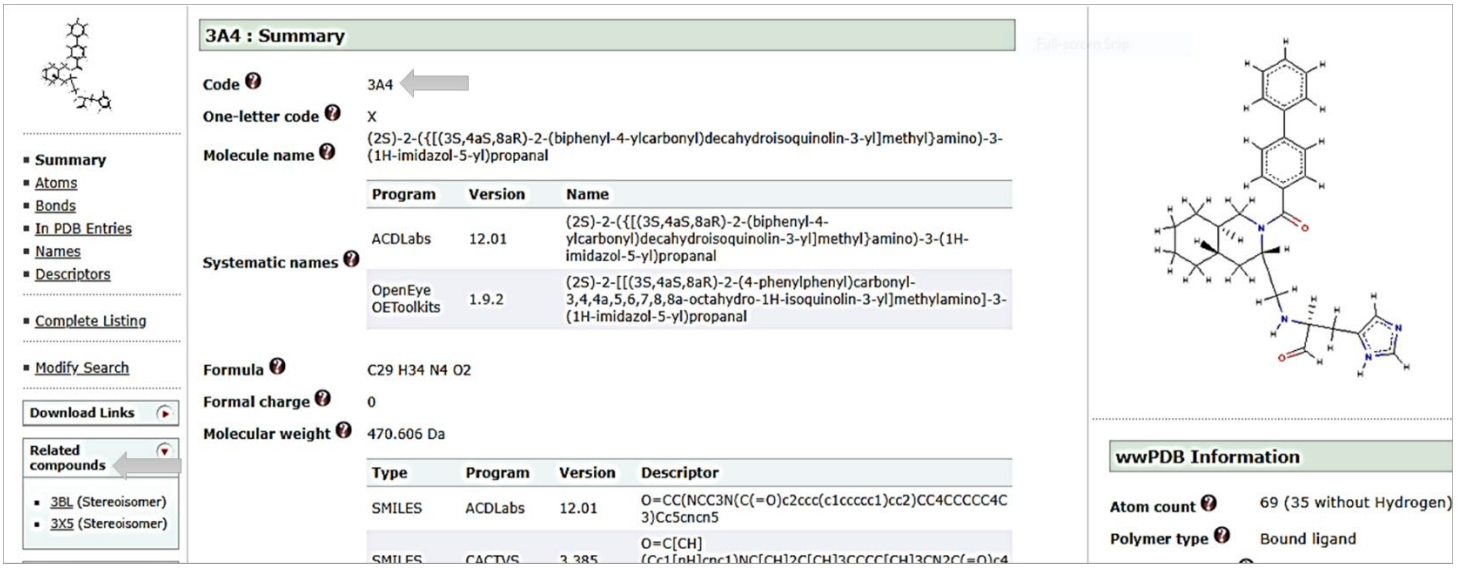
Data Processing and the Construction of Reduced Chiral Graphs: Python scripts were used to process coordinates and atom connections data from the PDB files. The Unicon server was used to convert the 3D coordinates into 2D coordinates and creating mol2 files for each ligand. These 2D coordinates, along with the new reduced atom connections, were then processed into DOT files using python scripts **(**Supplementary **Section 4).**

Visualization: To visualize the 2D *Reduced Chiral Graphs (RCG)* and the *Chiral Graph (CG)* of the non-reduced ligand, we used GraphViz software with Neato utility to display the *RCGs* as undirected graphs. GraphViz was also used to display protein-ligand interactions that were obtained from PLIP server.

The **ChiraLig (**http://chiralig.abrollab.org) web-based server has been created for visualizing and downloading of Chiral Graphs, Reduced Chiral Graphs, and other related files for each of the ~26,000 ligands. More details on the webserver are contained in the Supplementary **Section 5**.

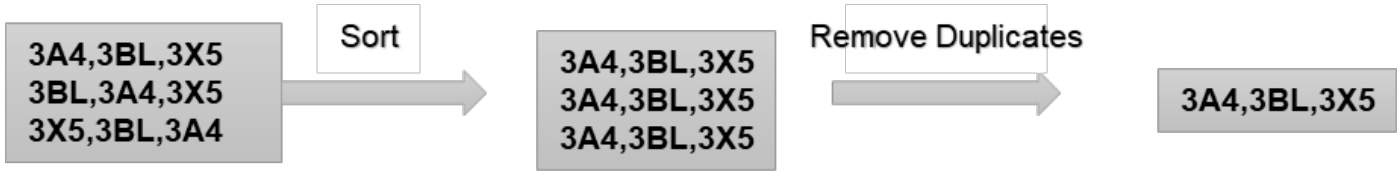
Obtaining Ligands:

Step	Input(s)	Python Script	Output(s)
Step1	**`AllLigands.txt`** 3-letter ligand code (3A4) for the 26,437 ligands in the PDB	**`ligCode.py`** Parses the ligands and stereoisomers form the PDBeChem dictionary of ligands.	**`ligands_isomers.csv`** ligand, stereoisomer 1, stereoisomer 1, ...etc. (3A4,3BL,3X5)



Supplementary Figure S2. PDBeChem ligands dictionary showing an example of parsing the ligand 3A4, and related compound (stereoisomers) 3BL and 3X5.

