Chapter 2

THE DARWINDOCK ALGORITHM FOR SMALL MOLECULE DOCKING

Adam R. Griffith, Ismet Caglar Tanrikulu, Ravinder Abrol, William A. Goddard III

Introduction

The first step in understanding how binding of a ligand to a protein affects its function is to determine its binding site, conformation, and binding energy. The ultimate arbiter for the structure is the x-ray structure, and the arbiter for the binding energy uses radioisotopes to measure an equilibrium constant or competitive binding to obtain an IC50. However, such experimental procedures are far too slow for broad searching for new ligand scaffolds (e.g. virtual ligand screening) or for optimizing hits from such a screening. Here we use theory and computation to identify the most likely binding sites and configurations and to rank ligands in terms of binding (or, ideally, in terms of function). This computational process is referred to as "docking". Various methods for docking have been developed over the last 50 years and are widely practiced in industry and academia. A typical application involves a coupling of theory and experiment where a number of putative poses from energy minimization and molecular dynamics computations might be tested by doing mutation experiments to identify which poses are most likely correct, followed by computational modifications of the ligand to improve binding by competitive binding experiments to select the best ligand.

My goal has been to develop a purely computational algorithm that can predict the best ligands and binding sites, without the intervention of experiments. This requires a very

complete sampling of ligand poses throughout the possible binding region and requires an energy scoring that is accurate enough to discriminate the binding strength of various ligands to various binding sites. But carrying out accurate calculations of binding strength generally involves extensive calculations for every possible pose and every possible ligand conformation, which is not practical. DarwinDock is an algorithm that I in collaboration with others in the Goddard group have developed to solve this conundrum.

The DarwinDock method is a new strategy for docking that we refer to as Complete Sampling-Hierarchical Scoring (CS-HS). The idea is:

1. **Alanization** – First, properly prepare the protein system to minimize the chances of bad contacts due to improper sidechain placement and maximize the interaction of the ligand with polar groups in the protein. We achieve this via a process referred to as "alanization", which is the replacing of bulky, nonpolar sidechains with alanine.
2. **Completeness** – Then, a complete set of ligand poses is generated that completely samples the putative binding site, but is done so quickly by eschewing any energy calls. Poses are generated in iterations of 5000 and clustered by RMSD into families. When the number of new families reaches a convergence threshold we consider this to be a complete sampling of the binding site. This typically yields ~50,000 poses.
3. **Scoring** – Finally, the poses are scored in a hierarchical manner in order to minimize the number of energy calls necessary. We score each of the family heads

(as determined by RMSD) using a single energy calculation. Based on these energies we eliminate 90% of the families. The children of the remaining families are then scored and all remaining poses are ranked. From this set we choose the best 120 poses based on the scoring energy and output them for detailed examination

This overall procedure is referred to as DarwinDock. DarwinDock is aimed at being automatic, relying on our scoring algorithms to select interesting ligand candidates without human intervention. This makes DarwinDock useful for virtual ligand screening (VLS) applications where the DarwinDock method might be used to rank the output of a pharmacophore-driven VLS process.

Indeed, as I have been developing and optimizing DarwinDock over the last few years, it has been used in the Goddard group for numerous successful applications.[1-9]

The goal of this chapter is to document and explain the full DarwinDock procedure, in particular how the optimum settings for the procedure are determined. That is, to identify the settings that provide the highest probability of success with minimal computational time.

Evaluating the performance of such an algorithm is difficult since biological systems are complex; therefore, we use sets of pre-determined systems for validation. Specifically, we use the Directory of Useful Decoys[10] (DUD set), which is a selection of 40 diverse systems

based on high quality x-ray structures of ligand/protein co-crystals and is intended for validation of docking programs and methods.

DarwinDock

The first stage of DarwinDock – **alanization** – addresses a critical problem in docking a ligand to a protein. The optimum binding site and conformation of the ligand depend on the conformations of the protein sidechains; simultaneously, the conformations of the protein sidechains depend on the binding site and conformation of the ligand. Thus we need to dock the ligand simultaneously with optimized protein sidechain conformations. We have solved this "chicken-egg" problem by replacing the bulky, nonpolar sidechains with alanine prior to docking, which we refer to as "alanization". We consider valine, leucine, isoleucine, methionine, phenylalanine, tyrosine, and tryptophan to be primarily nonpolar in character and thus less likely to be essential to orienting a ligand in a binding site. Furthermore, these residues are bulky enough that they might block significant portions of the binding site if placed incorrectly, eliminating what might be the ideal binding pose. Alanizing these residues additionally opens up the binding site, allowing the ligand to sample a larger space in the binding site without being bumped so that it has the best chance of interacting more directly with polar sidechains. The tradeoffs are that we miss a significant part of van der Waals (VDW) interactions with the ligand, and we must do greater sampling to span the more open binding site. While not a required part of preparing a system for DarwinDock, it is recommended and we consider this as the default approach.

The second stage of DarwinDock addresses **complete sampling** of the binding. Ligand poses are generated in iterations of 5000 using DOCK 6[11] and are clustered by RMSD into families with a diversity of 2Å. When the number of new families falls below a threshold of 2% we consider the set of poses to be a complete sampling of the putative binding site. This procedure typically results in ~50,000 poses and ~5,000 families. The only evaluation performed on poses during the "completeness" stage of DarwinDock is a bump test to ensure that the ligand does not clash with the protein. *No energy calls* are made during this stage.

The number of poses generated in the completeness stage is quite large, making evaluating accurate energies for each pose impractical. Instead we use a hierarchical approach for scoring the poses in the final, **scoring** stage of DarwinDock, beginning with scoring of the family heads. The head of each 2Å family from the completeness stage is selected as the centroid of the family based on the heavy-atom RMSD and then its energy is evaluated. We use the DREIDING[12] forcefield in MPSim[13] to evaluate the non-bond energy between the ligand and the rest of the protein. Based on the energy of each family head, we eliminate 90% of the families and focus on the children of the remaining 10%. Assuming that the family head is broadly representative of the children, this allows us to dramatically reduce the number of energy evaluations necessary to finally select the best pose. At the end of the scoring stage we obtain 120 best poses for further, more accurate evaluations.

