Small Molecule Rotamers Enable Simultaneous Optimization of Small Molecule and Protein Degrees of Freedom in ROSETTALIGAND Docking

Kristian W. Kaufmann, Kristin Glab, Ralf Mueller, Jens Meiler

Department of Chemistry
Vanderbilt University
465 21st Ave. South
Nashville, TN 37209
kristian.w.kaufmann@vanderbilt.edu
jens.meiler@vanderbilt.edu

Abstract: We introduce small molecule rotamers into the rotamer search protocol used in ROSETTA to model small molecule flexibility in docking. Rosetta, a premier protein modeling suite, models side chain flexibility using discrete conformations observed in the Protein Data Bank (PDB). We mimic this concept and build small molecule rotamers based on conformations from the Cambridge Structural Database. We evaluate the small molecule rotamer generation protocol on a test set of 628 conformations, taken from the PDBBind database, of small molecules with ≤ 6 rotatable bonds. Our protocol generates ensembles in which the closest conformation on average is 0.45 ± 0.31 Å RMSD from the crystallized conformation. Furthermore, in two sets of docking benchmarks the native ligand position and conformation is found within the top 1 % of models by energy in 72% and 90% of all cases.

1 Introduction

Representing protein flexibility through side chain rotamers [DK93] has been central to the success of protein structure prediction, protein docking, and protein design. This discretization of protein side chain conformations observed in the Protein Databank is, for example, used by the ROSETTA program in the *de novo* prediction of protein structure [BMB05]. Furthermore, rotamers form critical components of successful protein docking and protein design strategies such as ROSETTADOCK [GMW+03] [SFWB05] and ROSETTADESIGN [KDI+03][KOK+01][DKC+03]. Finally, ROSETTA incorporates the rotamer probability when performing alanine scanning mutagenesis to identify key residues in protein-protein interfaces (hot spots) [KKB04]. The above success of rotamers for modeling protein side chain flexibility makes adapting the concept for small molecule flexibility attractive.

Leach first introduced using rotamers in modeling small molecule flexibility during docking [Lea94]. He took small molecule conformations in local minima of a molecular dynamics forcefield as rotamers. However, Leach observed a failure of the energy function on docking of phosphocholine to the antibody McPc 603. We independently implemented a similar method using rigid ligands and full side chain flexibility in the ROSETTA [MB06] protein modeling suite. The ROSETTALIGAND energy function identified native conformations for 71 of 100 small molecule protein complexes in a self docking test and 14 of 20 small molecule protein complexes in a ”cross docking“ benchmark. In the cross docking benchmark, a small conformational ensemble containing 10 conformations, one of which was close to the crystallized conformation, was used to simulate small molecule flexibility. In the present work it is our objective to simulate small molecule flexibility using small molecule rotamers generated from a crystal structure database of small molecules. This setup mirrors the amino acid side chain rotamer approach used in ROSETTA for small molecules and thus capitalizes on “knowledge based” rotamers and energy functions deemed responsible for the success of ROSETTA.

In an analogous manner to the amino acid side chain rotamers, we employ small molecule crystal structures from the Cambridge Structural Database (CSD) [All02] to construct small molecule rotamers. Unlike amino acid side chains in the Protein Data Bank (PDB), in the case of small molecules we lack multiple conformations of the same configurational chemistry. Instead, torsion profiles are created from chemical similar groups. OMEGA, a highly regarded program for generating small molecule conformations, makes use of such torsion profiles extracted from the CSD. OMEGA generates conformational ensembles from overlapping fragments in a rule based manner using torsion profiles[BGG03]. Perola and Charifson, in a study of crystallized bioactive small molecules, found OMEGA to be the best available tool for generating ensembles containing the bioactive conformation [PC04].

Most current small molecule docking programs approach docking from the perspective of the small molecule. In contrast, ROSETTALIGAND approaches small molecule docking from the perspective of protein modeling. We hypothesize that ROSETTALIGAND will more accurately represent small molecule protein interactions because of its focus on the protein point of view which allows the accurate simulation of protein flexibility and associated energetics. In our previous paper we demonstrated the utility of the ROSETTA energy function to discriminate native-like models [MB06]. Here our objective is to demonstrate that the concept of rotamers in protein structure prediction can be extended to small molecules. We show that small molecule rotamers can be created using crystal structure data. In addition, these small molecule rotamer ensembles contain conformations similar to the bioactive conformation, in particular for small molecules with a number of torsions similar to those in protein side chains ($\leq$ 6 rotatable bonds). We show that these rotamer ensembles successfully simulate small molecule flexibility in small molecule docking benchmarks.