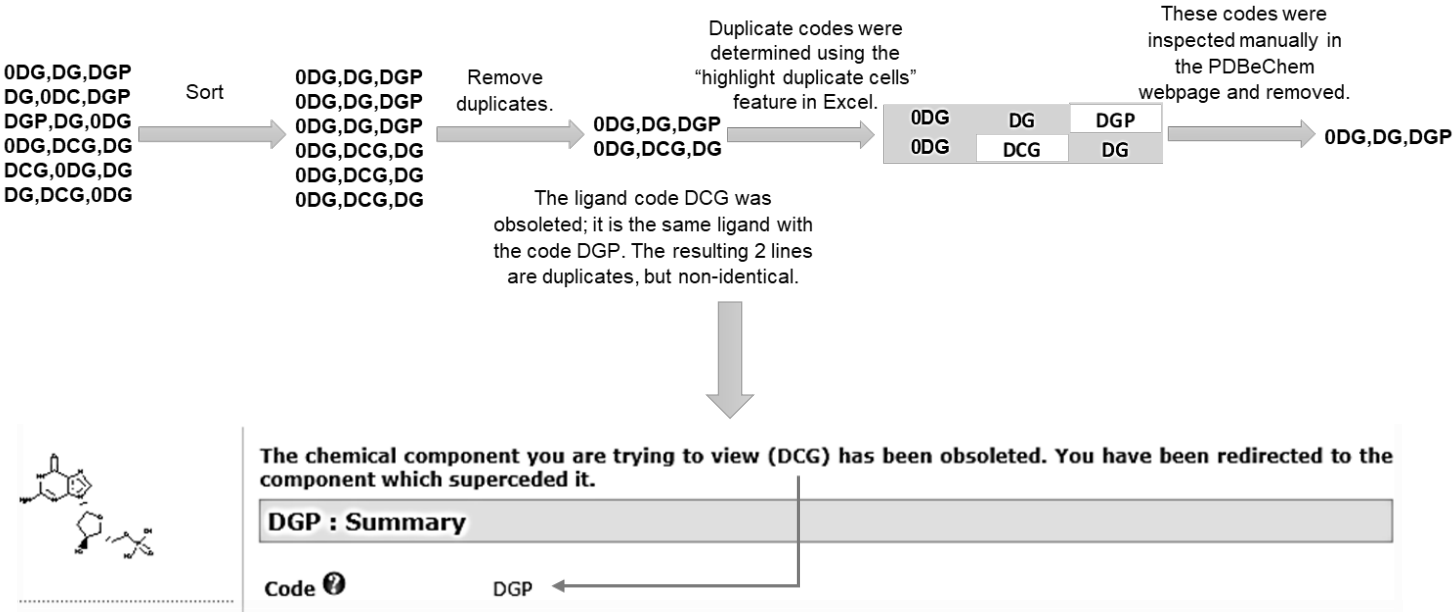
Step 2	**`AllLigands.txt`**, and **`ligands_isomers.csv`**	**`getNo_isomers.py`** Compares stereoisomers in the csv file with the original 3-letter codes file and prints the difference as non-stereoisomeric ligands.	**`Lig_noIsomers.txt`** The remaining ligands after removing the stereoisomeric ligands.
Step 3	**`ligands_isomers.csv`**	**`sort_rmDupl.py`** Removes the duplicates lines.	**`stereoisomers.csv`** Stereoisomers after removing duplicates.



Supplementary Figure S3. An example showing the steps to remove duplicate parsed 3A4 and its stereoisomers. Python "sets" function sorts the names alphabetically, so it would become easier to remove identical lines.

Step 4 — Obsoleted ligand codes were manually removed from the stereoisomers csv file.