We the use closest-neighbor seeded (CNS) algorithm to cluster poses based on heavy-atom RMSD. The pairwise RMSDs for all ligand poses are calculated and the pairs are ordered

by increasing RMSD. Each pose is initially placed in its own family. The list of RMSDs is then traversed for each *i, j* pair of poses. If the RMSD of each member of family *i* to each member of family *j* is less than the cutoff value (2Å), then the families are merged. This builds up clusters of poses starting from the most closely related poses. The centroid of the family is then labeled as the family head.

It should be noted that our choices of DOCK for pose generation, the CNS algorithm for clustering, and DREIDING force field for energy evaluation are made for convenience in our implementation. Any other methods of pose generation, clustering, and scoring could be used.

Many of the parameters in DarwinDock can be adjusted in order to better suit a particular project. For instance, the default completeness threshold of 2% can be increased to produce a faster, less complete calculation. The clustering diversity can be increased from 2Å so that the family head is more representative of the children. The percent of families scored can be increased above 10% for a slower, but more thorough consideration of possible poses. The polar and nonpolar components of the scoring energy can be adjusted based on the composition of the ligand. Changing forcefield parameters such as the dielectric constant can also alter scoring. The DUD set provides a straightforward test for identifying the default settings for DarwinDock. Specifically, we will derive the defaults for:

- Completeness threshold (2%)
- Clustering diversity (2Å)

- Percent of families fully scored (10%)
- Polar (100%) and nonpolar (10%) scaling for scoring
- Dielectric constant (2.5, distance dependent)

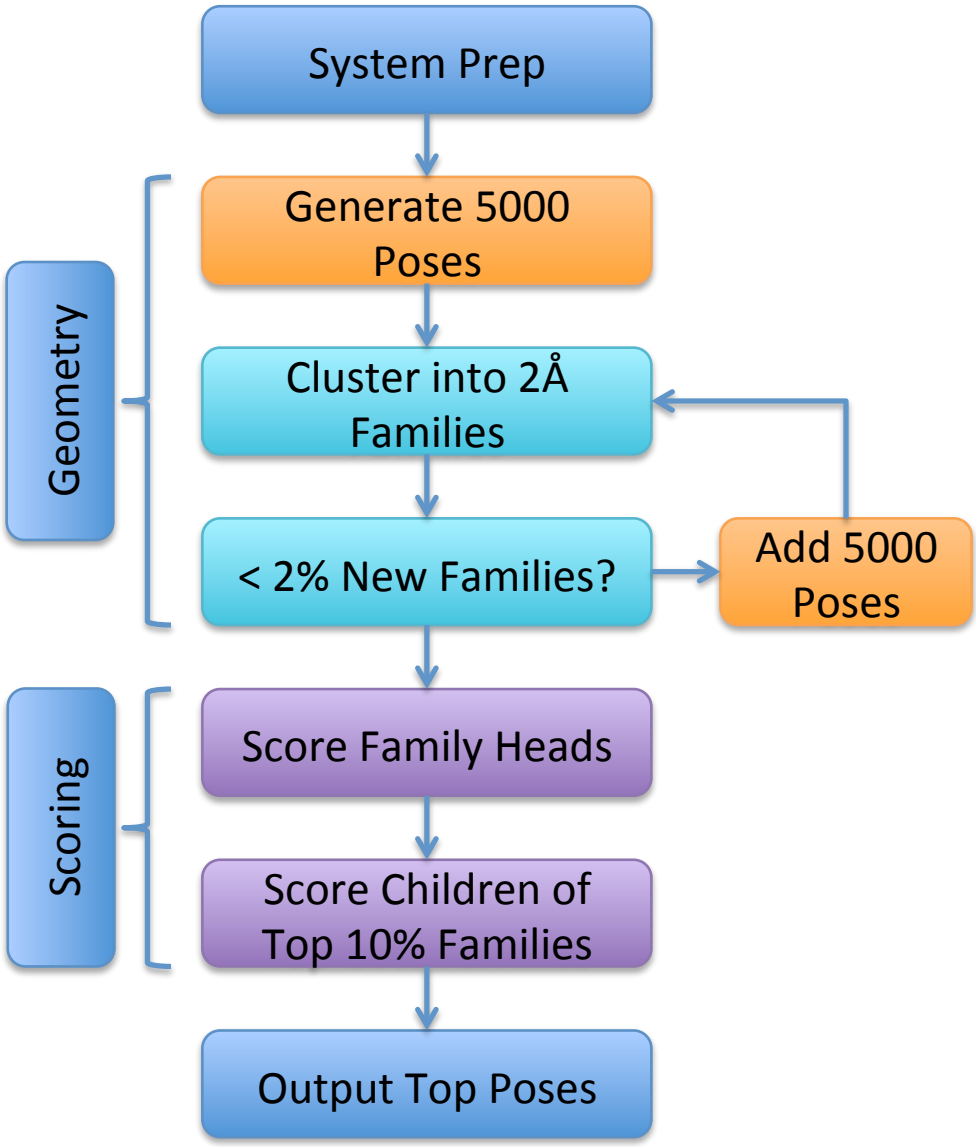

Figure 2-1 - Diagram of the DarwinDock algorithm. Pose generation and scoring are partitioned into two completely separate stages. The geometry or "completeness" stage generates 5000 poses, clusters them into 2Å families, and adds 5000 additional poses until the completeness threshold has been reached. The scoring stage initially only evaluates the 2Å family heads. 90% of the families are eliminated based on the family head energy, and the children of the remaining 10% are scored fully. Typically, 120 interesting candidate poses are output from this final list.

The DUD Set

The DUD set contains 40 systems for which accurate x-ray structures are available for a co-crystal of a small molecule ligand bound to a protein. It was intended to provide a reasonable test for docking programs and methods so that their accuracy can be assessed. Of these 40 systems we rejected three of the systems as inappropriate for small-molecule docking validation.

1. The version of the system 'thrombin' (pdb: 1ba8) included in the DUD set contains a covalent ligand, which represents a wholly different class of ligands than standard small-molecules.
2. The system 'pdgfrb' in the DUD set is in fact derived from a computational model and not an experimental structure; therefore, it cannot be used to provide accurate validation of another docking model.
3. The ligand in 'cdk2' is not completely resolved, with several missing heavy atoms.
4. Additionally, while we do not reject the 'comt' system, it should be noted that it contains two copies of the target ligand within the binding site. The presence of the second ligand is only obvious when one includes neighboring unit cells from the crystal in the structure, which may explain why this system was included in the DUD set. The positioning of the two ligands is shown in Figure 2.