2 Methods

2.1 Creating Torsion Profiles from the Cambridge Structural Database

We use 28 atom types to describe atoms in small molecules defined by element type, hybridization state, and number of bonded hydrogens [MMWM02]. We measure non-hydrogen atom torsions for each atom type pairing for all structures in the Cambridge Structural Database (CSD) [All02], excluding torsions in ring systems. Histograms are constructed for every pair of the 28 atom types using bins with a width of 10°. Histograms with less than 100 data points are excluded as containing too little information. The distributions are symmetrized by summing counts of symmetry equivalent bins.

A knowledge based torsion energy to model the interactions between atoms separated by three bonds is calculated using the inverse Boltzmann equation (Eq. 1)

$$E_i = -\log P_{torsion} = -\log\left(\frac{N_i + 1}{N_{tot} \times P_{i,ran}}\right) \qquad (1)$$

where $P_{torsion}$ is the propensity of the torsion, N_i is the number of counts in a bin, N_{tot} is the total number of torsions observed for this type, $P_{i,ran}$ is the probability of selecting the bin from a uniform distribution. The propensity of the torsion is the probability of the torsion divided by the background probability the torsion value occurring by chance. The pseudo count of 1 is added to avoid zero probability bins, which result in infinite energies. The background probability is drawn from the uniform distribution since we assume that no other forces bias the torsions observed in the CSD. The weighting of the other internal energies will counterbalance the error introduced by this assumption. The minima in the energy profiles form the set of allowed dihedral angles for the rotamer ensemble generator.

2.2 Small Molecule Rotamer Ensemble Generation

The small molecule ensemble generation protocol (see Figure 1a) creates an ensemble of acceptable energy rotamers. The protocol maximizes coverage of the conformational space accessible to the small molecule by maximizing pair-wise root mean squared deviation (RMSD) for all rotamers. Starting from a conformation with idealized bond lengths and angles, a set of dihedral angles is chosen from the minima of the appropriate torsion profiles. Rotamers containing overlapping atoms are discarded. If the energy is acceptable then the rotamer is provisionally accepted. Otherwise, a new set of dihedral angles are chosen. Using this protocol a list of 10,000 candidates is obtained for pruning. The pruning first selects the lowest energy rotamer from this list and makes it the first rotamer of the ensemble considered for docking.

The energy incorporates van der Waals interaction for atoms separated by four or more covalent bonds [KOK+01], the knowledge based torsion energy described in the previous section, an intra-molecular hydrogen bonding term [MK05], a desolvation energy based on the Lazaridis-Karplus approximation [LK99], and a coulomb electrostatics term [KOK+01].

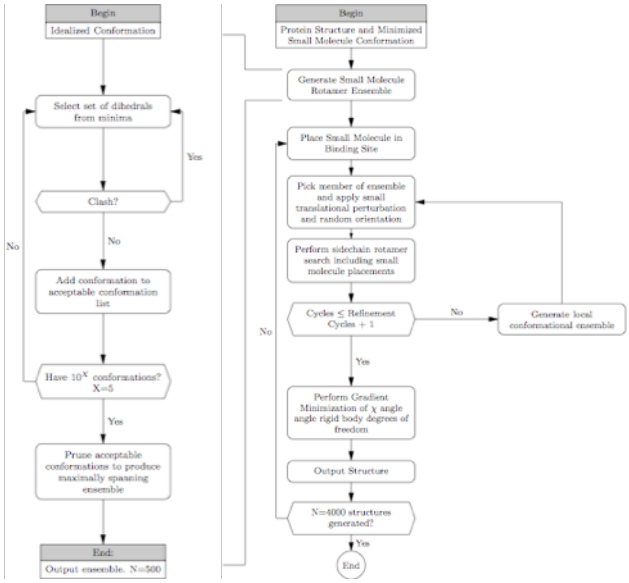

Figure 1. Small molecule docking protocol incorporates a) a rotamer ensemble generating protocol into b) a monte carlo search of protein side chain conformations**.**

Next the protocol iteratively identifies the candidate rotamer that has the largest RMSD to all current members of the docking ensemble and adds it to this ensemble. The protocol continues until the desired number of 500 rotamer has been reached or all candidate rotamers are within a user defined cutoff of 0.2 Å RMSD of one member of the docking ensemble.