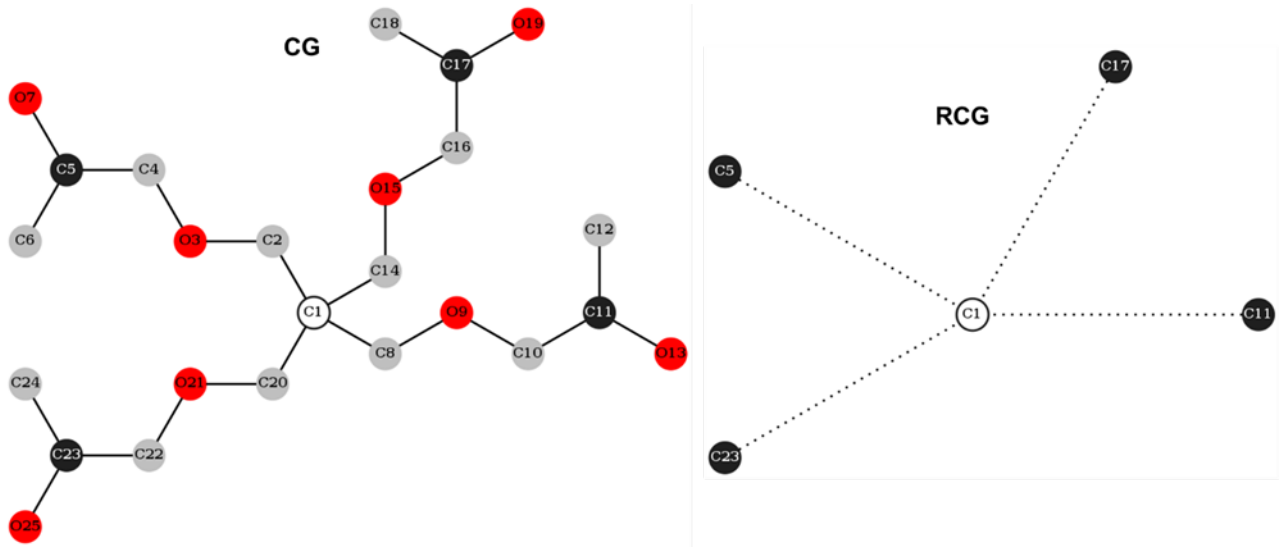
`stereoisomers.csv`
The updated stereoisomers csv file after removing the obsoleted structures.



Supplementary Figure S4. An example of manually removing the obsoleted ligand code DCG (DGP is the same ligand).

Step 5 — **`stereoisomers.csv`** — **`divide.py`**
Divides the stereoisomers csv file into ligand groups based on the number of stereoisomers each ligand has

Ligand Groups:
`pairs.csv`
stereoisomer 1, stereoisomer 2
`three.csv`
stereoisomer 1, stereoisomer 2, stereoisomer 3
`four.csv`
`five.csv`
`six.csv`
`more.csv`

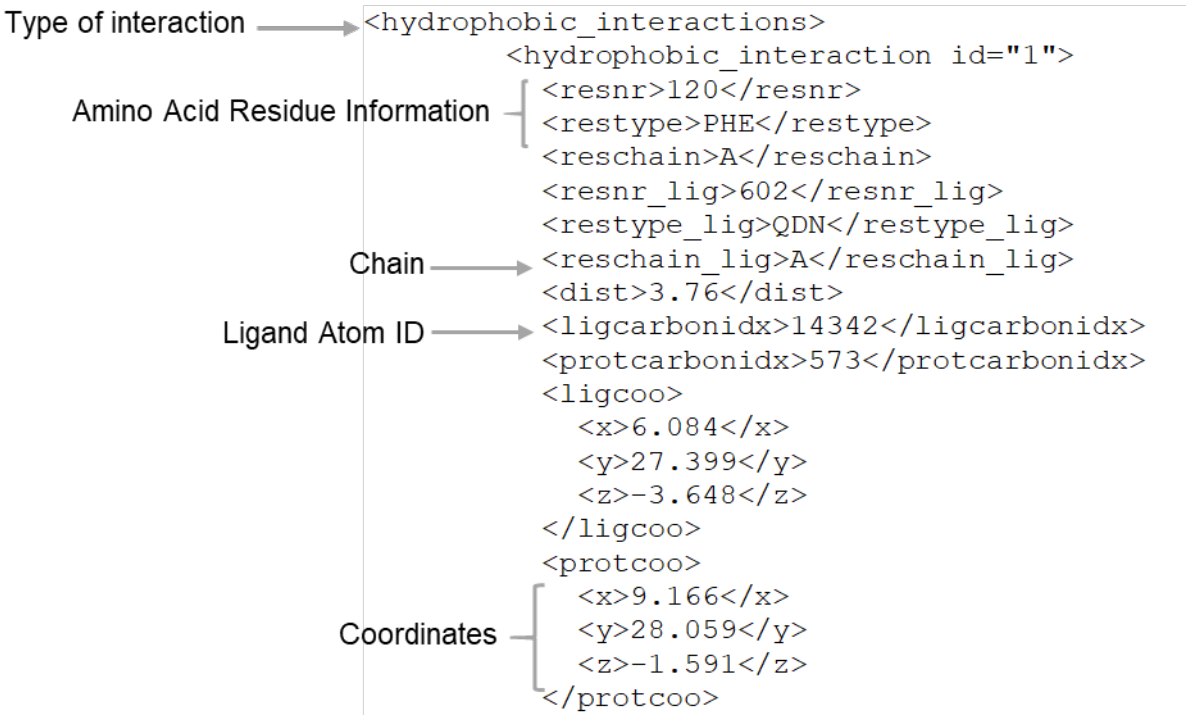


Supplementary Figure S5. *Graph representations* for the symmetrically branched ligand (2S)-1-[3-[(2S)-2-oxidanylpropoxy]-2-[[(2S)-2-oxidanylpropoxy] methyl]-2-[[(2R)-2-oxidanylpropoxy] methyl] propoxy] propan-2-ol. *Ligand Code*: 4TD. *Left: Chiral Graph (CG)*. *Right: Reduced Chiral Graph (RCG)* showing the *prochiral* atom C1 (bold empty circle) as a *pseudonode.*

Supplementary Section 2. Detailed information about the ligand's name, number of chiral atoms, and names of chiral atoms.