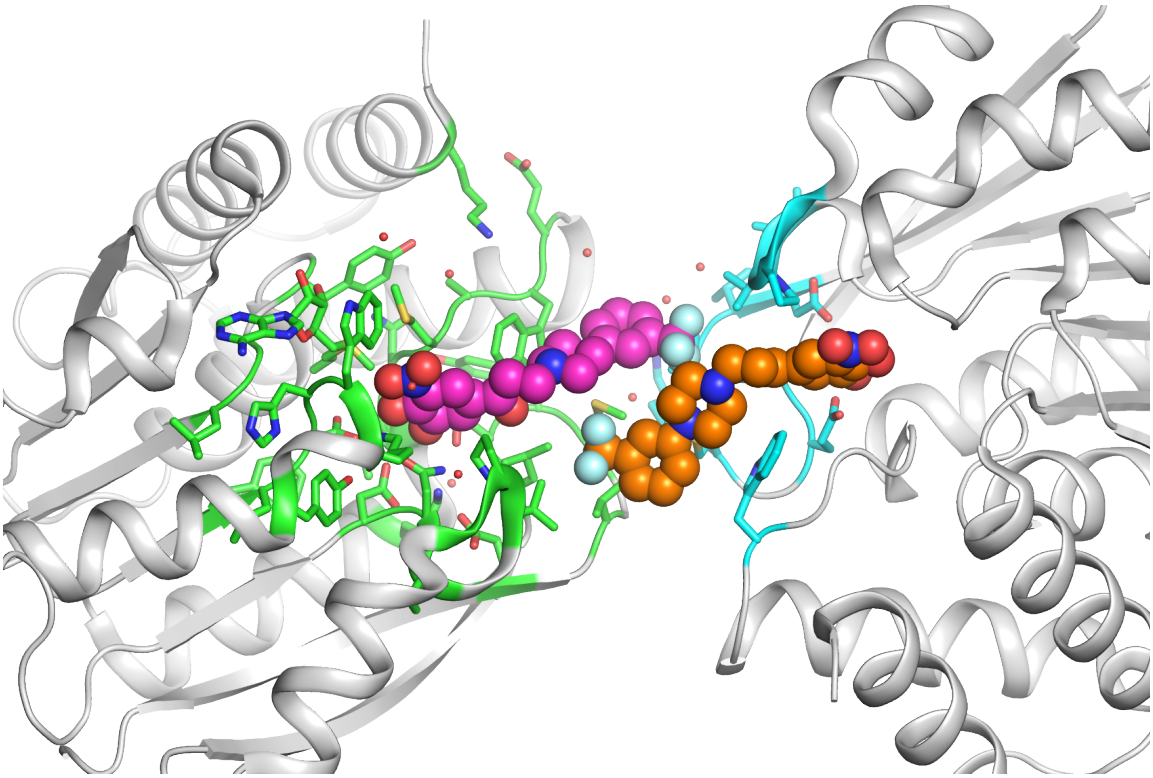

Figure 2-2 - The system 'comt' contains two copies of the target ligand within the binding site, shown as spheres in magenta and orange. This oddity is only obvious when one includes neighboring unit cells from the crystal in the structure.

Table 2-1 - DUD Systems used in validation with corresponding PDB ID number, the residue ID information of the ligand, and general system information. Three of the forty systems have been rejected as inappropriate for inclusion in benchmarking, with reasons listed.

Name	PDB ID	Lig ID	Lig Num	Lig Chn	System Name	System Class
ace	1o86	LPR	702	A	Angiotensin-converting enzyme	Metalloenzyme
ache	1eve	E20	2001	A	Acetylcholine esterase	Other enzyme
ada	1ndw	FR2	1001	A	Adenosine deaminase	Metalloenzyme
alr2	1ah3	TOL	320	A	Aldose reductase	Other enzyme
ampc	1xgj	HTC	777	A	AmpC beta lactamase	Other enzyme
ar	2ao6	R18	1001	A	Androgen receptor	Nuclear hormone receptor
comt	1h1d	BIA	335	A	Catechol O-methyltransferase	Metalloenzyme
cox1	1q4g	BFL	701	A	Cyclooxygenase 1	Other enzyme
cox2	1cx2	S58	701	A	Cyclooxygenase 2	Other enzyme
dhfr	3dfr	MTX	164	A	Dihydrofolate reductase	Folate enzyme
egfr	1m17	AQ4	999	A	Epidernam growth factor receptor kinase	Kinase
er_ag	1l2i	ETC	600	A	Estrogen receptor agonist	Nuclear hormone receptor
er_ant	3ert	OHT	600	A	Estrogen receptor antagonist	Nuclear hormone receptor
fgfr1	1agw	SU2	1001	A	Fibroblast growth factor receptor kinase	Kinase
fxa	1f0r	815	401	A	Factor Xa	Serine protease
gart	1c2t	NHS	222	A	Glycinamide ribonucleotide transformylase	Folate enzyme
gpb	1a8i	GLS	998	A	Glycogen phosphorylase beta	Other enzyme
gr	1m2z	DEX	301	A	Glucocorticoid receptor	Nuclear hormone receptor
hivpr	1hpx	KNI	900	B	HIV protease	Other enzyme
hivrt	1rt1	MKC	999	A	HIV reverse transcriptase	Other enzyme
hmga	1hw8	114	3	D	Hydroxymethylglutaryl-CoA reductase	Other enzyme
hsp90	1uy6	PU3	1224	A	Human heat shock protein 90 kinase	Kinase
inha	1p44	GEQ	350	A	Enoyl ACP reductase	Other enzyme
mr	2aa2	AS4	201	A	Mineralcorticoid receptor	Nuclear hormone receptor
na	1a4g	ZMR	466	A	Neuraminidase	Other enzyme
p38	1kv2	B96	391	A	P38 mitogen activated protein kinase	Kinase
parp	1efy	BZC	201	A	Poly(ADP-ribose) polymerase	Other enzyme
pde5	1xp0	VDN	201	A	Phosphodiesterase V	Metalloenzyme
pnp	1b8o	IMH	600	A	Purine nucleoside phosphorylase	Other enzyme
ppar	1fm9	570	200	D	Peroxisome proliferator activated receptor gamma	Nuclear hormone receptor
pr	1sr7	MOF	302	B	Progesterone receptor	Nuclear hormone receptor
rxr	1mvc	BM6	200	A	Retinoic X receptor alpha	Nuclear hormone receptor
sahh	1a7a	ADC	435	A	S-adenosyl-homocysteine hydrolase	Other enzyme
src	1y57	MPZ	600	A	Tyroside kinase SRC	Kinase
tk	1kim	THM	2	B	Thymidine kinase	Kinase
trypsin	1bju	GP6	910	A	Trypsin	Serine protease
vegfr2	1fgi	SU1	1001	A	Vascular endothelial growth factor receptor kinase	Kinase