2.3 Flexible Small Molecule Docking

Given a protein structure and small molecule conformation the protocol (see Figure 1b) first generates a conformational ensemble for the small molecule. Next iteratively conformations are chosen from the ensemble and placed at a random position and orientation within the manually defined biding site.

The first 100 non-clashing placements are incorporated as small molecule rotamers into the protein side-chain rotamer conformational optimization. After completion of this optimization a local ensemble of up to 100 rotamers is created for refinement by allowing small random changes sampled from a uniform distribution of [-5°, +5°] to all rotatable bonds of the optimized small molecule conformation. After four rounds of side chain optimization with this local discrete conformational ensemble, a gradient minimization of the amino acid side chain χ angles and the small molecule position and orientation take the structure to a local minimum. This structure is then written out. The sequence is repeated until 4,000 models have been generated.

2.4 Small Molecule Flexibility Benchmark Sets

Compounds for the ensemble generation test set were taken from the 2007 PDBBind database [WFLW04]. All molecules with ≤ 6 rotatable non-hydrogen atom torsions were selected.

Two docking benchmarks were carried out. The self docking benchmark tests whether our protocol recovers the correct position, orientation, and conformation of a small molecule in the protein crystal structure solved with that same small molecule. Using the protein structure crystallized with the small molecule ensures the backbone of the protein is in the correct conformation for binding of this substance. The structures used in the self docking benchmark are listed in Table 1. The set contains 10 small molecules crystallized in 7 proteins. The cross docking benchmark is comprised of the same small molecules, but uses protein coordinates from other crystal structures. Hence, the cross docking benchmark assesses the capacity of the protocol to recover the placement of a small molecule in a real world situation where the protein was not crystallized with the small molecule of interest. The structures used are listed in Table 1. Meiler and Baker previously evaluated all structures in both docking benchmarks [MB06]. The set was reduced to contain only small molecules with ≤ 6 rotatable non-hydrogen atom torsions.

Table 1: Crystal Structures forming the small molecule docking benchmark sets

Self docking protein structure	small molecule	# of torsions	Cross docking protein structures
1aq1 human Cyclin Dependent Kinase 2	Straurosporine	1	1dm2
1dm2 human Cyclin Dependent Kinase 2	Hymenialdisine	0	1aq1
1dbj IGG1-κ DB3 FAB	Aetiocholanolone	0	2dbl
2dbl IGG1-κ DB3 FAB	5-α-pregnane-3-β-ol-hemisuccinate	6	1dbj
1pph Trypsin	m-aminophenyl-3-alanine	5	1ppc
1p8d Liver X receptor β small molecule binding domain	24(S),25-epoxycholesterol	4	1pq6,1pqc
2ctc carboxypeptidase A	L-phenyl lactate	3	7cpa
2prg small molecule binding domain of peroxisome proliferator activated receptor γ	2,4-thiazolidiinedione, 5-[[4-[2-(methyl-2-pyridinylamino) ethoxy] phenyl]methyl]-(9cl)	5	1fm9
4tim Triosephosphate isomerase	2-phosphoglyceric acid	4	6tim
6tim Triosephosphate isomerase	SN-Glycerol-3-phosphate	4	4tim

3 Results and Discussion

3.1 Small Molecule Flexibility Benchmark Sets

The torsion profiles generated cover 103 common bond types (see supplement). The profiles obtained show similar characteristics to profiles seen in the AMBER [WWC+04] forcefield (see Figure 2a).