Step	Input(s)	Python Script	Output(s)
Step 1	**pairs.csv** **three.csv** **four.csv** **five.csv** **six.csv** **more.csv**	**main_info.py** Parses the detailed stereocenters information (counts, and atom names) for each ligand from the PDB, including the corresponding ligand name for each 3-letter ligand code.	**main_pairs.csv** stereoisomer 1, molecule name, # chiral atoms, names of chiral atoms, stereoisomer 2 stereoisomer 2, molecule name, # chiral atoms, names of chiral atoms, stereoisomer 1 **main_three.csv** **main_four.csv** **main_five.csv** **main_six.csv** **main_more.csv**
Step 2	**main_pairs.csv** **main_three.csv** **main_four.csv** **main_five.csv** **main_six.csv** **main_more.csv**	**cis_trans.py** Removes the cis/trans isomeric ligands in a separate file, and updates the main ligand groups stereoisomeric ligand files.	**cis_trans.csv**, and Updated: **main_pairs.csv** **main_three.csv** **main_four.csv** **main_five.csv** **main_six.csv** **main_more.csv**

Supplementary Section 3. Xml file format and selected sections of the Xml parser form the python script.



Supplementary Figure S6. Sections from the Xml file from PLIP server containing protein-ligand interactions. An example of one interaction between the ligand **QDN** and the protein it's bound to; **4wnu**, chain **A**.

Supplementary Section 4. Detailed steps of the *reduced chiral graphs* construction.

Step	Input(s)	Python Script	Output(s)
Step 1	**`main_info.csv`** Stereoisomer 1, molecule name, # chiral atoms, names of chiral atoms, stereoisomer 2. Stereoisomer 2, molecule name, # chiral atoms, names of chiral atoms, stereoisomer 1	**`getPDB.py`** Gets ligand PDB file containing atom connections and coordinates.	**`lig.PDB`**
Step 2	**`lig.PDB`**	**`pdb2D.py`** Uses Unicon utility to convert 3D coordinates into 2D coordinates.	**`lig_2D.mol2`**

PDB File

```
                      X       Y       Z
ATOM      1  C1  FRU     0       0.791   1.055  -1.776  1.00 20.00           C+0
ATOM      2  C2  FRU     0       0.144  -0.149  -1.090  1.00 20.00           C+0
ATOM      3  C3  FRU     0      -1.304   0.188  -0.689  1.00 20.00           C+0
ATOM      4  C4  FRU     0      -1.307   0.081   0.857  1.00 20.00           C+0
ATOM      5  C5  FRU     0       0.198   0.230   1.193  1.00 20.00           C+0
ATOM      6  C6  FRU     0       0.518  -0.427   2.536  1.00 20.00           C+0
ATOM      7  O1  FRU     0       2.138   0.735  -2.130  1.00 20.00           O+0
ATOM      8  O2  FRU     0       0.154  -1.273  -1.972  1.00 20.00           O+0
ATOM      9  O3  FRU     0      -2.215  -0.752  -1.261  1.00 20.00           O+0
ATOM     10  O4  FRU     0      -2.072   1.137   1.440  1.00 20.00           O+0
ATOM     11  O5  FRU     0       0.858  -0.467   0.115  1.00 20.00           O+0
ATOM     12  O6  FRU     0       1.919  -0.319   2.798  1.00 20.00           O+0
CONECT    1    2    7
CONECT    2    1    3    8   11
CONECT    3    2    4    9
CONECT    4    3    5   10
CONECT    5    4    6   11
CONECT    6    5   12
CONECT    7    1
CONECT    8    2
CONECT    9    3
CONECT   10    4
CONECT   11    2    5
CONECT   12    6
```

Mol2 File

```
                X          Y
 1 O6          39.3103     8.1696     0.0000 O.3          1    LIG    0.0000
 2 O1         -36.2675     6.5423     0.0000 O.3          1    LIG    0.0000
 3 O4          21.6252    32.2318     0.0000 O.3          1    LIG    0.0000
 4 O3         -10.6066    37.3368     0.0000 O.3          1    LIG    0.0000
 5 O2         -22.2638     1.1668     0.0000 O.3          1    LIG    0.0000
 6 O5           0.0000     0.0000     0.0000 O.3          1    LIG    0.0000
 7 C6          26.7302     0.0000     0.0000 C.3          1    LIG    0.0000
 8 C1         -24.6103    15.9821     0.0000 C.3          1    LIG    0.0000
 9 C4          11.0186    21.6252     0.0000 C.3          1    LIG    0.0000
10 C5          13.3651     6.8099     0.0000 C.3          1    LIG    0.0000
11 C3          -3.7967    23.9717     0.0000 C.3          1    LIG    0.0000
12 C2         -10.6066    10.6066     0.0000 C.3          1    LIG    0.0000
 1        1        7 1
 2        1       13 1
 3        2        8 1
 4        2       14 1
 5        3        9 1
 6        3       15 1
 7        4       11 1
 8        4       16 1
 9        5       12 1
10        5       17 1
11        6       10 1
12        6       12 1
```

Supplementary Figure S7. PDB and mol2 file formats.