Rejected					
thrombin	1ba8	0IT	1	B	covalent ligand
pdgfrb					computational model
cdk2	1ckp	PVB	299	A	incompletely resolved ligand

System Preparation

Despite being part of a curated set, the DUD systems required careful preparation before being used in validating our method. While DUD provides pre-prepared files for each system, we found it useful to return to the original PDB source. In particular this allowed

us to use the full protein structures and symmetry information in order to generate the final systems used for docking. The general procedure is as follows:

1. Generate neighboring unit cells using the symmetry information
2. Remove parts of the system that are distant from the target ligand (8Å cutoff)
3. Add hydrogens
4. Optimize ligand protonation states
5. Optimize asparagine, glutamine, and histidine flips, as well as histidine protonation
6. Minimize ligand separately and assign partial charges
7. Assign forcefield types
8. Perform conjugate-gradient energy minimization on the system
9. Alanize bulky residues (V, L, I, M, F, Y, W)
10. Generate sphere points for use with DOCK

Steps 1 and 2 were performed using PyMol[14]. Steps 3-5 were performed using the Maestro Protein Preparation Wizard[15-17]. CHARMM[18] charges were used for protein atoms. Ligands were minimized using the Maestro OPLS forcefield minimization[19] before generating Mulliken charges. Single-point energy calculations were performed using Jaguar[20] and the B3LYP level of DFT with the 6-311G** basis set except for the ligand containing bromine, where the ERMLER**++ basis set was used. Conjugate gradient energy minimization of the final system was performed using MPSim. Sphere generation was performed using the standard DOCK *sphgen* parameters and methods, with the exception that the maximum sphere radius is set to be 12Å instead of 4Å. This allows the

spheres to span larger binding sites without gaps or voids in the binding site. All spheres within 5Å of the ligand were selected, and were then clustered using the "cns" algorithm to reduce the number of spheres to below 400 spheres. This is necessary for the DOCK calculation to fit within available memory.

Systems prepared thus represent the final pre-docking crystal structures; however, there are other considerations that must be made. In a real-world use of DarwinDock there would be no pre-existing crystal structure containing both protein and ligand. At best there would be an apo crystal structure or a structure containing a different ligand. Thus we will not know the conformations of protein sidechains in a real-world application of any docking method. Therefore, in addition to testing DarwinDock against structures with crystal sidechains, our most important tests are for systems in which the sidechain conformations are predicted. Here we use the SCREAM[21] method to predict the sidechain conformations of the apo-protein. This allows us to test how well DarwinDock would do in a real ligand discovery project. Some sidechains were kept fixed during the predictions due to obvious strong interactions with ions or non-target ligands in the protein. These are listed in Table 2. Alanization of the bulky, nonpolar residues was applied after sidechain placement with SCREAM. It should be noted that while many of the structures showed waters present in the binding sites, we removed all waters prior to any calculations. In a real-world test the placement of waters in the binding site would not be known before docking. Coordinated ions and other ligands, however, might be known or inferred from related structures; therefore, these were left in place.

We tested several combinations of parameters in SCREAM in order to identify the best way to predict sidechains without the presence of the ligand. Specifically, we considered flat dielectric constants of 2.5, 3.33, and 5.0, as well as distance-dependent dielectric constants of 1.0, 2.5, 3.33, and 5.0.

The presence of histidines within the binding sites of some systems required special consideration. No system had more than two histidines within 4Å of the ligand, excluding histidines that were fixed due to interactions with ions. Therefore each possible combination of neutral and positively charged histidine (denoted as "B" instead of "H" in our terminology) was attempted. We also tested an additional combination where all flexible histidines in the binding site were replaced with alanine. The histidines treated in this way are listed in Table 2. Using this approach up to five different sidechain predictions were made for each dielectric constant.

Table 2-2 - Residues in the binding site that are fixed due to interactions with ions or non-target ligands, and histidines in the binding site that are tested as both neutral and charged.

System	Fixed Residues	Histidines
ace	H383_A H387_A E411_A	H353_A H513_A
ache		H440_A
ada	H15_A H17_A H214_A H238_A D295_A	H157_A
alr2		H110_A
comt	D141_A D169_A	H142_A
cox2		H90_A
dhfr		H28_A
er_ag		H524_A
er_ant		H524_A
fxa		H57_A
gart		H108_A H137_A
gpb		H377_A
hivrt		H235_A
p38		H148_A
parp		H862_A
pde5	H617_A H653_A D654_A D764_A	H613_A
pnp	S33_A H64_A R84_A H86_A S220_A	H257_A
ppar		H323_D H449_D
rxr		H435_A
sahh		H55_A H353_A
src	R388_A	
tk		H58_B
trypsin	H57_A	

Predicted Sidechain Sets

We assessed predicted sidechains for each system using the dielectric constants and histidine considerations mentioned above, both with and without alanization of the nonpolar residues. The calculations were performed in the absence of the ligand, but the

ligand was replaced in order to evaluate the number of heavy atoms with close contacts to the ligand. Based on these calculations we identified three sets of sidechains to dock to.