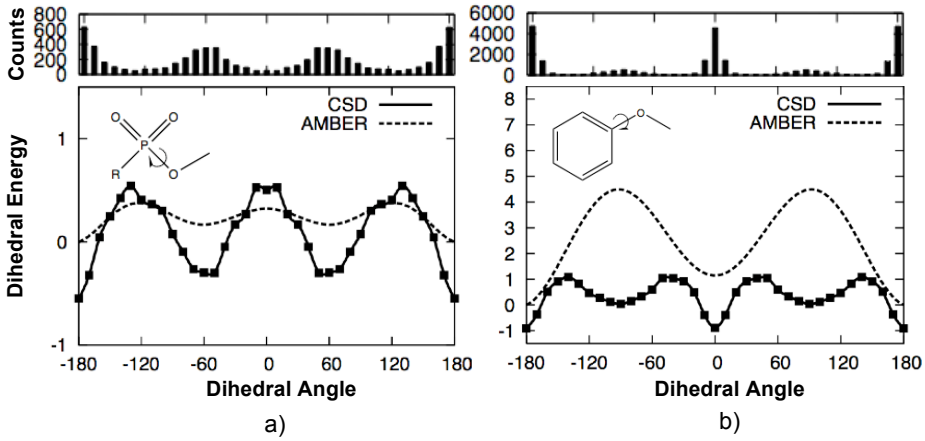

Figure 2 Torsion profiles for phosphate ether and aromatic carbon oxygen bond. Histograms for a) phosphate ether torsion and b) aromatic carbon oxygen torsion yield energy profiles for phosphate ether torsion and aromatic carbon oxygen torsion using the inverse Boltzmann equation

However, some profiles exhibit minima not present in the AMBER forcefield. The aryl oxygen profile, shown in Figure 2b, displays additional minima at ± 90º. Klebe and Meitzner found that these additional minima arise from meta substituted compounds [KM94]. The additional minima give the CSD torsion profiles an advantage, since they allow the ensemble generator to sample conformations that might otherwise be excluded.

3.2 Small Molecule Rotamer Ensemble Generation

The ensemble generator created ensembles for 628 small molecules with ≤ 6 rotatable bonds. The atomic coordinates of no two conformations within the ensemble were allowed to be closer 0.2 Å RMSD. Ten thousand conformations were generated while constructing the ensemble. The final ensembles contained up to 500 conformations. On the set of 628 molecules, the ensemble generator produced a rotamer with 0.46 ± 0.31 Å RMSD to the crystallized conformation. As expected, the accuracy decreases from 0.14 ± 0.16 Å RMSD to 0.79 ± 0.32 Å RMSD as the number of rotatable torsion angles increases from 1 to 6 (see Table 2). Improvement of these numbers might be possible by increasing the size of the ensemble,

Table 2. Performance of Rotamer ensemble Generator evaluated by computing rmsd of closest and furthest conformation in the ensemble to the crystallized conformation of the small molecule

Number of Torsions	Number of Molecules	Average RMSD of closest conformation	Average RMSD of furthest conformation
1	92	0.14 ± 0.16	1.12 ± 0.47
2	118	0.33 ± 0.26	1.74 ± 0.69
3	118	0.41 ± 0.22	2.13± 0.62
4	135	0.47 ± 0.21	2.45 ± 0.69
5	97	0.61 ± 0.30	2.83± 0.81
6	118	0.79 ± 0.32	3.07 ± 0.87
Overall Totals	628	0.46 ± 0.31	2.23± 0.94

and increasing the number of rotamers generated during construction of the ensemble. The additional cost of such increases may outweigh the benefits.

3.2 Flexible Small Molecule Docking

The small molecule docking results are summarized in Table 3. For the self docking, 9 of the 10 cases show a native-like model in the top 1 % by energy. In 7 of the 10 cases the top ranked model is native-like. For the cross docking benchmark 8 of 11 cases show a native-like structure in the top 1 % by energy.

Table 3. Summary of Small Molecule docking benchmark

Source Structure for small molecule	Source Structure for protein	Rank by energy of best structure recapturing the binding mode	RMSD of best structure recapturing the binding mode
Self Docking Results			
1aq1	1aq1	**1**	**0.56**
1dm2	1dm2	**1**	**0.31**
1dbj	1dbj	**1**	**1.36**
2dbl	2dbl	**1**	**1.45**
1p8d	1p8d	**1**	**1.63**
1pph	1pph	*6*	*1.49*
2prg	2prg	639	1.94
2ctc	2ctc	**3**	**0.82**
4tim	4tim	**1**	**1.87**
6tim	6tim	**1**	**1.77**
Cross Docking Results			
1aq1	1dm2	4296	1.87
1dm2	1aq1	**1**	**0.56**
1dbj	2dbl	**1**	**1.80**
2dbl	1dbj	468	3.49
1p8d	1pq6	181	1.62
1p8d	1pqc	*10*	*1.28*
1pph	1ppc	*2*	*1.96*
2ctc	7cpa	*3*	*0.95*
2prg	1fm9	*16*	*2.02*
4tim	6tim	*2*	*1.90*
6tim	4tim	*5*	*1.77*

In only 2 of the 11 cases was the top ranked model a native-like model. In Figure 3a the RMSD energy plot demonstrates that the scoring function identifies the native binding mode as the most favorable (see Figure 3c). However, in other cases the RMSD energy plots appear like that of Figure 3b. Some models are present in the native binding mode (see Figure 3d), but low energy does not imply low RMSD.