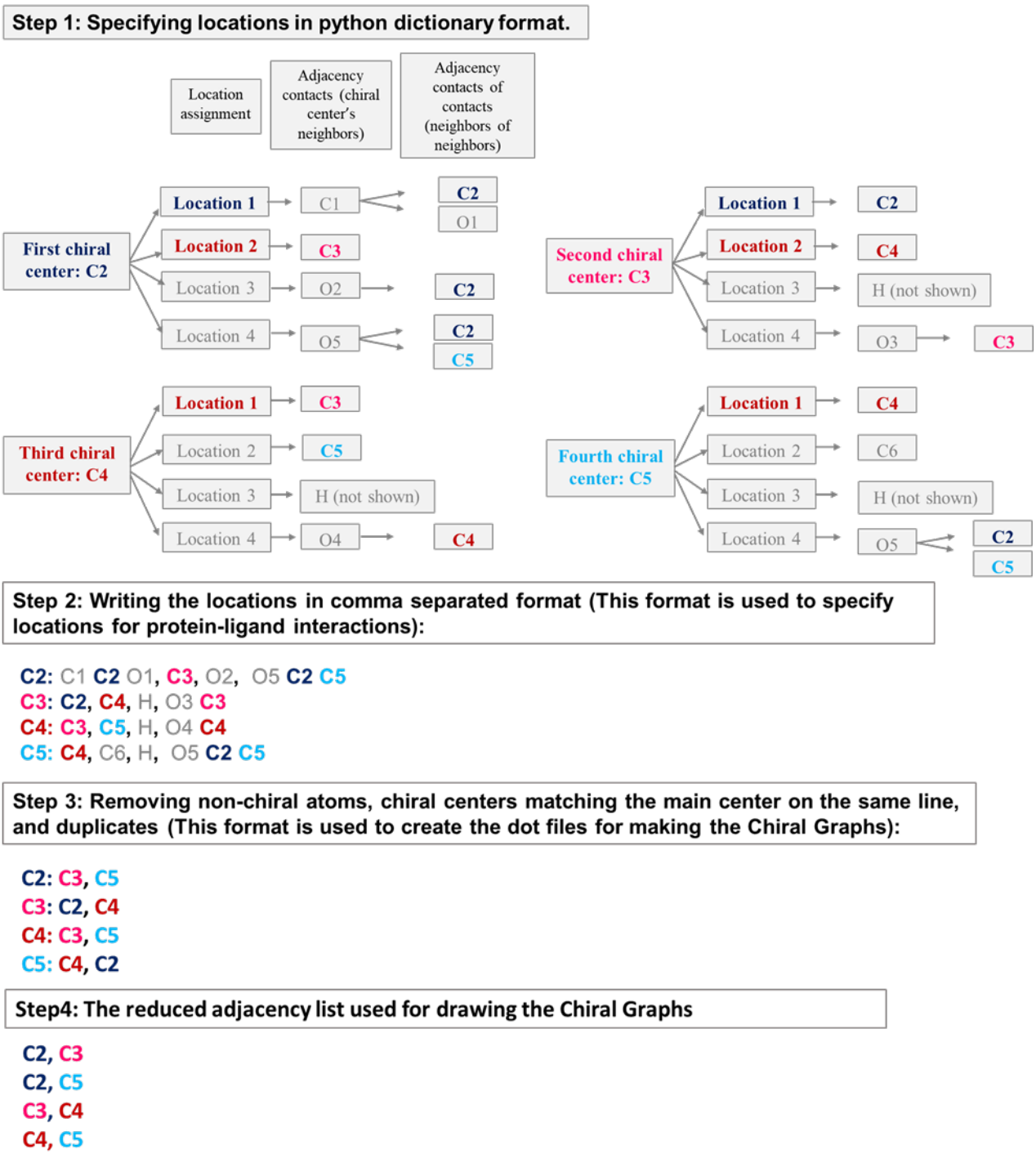
Step 3	**lig.PDB**	**adjList.py** Converts atom IDs in the CONECT section into atom names. Extracts nitrogen heteroatoms as potential pseudonodes, if any.	**lig-cgCONECT.txt** **lig-PN.txt**

```
CONECT   1    2    7
CONECT   2    1    3    8   11
CONECT   3    2    4    9
CONECT   4    3    5   10
CONECT   5    4    6   11
CONECT   6    5   12
CONECT   7    1
CONECT   8    2
CONECT   9    3
CONECT  10    4
CONECT  11    2    5
CONECT  12    6
```

```
CONECT C1 C2 O1
CONECT C2 C1 C3 O2 O5
CONECT C3 C2 C4 O3
CONECT C4 C3 C5 O4
CONECT C5 C4 C6 O5
CONECT C6 C5 O6
CONECT O1 C1
CONECT O2 C2
CONECT O3 C3
CONECT O4 C4
CONECT O5 C2 C5
CONECT O6 C6
```