The first set is referred to as the "best case". We identified the best sidechain predictions for each of the 37 systems using the number of close contacts and the sidechain RMSD for each combination of dielectric constant and histidine treatment. The systems were alanized, therefore only polar sidechains were used in the bump and RMSD analysis. These predictions represent a "best case scenario" for predicted sidechains and are a reasonable set to use for identifying the optimum default settings for DarwinDock. However, this set doesn't represent a true real-world test because information about the ligand and sidechains are used to identify which prediction method to use for each system. Table 3 shows the analysis of bumps for different settings. It is obvious from the analysis of the bumps that there are clear cases where alanization of bulky residues dramatically decreases the number of bumps with the ligand.

For real-world testing of DarwinDock we used two additional sets of sidechain predictions. The first used a constant dielectric of 2.5. While not the best performer in terms of bumps, it represents the default settings that have been used in previous applications of DarwinDock. The set with the fewest average number of bumps used a distance-dependent dielectric of 2.5. Both sets were tested with and without alanization.

For reference only we also tested DarwinDock against the systems using crystal sidechains with and without alanization.

Table 2-3 - This table shows the number of bumps with the ligand for each type of sidechain prediction using different dielectrics and with or without alanization. The left half of the table shows results with alanization, the right without alanization. The table has been color-coded with large numbers of bumps shown in red. It is clear when comparing the left (alanized) and right (not alanized) portions of the table that alanization is key to reducing the number of bumps with the ligand. Even small numbers of bumps can make it impossible for the ligand to be placed correctly during docking.

	alanized									all sidechains								
	best	worst	2.50/flat	3.33/flat	5.00/flat	1.00/dist	2.50/dist	3.33/dist	5.00/dist	best	worst	2.50/flat	3.33/flat	5.00/flat	1.00/dist	2.50/dist	3.33/dist	5.00/dist
average			1.57	1.14	1.16	1.78	1.11	1.54	1.59			6.62	5.22	5.62	6.19	5.03	4.70	5.49
worst			26	10	10	14	10	22	29			35	21	35	19	20	24	29
ace	0	1	0	0	0	0	1	1	1	0	1	0	0	0	0	1	1	1
ache	0	0	0	0	0	0	0	0	0	7	12	7	8	7	8	11	12	12
ada	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
alr2	0	0	0	0	0	0	0	0	0	0	13	0	0	0	13	2	2	2
ampc	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ar	2	6	2	2	2	2	2	6	2	10	23	23	17	22	16	18	10	16
comt	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
cox1	0	0	0	0	0	0	0	0	0	0	1	1	1	0	1	1	1	1
cox2	0	10	2	10	8	4	0	0	0	0	10	2	10	8	4	0	0	0
dhfr	0	6	0	0	0	3	6	6	6	6	20	14	15	14	17	20	6	6
egfr	0	0	0	0	0	0	0	0	0	0	5	0	0	0	0	2	2	5
er_ag	0	0	0	0	0	0	0	0	0	8	11	9	8	8	8	11	8	8
er_ant	0	0	0	0	0	0	0	0	0	0	7	0	1	4	1	6	6	7
fgfr1	0	0	0	0	0	0	0	0	0	0	6	0	4	0	6	0	0	6
fxa	0	0	0	0	0	0	0	0	0	3	7	7	3	3	3	7	7	3
gart	0	4	1	0	2	1	1	4	0	0	4	1	0	2	1	1	4	0
gpb	2	10	3	3	10	10	10	2	2	2	13	10	6	13	10	10	2	2
gr	1	8	2	2	2	8	1	1	1	12	24	24	21	21	19	12	12	12
hivpr	0	8	0	0	0	8	1	1	5	0	9	2	2	0	9	2	1	6
hivrt	0	0	0	0	0	0	0	0	0	2	17	14	14	3	17	3	2	3
hmga	0	2	2	2	1	0	0	1	1	2	7	7	3	2	5	5	6	6
hsp90	0	0	0	0	0	0	0	0	0	2	13	6	6	6	2	2	2	13
inha	0	0	0	0	0	0	0	0	0	12	35	35	20	35	12	20	20	20
mr	0	1	0	0	1	0	0	0	0	1	9	3	3	1	9	2	2	3
na	2	29	26	6	10	2	5	22	29	3	29	27	6	11	3	7	24	29
p38	1	14	10	5	1	14	4	2	2	4	17	11	8	4	17	5	5	5
parp	0	1	0	0	0	1	0	0	0	0	12	0	2	2	12	0	0	0
pde5	0	0	0	0	0	0	0	0	0	0	15	15	7	4	4	1	2	0
pnp	0	1	1	1	1	0	1	1	0	1	4	1	4	4	3	4	3	3
ppar	0	2	0	0	0	0	0	0	2	0	7	2	0	2	2	3	7	2
pr	0	0	0	0	0	0	0	0	0	0	3	3	2	3	0	1	0	0
rxr	0	3	3	3	0	1	0	0	0	2	10	10	4	7	2	8	8	8
sahh	0	6	3	2	0	6	3	4	3	0	6	3	2	0	6	5	4	3
src	0	1	1	1	0	0	0	0	0	0	1	1	1	0	0	0	0	0
tk	0	4	0	3	3	4	4	4	4	0	15	0	8	15	12	9	9	15
trypsin	0	1	1	1	1	1	1	1	0	0	1	1	1	1	1	1	1	0
vegfr2	0	0	0	0	0	0	0	0	0	4	5	5	5	5	5	5	4	5

"Best Case" Predicted Sidechain Results

This set of sidechain predictions represents a "best case scenario" for predicted sidechains. Such a set allowed us to identify the best default settings for the percent of families scored, the composition of the scoring energy, the clustering diversity, and the completeness threshold. For the percent of families scored we tested 10, 25, 33, 50, and 100%. As with the sidechain predictions, we assessed both a flat dielectric of 2.5 and a distance-dependent dielectric of 2.5. We also tried various scalings of the polar (Coulomb and hydrogen bond) component and the nonpolar (van der Waals) component of the scoring energy. The

scalings tested were 100% polar with 0-100% phobic energy in 10% increments, and 100% phobic with 0-100% polar energy in 10% increments. In order to assess the accuracy of the docking calculations we consider how many of the final 120 poses are within a 2.0Å heavy-atom RMSD of the crystal ligand. Through experience we have found that a pose within 2Å of the crystal ligand is sufficient to identify the pharmacophore. It represents both the orientation of the ligand in the binding site as well as the key interactions with the protein. While it is of course ideal to have as small an RMSD as possible, a 2Å RMSD is sufficient to have predictive value and acknowledges that the crystal structure is only a snapshot of what is really a dynamical system..