The self docking results are comparable to those in Meiler and Baker [MB06]. Meiler and Baker achieved a 71% success rate in a self docking benchmark of 100 small molecules. We see the same success rate on our reduced set despite the increased conformational space sampled for the small molecule. However in the cross docking benchmark our results fall short. One possible cause is the much increased conformational space sampled for small molecules in the present protocol.

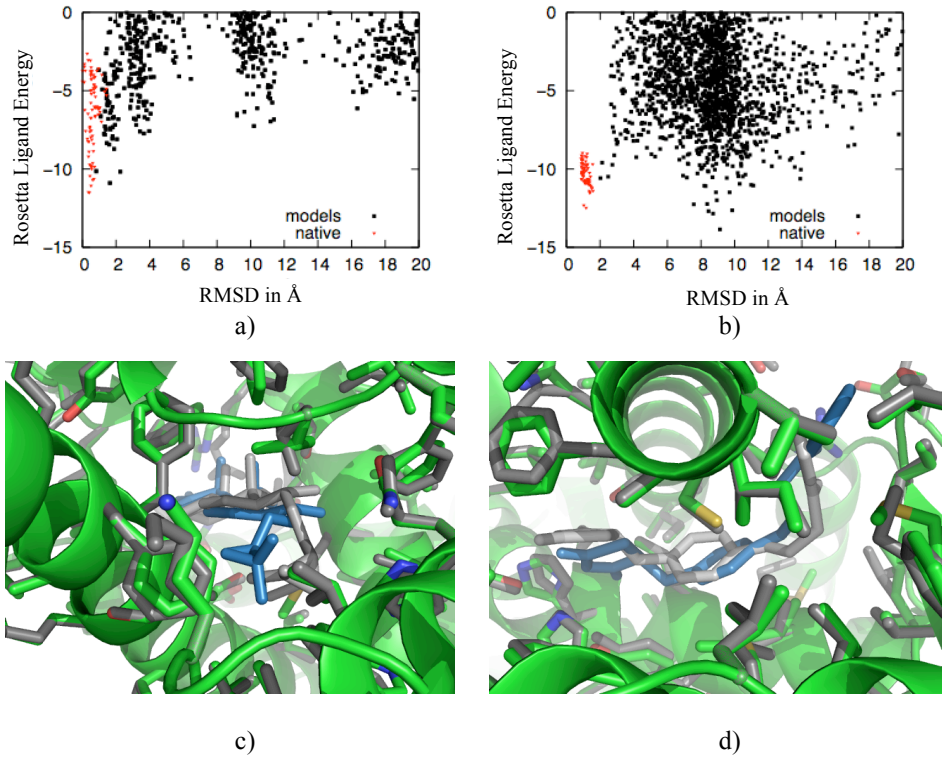

Figure 3. RMSD energy funnels show successful discrimination of binding funnel in a) for the case epoxycholesterol into Liver X receptor from both from 1p8d, and failure to a singular binding in b) for the thiazolidiinedione from 2prg in the 1fm9 structure of the peroxisome proliferator activated receptor. c) shows the best scoring model (in green and blue) overlaid on the on the 2prg structure (shown in grey). d) best scoring model (in green and blue) under 2 Å RMSD from the atomic coordinates of the small molecule crystallized in 2prg (in grey).

The previous evaluation used an ensemble size of only ten in which one conformation was close to the crystallized conformation. Here, we create unbiased ensembles with up to 500 conformations. The increase in conformational diversity represents an increased challenge to the search process as well as the scoring function.

4 Conclusion

We have extended the amino acid concept of rotamers to include small molecules. When the number of torsions is in the same range as those seen in amino acids, small molecule rotamer ensembles contain conformations close to those seen in crystal structures of protein small molecule complexes. In such cases rotamer ensembles can efficiently simulate flexibility for small molecules.

However, as the number of rotamers grow (due to increased flexibility) and the precision of the protein structures decrease (due to inaccuracy in the protein backbone), the discriminatory power of the scoring function decreases. The components of the scoring function have not been optimized for the increased flexibility; doing so may yield increased discrimination. Improved fine grain sampling of protein backbone motion may also assist in the docking process.