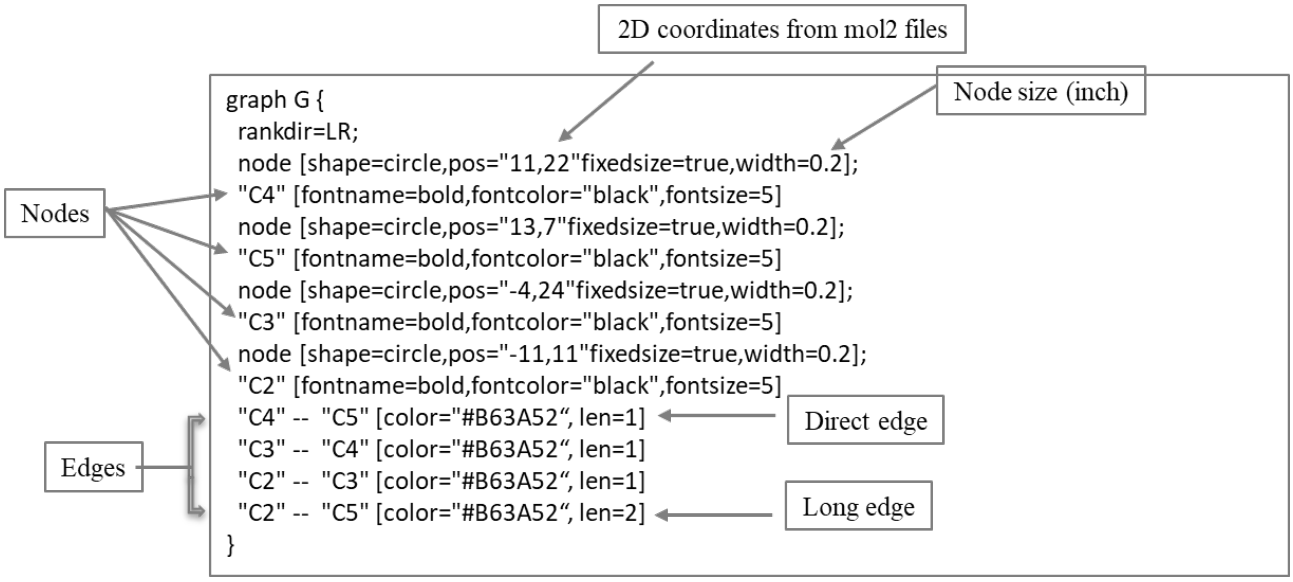
Supplementary Figure S8. Converting atom IDs into atom names to facilitate reading the chiral atoms from the chiral atoms lists, and to facilitate dealing with the different numbering schemes in PLIP server files.

Step 4	`lig-cgCONECT.txt`	`reduce_list.py` Determines the 4 locations around each tetrahedral chiral center. Reduces atom connections into an adjacency list with chiral centers only.	`lig-reduced.txt`



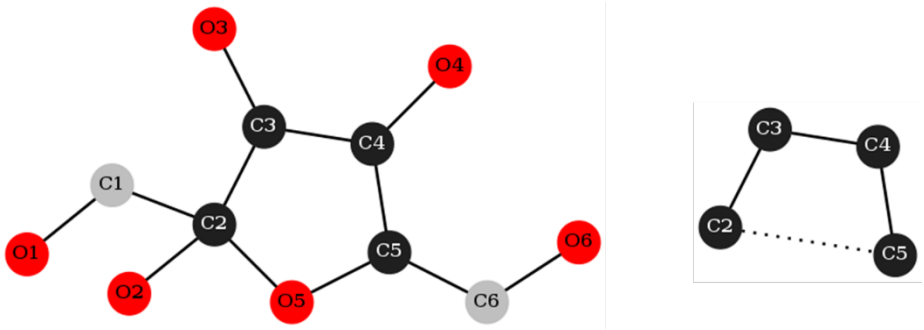
Supplementary Figure S9. Steps of generating *Reduced Chiral Graphs* with an illustrated example of the ligand fructose. Three-letter Ligand Code: FRU.