Figure 3 summarizes the average number of 2Å hits across all 37 DUD systems docked to, out of a possible 120 hits. There was a clear preference for the scoring energy using 100% polar (Coulomb + hydrogen bonding) and 10% phobic (van der Waals), regardless of the dielectric constant used or the percent of families scored, therefore this was set as the default for remaining calculations.

Figure 4 examines the impact of the percent of families scored and the dielectric constant when the 100% polar / 10% phobic scoring energy is used. The average number of 2Å hits was slightly higher for the distance-dependent 2.5 dielectric than for the constant 2.5 dielectric. As one would expect, increasing the percent of families scored increased the average number of hits; however, the impact was not significant. Increasing the percent of families scored from 10% to 100% yielded a little more than 1 extra hit on average while dramatically increasing the computational cost and calculation time. As 25, 33, and 50%

yielded even smaller increases, there is no reason to set the default percent families scored above 10%. Therefore, the default parameters for the remaining calculations were set at:

- Scoring energy: 100% polar, 10% phobic
- Distance-dependent dielectric, 2.5
- 10% families scored

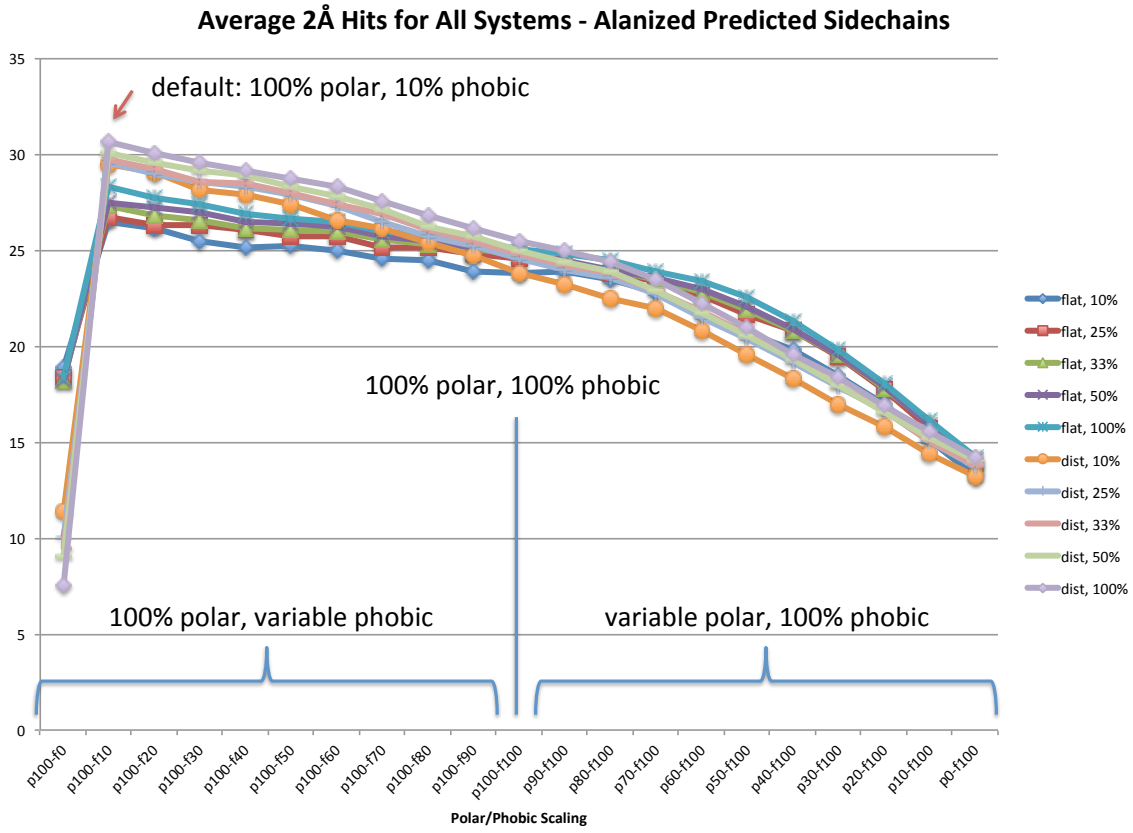

Figure 2-3 - Analysis of docking to the "best" sidechain predictions. The series show either flat or distance-dependent dielectric constants for varying percentages of families scored. The left half of the chart shows scoring energies with 100% polar and the phobic energy scaled from 0 to 90%. The right half of the chart shows scoring energies with 100% phobic energy and the polar energy scaled from 90 to 0%. The center point has 100% polar and 100% phobic scaling. There is a clear preference for 100% polar and 10% phobic regardless of dielectric and percent of families scored.