Additionally, the method must be extended to larger small molecules. We intend on expanding our method by breaking small molecules into multiple residues. The residues would then be reassembled in the protein binding site to form the small molecule. Thereby, we decrease the conformational complexity and incorporate information from the protein environment.

References

[All02] F. H. Allen. The Cambridge Structural Database: a quarter of a million crystal structures and rising. *Acta Crystallographica B*, 58(Pt 3Pt 1):380–8, 2002.

[BGG03] J.Bostrom,J.R.Greenwood,andJ.Gottfries. Assessing the performance of OMEGA with respect to retrieving bioactive conformations. *Journal Molecular Graphics and Modelling*, 21(5):449-462,2003.

[BMB05] P. Bradley, K. M. Misura, andD. Baker. Toward high-resolution *de novo* structure prediction for small proteins. *Science*, 309(5742):1868–1871,2005.

[DK93] R. L. Dunbrack and M. Karplus. Backbone-Dependent Rotamer Library for Proteins Application to Side-Chain Prediction. *Journal of Molecular Biology*, 230(2):543-574, 1993.

[DKC+03] G. Dantas, B. Kuhlman, D. Callender, M. Wong, and D. Baker. A large scale test of computational protein design: Folding and stability of nine completely redesigned globular proteins. *Journal of Molecular Biology*, 332(2):449–460, 2003.

[GMW+03] J. J. Gray, S. Moughon, C. Wang, O. Schueler-Furman, B. Kuhlman, C. A. Rohl, and D. Baker. Protein-protein docking with simultaneous optimization of rigid-body displacement and side-chain conformations. *Journal of Molecular Biology*, 331(1):281–99, 2003.

[KDI+03] B. Kuhlman, G. Dantas, G. C. Ireton, G. Varani, B. L. Stoddard, and D. Baker. Design of a novel globular protein fold with atomic-level accuracy. *Science*, 302(5649):1364-1368, 2003.

[KKB04] T. Kortemme, D. E. Kim, and D. Baker. Computational alanine scanning of protein-protein interfaces. *Science STKE*, 2004(219):pl2, 2004.

[KM94] G. Klebe and T. Mietzner. A fast and efficient method to generate biologically relevant conformations. *Journal of Computer Aided Molecular Design*, 8(5):583-

606, 1994.

[KOK+01] B. Kuhlman, J. W. O'Neill, D. E. Kim, K. Y. J. Zhang, and D. Baker. Conversion of monomeric protein L to an obligate dimer by computational protein design. *Proceedings of the National Academy of Sciences of the United States of America*, 98(19):10687–10691, 2001.

[LK99] T. Lazaridis, M. Karplus Effective energy function for proteins in solution. *Proteins,* 35(2):133-152, 1999

[Lea94] A. R. Leach. Ligand docking to proteins with discrete side-chain flexibility. *Journal Molecular Biology*, 235(1):345–56, 1994.

[MB06] J. Meiler and D. Baker. ROSETTALIGAND: protein-small molecule docking with full side-chain flexibility. . *Proteins-Structure Function and Bioinformatics*, 65(3):538–48, 2006.

[MMWM02] J. Meiler, W. Maier, M. Will, and R. Meusinger. Using neural networks for ^{13}C NMR chemical shift prediction-comparison with traditional methods. *Journal of Magnetic Resonance*, 157(2):242–52, 2002.

[MK05] A. V. Morozov, T. Kortemme Potential functions for hydrogen bonds in protein structure prediction and design. *Advances in Protein Chemistry*, 72:1-38, 2005

[SFWB05] O. Schueler-Furman, C. Wang, and D. Baker. Progress in protein-protein docking: Atomic resolution predictions in the CAPRI experiment using RosettaDock with an improved treatment of side-chain flexibility. *Proteins-Structure Function and Bioinformatics*, 60(2):187–194, 2005.

[PC04] E. Perola and P. S. Charifson. Conformational analysis of drug-like molecules bound to proteins: an extensive study of ligand reorganization upon binding. *Journal of Medicinal Chemistry*, 47(10):2499-2510, 2004.

[WFLW04] R. Wang, X. Fang, Y. Lu, and S. Wang. The PDBbind database: collection of binding affinities for protein-ligand complexes with known three-dimensional structures. *Journal of Medicinal Chemistry*, 47(12):2977–80, 2004.

[WWC04] Wang, R. M. Wolf, J. W. Caldwell, P. A. Kollman, and D. A. Case. Development and testing of a general amber force field. *Journal of Computational Chemistry*, 25(9):1157–74, 2004.