Step 5	**lig-PN.txt** **main_info.csv**	**PN.py** and **addPN.py** Determines whether the nitrogen atom is a pseudonode by counting the number of its occurrences. 3 or more occurrences means that the nitrogen atom connects three chiral centers, and it can be added to the main file if it wasn't already listed as a chiral center.	**PNmain_info.csv**
Step 6	**PNmain_info.csv** **lig-reduced.txt**	**PNreduce_list.py** Updates the reduced adjacency list after adding pseudonodes.	**PNlig-reduced.txt**
Step 7	**PNlig-reduced.txt** **lig_2D.mol2**	**makeDOT.py** Uses the 2D coordinates and the reduced chiral centers connections to construct new file in DOT language.	**lig-rcg.dot**

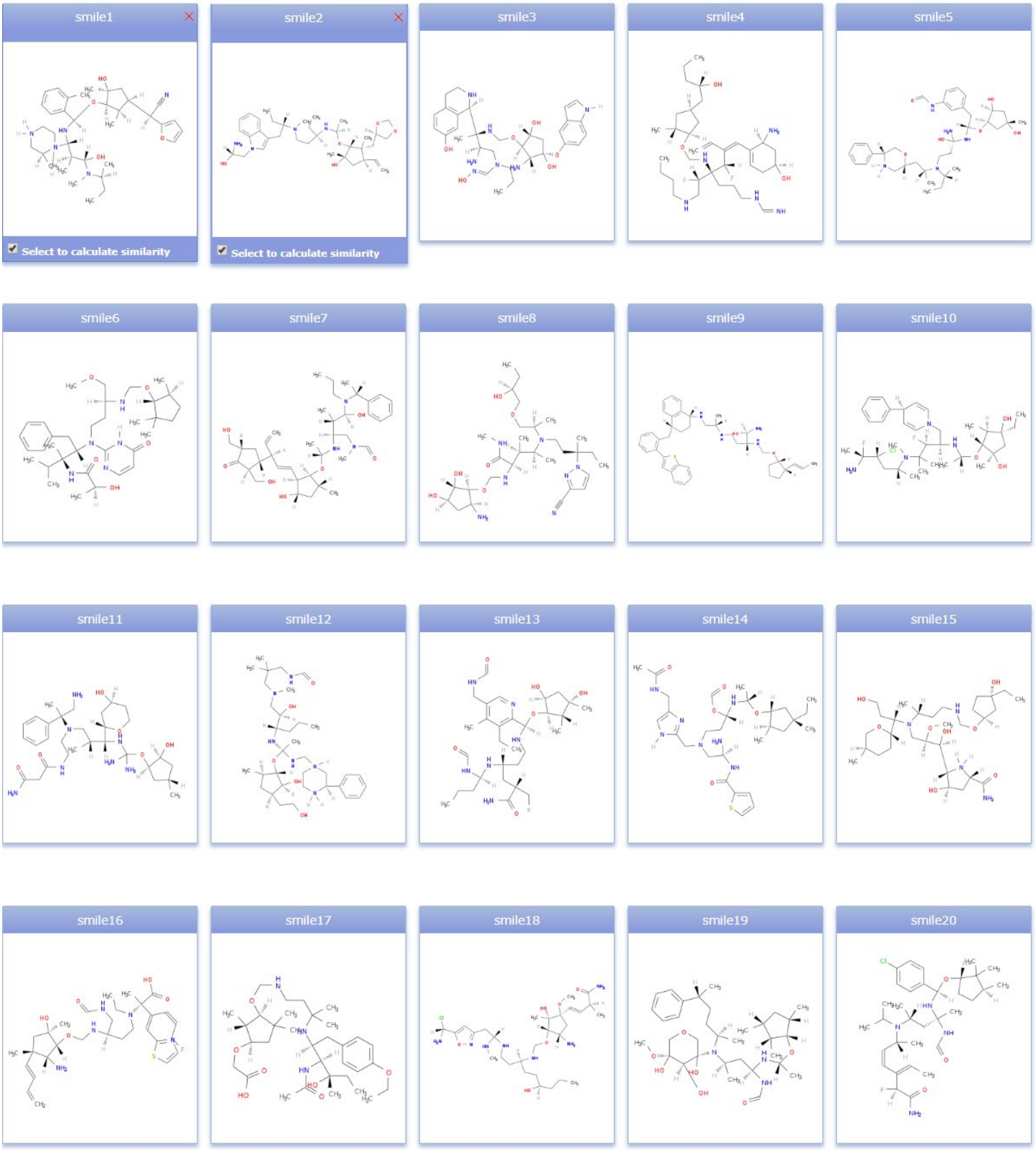


Supplementary Figure S10. Dot file components for the ligand FRU (fructose)

Step 8	**lig-rcg.dot**	**graph.py** Uses Neato utility with GraphViz for RCG visualization.	**lig-rcg.pdf** **lig-rcg.png**