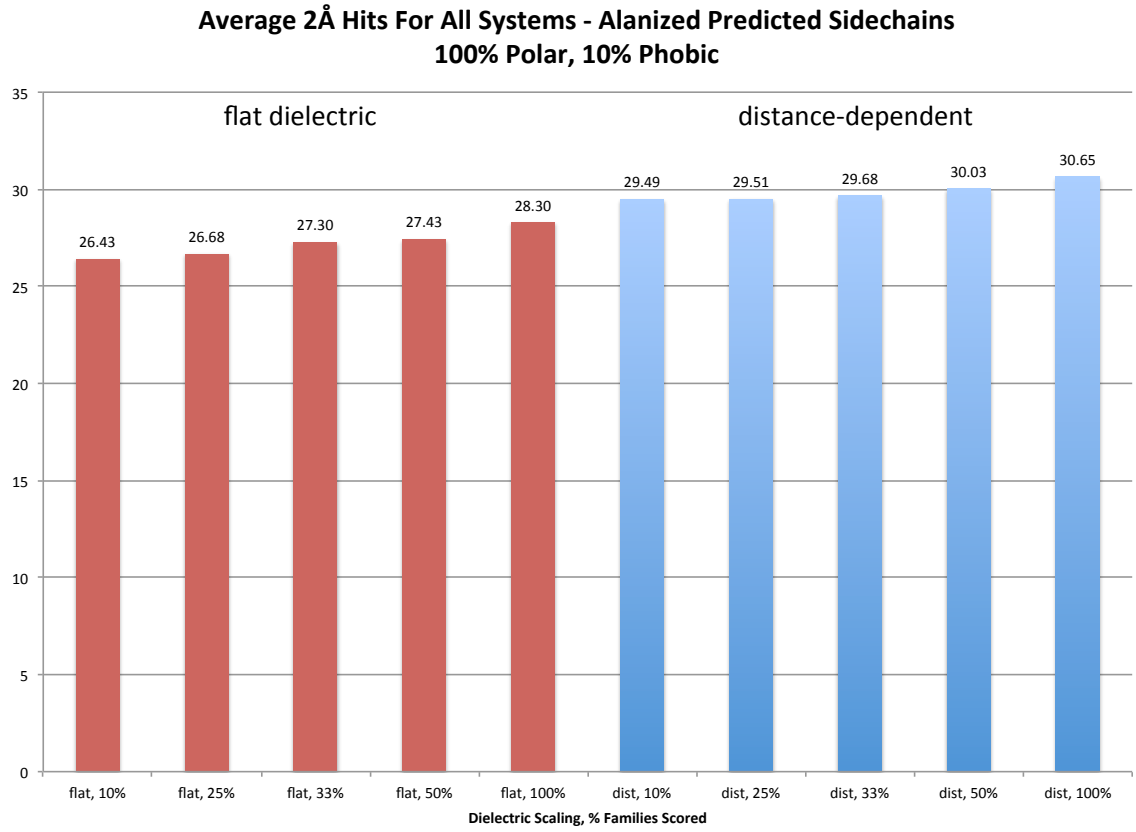

Figure 2-4 - Analysis of docking to the "best" sidechain predictions with the scoring energy set to 100% polar, 10% phobic. The red columns show a flat dielectric of 2.5 with the percent of families scored set to 10, 25, 33, 50, and 100%. The blue columns show the distance-dependent dielectric for the same percent of families scored. There is a clear, modest preference for the distance-dependent dielectric. There is also a slight, but insignificant improvement from increasing the percent of families scored. Due to the added computational cost, there is no reason to score more than 10% of the families for such insignificant improvement.

Using these parameters we assessed different possible completeness thresholds and clustering diversities. In addition to the accuracy of the calculations, these parameters can have a significant impact on the computational cost of the calculations. Using the default clustering diversity of 2Å, we tested completeness thresholds of 1, 2, 5, 10, and 25%. As expected, the accuracy of the calculations increases with the thoroughness of the calculations. That is, a higher completeness threshold represents a more complete sampling of the binding site and is more likely to produce correct poses. Conversely, a

looser completeness threshold is more likely to miss correct poses. The 1% completeness threshold produced a higher number of average hits (33.4) compared to the 2% completeness threshold (29.5), but had a much higher computational cost. The maximum number of poses used by a system increased from 60,000 to 100,000 poses and the average number of poses increased from 39,000 to 60,000. The number of systems that failed to produce any 2Å hits remained the same. Decreasing the threshold to 5% significantly reduced the computational cost, but the number of failed systems increased from 3 to 4. Therefore we feel that a completeness threshold of 2% represents a reasonable balance between accuracy and computational cost.

Using the 2% completeness threshold we assessed clustering diversities of 1, 2, and 3Å. The clustering diversity determines how closely a family head resembles the rest of the family members, and thus how accurate our assumption is that the energy of the family head can be used to eliminate the family members from further consideration. Clustering at 1Å yielded ~6.5 additional hits than the 2Å clustering and one fewer completely failed systems. However, the computational cost of clustering at 1Å was dramatically higher than that at 2Å. 1Å clustering had a maximum number of poses of 125,000 and an average of 86,000, while 2Å clustering had a maximum of 60,000 and an average of 39,000. Increasing the clustering to 3Å reduced the average number of hits by ~3 and decreased the maximum and average number of poses to 40,000 and ~27,000, respectively. However, despite the reduced number of poses, clustering at 3Å actually increased the number of energy calls. This is because the average number of children per family increased from 5.6 to 10. Our default of clustering at 2Å represents an optimum, given the other settings used.

Table 4 summarizes the results for the completeness threshold and clustering diversity tests.

Table 2-4 – Summary of results for various completeness thresholds and clustering diversities. Smaller completeness thresholds and clustering diversities yield improved results, but at significantly increased computational cost. The default settings are highlighted in green.

	Clustering Diversity: 2Å						
Threshold	Avg Hits	# Fails	% Fails	Max Poses	Avg Poses	Avg Fams	Poses/Fam
1%	33.4	3	8.1	100000	60214.3	9124.3	6.6
2%	29.5	3	8.1	60000	39082.1	6986.2	5.6
5%	25.5	4	10.8	35000	23984.4	5052.2	4.7
10%	22.8	4	10.8	20000	16416.8	3879.8	4.2
25%	18.9	4	10.8	15000	11011.4	2941.9	3.7

	Completeness Threshold: 2%						
Diversity	Avg Hits	# Fails	% Fails	Max Poses	Avg Poses	Avg Fams	Poses/Fam
1 Å	36.1	2	5.4	125000	86025.1	30701.3	2.8
2 Å	29.5	3	8.1	60000	39082.1	6986.2	5.6
3 Å	26.4	3	8.1	40000	26784.8	2675.4	10.0

% Scored	10% of Families
Scoring Energy	100% Polar / 10% Phobic
Dielectric	2.5, Distance-Dependent

Standard Predicted Sidechains

Based on the results of the "best case" predicted sidechains, we performed tests on the two sets of standard sidechain predictions. These sets used either a flat or distance-dependent dielectric of 2.5 for the sidechain predictions. The scoring energy for the docking calculations used the settings identified above: 100% polar, 10% phobic, distance-dependent dielectric of 2.5. Additionally, these sets included all of the various histidine

combinations mentioned above. These calculations represent a more authentic real-world approach to a docking problem. The "best" predicted sidechains relied on knowledge of the crystal structure in order to identify which set to use. The "standard" sidechain predictions only made use of the protein backbone from the crystal, the ligand position to identify the general binding site for sphere generation, and the crystal ligand conformation, thus making it much closer to a real-world scenario.