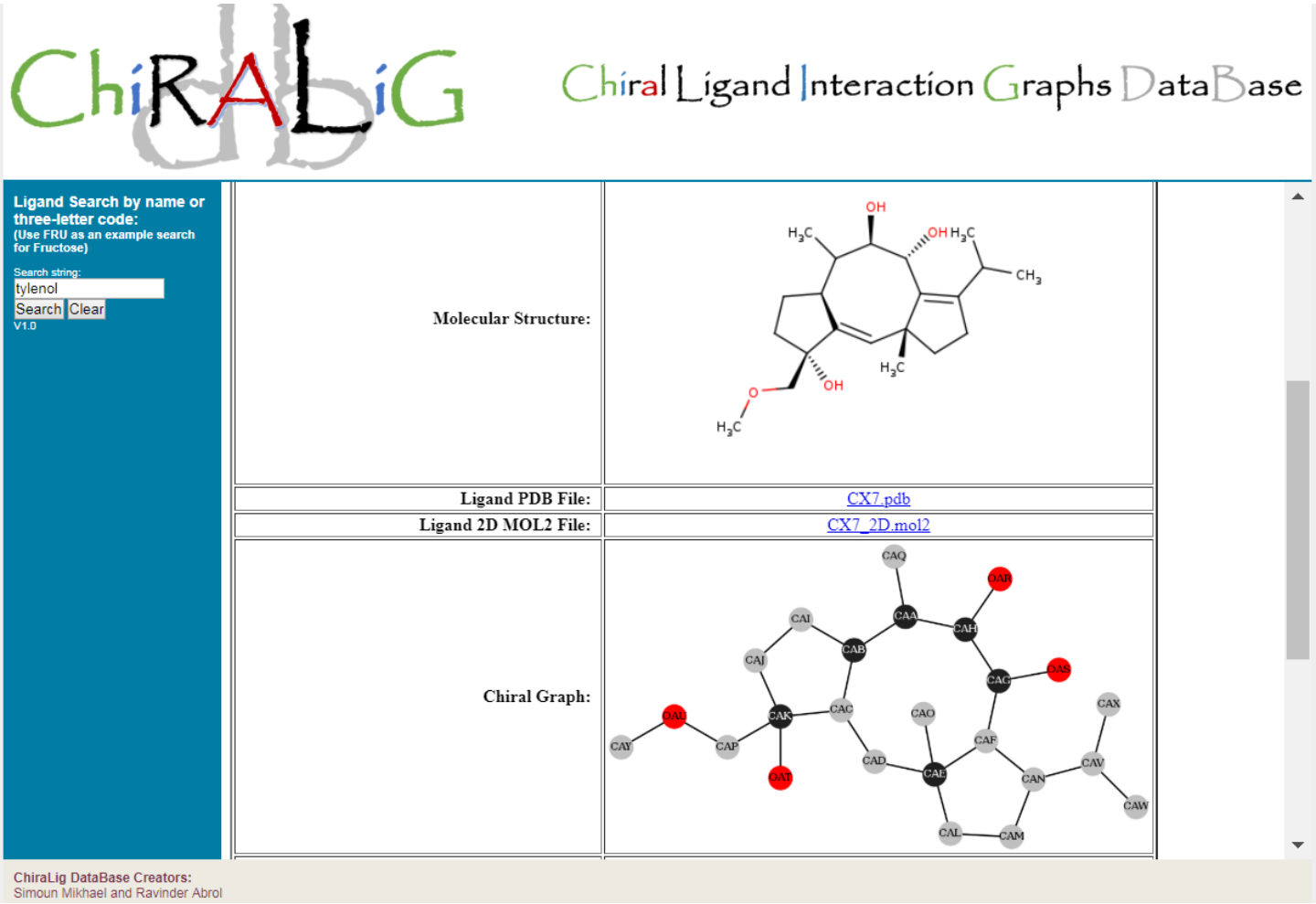
Supplementary Figure S11. Graph representations for the cyclic ligand Fructose. *Ligand Code*: FRU. *Left:* Chiral Graph (CG). *Right*: *Reduced Chiral Graph (RCG)*.



Supplementary Figure S12: The top 20 hits from scaffold exploration using OpenGrowth program starting from a reduced chiral graph based seed molecule.

Supplementary Section 5: Description of the ChiraLig webserver (http://chiralig.abrollab.org)

Here is an example search result from the webserver:



INPUT:

Left bar provides a search field to provide ligand name or three-letter code as input. If partial name is provided, then all ligands with the name match will be displayed.

CURRENT OUTPUT:

Chiral Graph dot file, CONECT file, and pdf file for the image

Reduced Chiral Graph dot file, CONECT file, and pdf file for the image

Ligand's mol2 file

Ligand's pdb file

Names of all chiral atoms in the ligand

Ligand code for other stereoisomers of the input ligand in the PDB

PLANNED ADDITIONAL OUTPUT:

Protein-Ligand interaction map as a chiral graph, dot file, and pdf file

Protein-Ligand interaction map as a reduced chiral graph, dot file, and pdf file

Link to and list of all PDB files containing the input ligand