These docking sets produced interesting results. First, on average the systems with sidechains predicted using the distance-dependent dielectric outperformed those using the constant dielectric by about 8 hits on average, despite not having a large difference in the bindsite sidechain RMSD. The number of bumps slightly favors the distance-dependent dielectric sidechains on average. The true surprise is from the non-alanized results. The non-alanized, flat dielectric sidechains outperformed the alanized sidechains by nearly 15 hits on average. The non-alanized, distance-dependent dielectric sidechains *underperformed* the alanized sidechains by about 20 hits on average. This result is puzzling. However, the average number of hits per system is not the only important criteria. It is also important to identify the number of systems that produced no 2Å hits, which represents a complete failure of docking. Flat dielectric with alanization had the fewest systems with no hits at 4 systems, followed by the distance-dependent dielectric with alanization at 7 systems. Both non-alanized sets had 18 of 37 systems with zero 2Å hits.

As with the crystal sidechains, the loss of the van der Waals energy contribution to the binding energy and the increase in the volume to sample can clearly have a detrimental impact on some systems. However, unless one can be *supremely* confident in their sidechain predictions, alanizing the binding site is the most reliable way to ensure that some number of good poses will be produced. There are some systems where alanization is essential (e.g. "ppar"), some where it is detrimental (e.g. "pnp"), and some where it doesn't make an impact (e.g. "trypsin"). Similarly, some systems do better with the flat dielectric (e.g. "dhfr") and some do better with the distance-dependent dielectric (e.g. "rxr"). The ideal approach when working on an individual system is likely to dock to several sets of diverse sidechain predictions in order to cover multiple possibilities.

Table 2-5 – Comparison of docking results for flat-dielectric and distance-dependent dielectric predicted sidechains with and without alanization. The average number of 2Å hits per system looks encouraging for the flat-dielectric set without alanization, but *half* of the systems produce no 2Å hits.

	Avg 2Å Hits				# Sys With 0 2Å Hits			
set	flat, ala	flat, full	dist, ala	dist, full	flat, ala	flat, full	dist, ala	dist, full
33%	45.8	64.9	53.4	36.3	3	17	6	16
25%	45.8	63.5	53.2	36.4	3	17	6	17
10%	45.7	59.4	53.3	33.6	4	18	7	18

Table 2-6 – Average number of 2Å hits for flat dielectric and distance-dependent dielectric predicted sidechains, with and without alanization. There are some systems where alanization is essential (e.g. "ppar"), some where it is detrimental (e.g. "pnp"), and some where it doesn't make an impact (e.g. "trypsin"). Similarly, some systems do better with the flat dielectric (e.g. "dhfr") and some do better with the distance-dependent dielectric (e.g. "rxr"). Note: "A" refers to a histidine replaced with alanine, "B" refers to a neutral histidine changed to protonated histidine.

system	flat, ala	flat, full	dist, ala	dist, full
ace.A353_A_A513_A	31	100	42	13
ace.B353_A	74	30	15	11
ace.B353_A_B513_A	106	0	17	16
ace.B513_A	115	69	61	73
ace	106	109	36	100
ache.A440_A	1	0	0	0
ache.B440_A	0	0	0	0
ache	0	0	0	0
ada.A157_A	10	43	5	30
ada.B157_A	10	31	5	23
ada	5	50	9	11
alr2.A110_A	0	0	0	0
alr2.B110_A	7	0	0	0
alr2	7	0	0	0
ampc	1	0	0	0
ar	0	0	0	0
comt.A142_A	18	49	6	3
comt.B142_A	9	28	7	5
comt	9	28	7	5
cox1	30	120	36	120
cox2.A90_A	12	116	0	0
cox2.B90_A	2	0	9	0
cox2	3	90	2	63
dhfr.A28_A	43	0	59	22
dhfr.B28_A	51	0	4	0
dhfr	51	0	4	0
egfr	14	42	3	0
er_ag.A524_A	5	73	5	119
er_ag.B524_A	0	0	1	120
er_ag	7	77	9	119
er_ant.A524_A	2	115	10	0
er_ant.B524_A	2	96	10	0
er_ant	2	96	10	0
fgfr1	5	34	4	15
fxa.A57_A	0	0	49	0
fxa.B57_A	14	0	58	0
fxa	14	0	58	0
gart.A108_A_A137_A	0	19	0	0
gart.B108_A	0	1	0	0
gart.B108_A_B137_A	0	0	0	0
gart.B137_A	0	0	0	0
gart	0	1	0	0
gpb.A377_A	0	0	2	13
gpb.B377_A	2	0	0	0
gpb	0	0	0	0

system	flat, ala	flat, full	dist, ala	dist, full
gr	8	0	45	0
hivpr	59	0	27	6
hivrt.A235_A	17	0	27	0
hivrt.B235_A	20	0	25	0
hivrt	18	0	26	0
hmga.A752_C	1	1	1	2
hmga.B752_C	3	2	8	3
hmga	0	0	12	6
hsp90	7	3	9	0
inha	51	0	59	0
mr	14	120	33	120
na	0	0	0	0
p38.A148_A	55	0	117	0
p38.B148_A	0	0	40	0
p38	44	0	10	0
parp.A862_A	5	0	2	0
parp.B862_A	13	1	11	1
parp	12	0	4	1
pde5.A613_A	0	0	0	11
pde5.B613_A	0	0	0	0
pde5	0	0	0	0
pnp.A257_A	11	120	23	0
pnp.B257_A	2	120	15	37
pnp	16	120	37	0
ppar.A323_D_A449_D	0	0	0	0
ppar.B323_D	112	0	110	2
ppar.B323_D_B449_D	65	0	109	2
ppar.B449_D	65	0	109	2
ppar	112	0	110	2
pr	26	0	24	3
rxr.A435_A	40	0	101	0
rxr.B435_A	13	0	96	0
rxr	16	0	90	0
sahh.A55_A_A353_A	1	0	1	0
sahh.B353_A	0	21	0	0
sahh.B55_A	0	0	0	0
sahh.B55_A_B353_A	0	0	0	0
sahh	0	21	0	0
src	6	7	11	54
tk.A58_B	0	16	1	0
tk.B58_B	0	0	8	0
tk	6	118	3	0
trypsin	102	109	118	110
vegfr2	3	0	6	0

Crystal Sidechains

Tests of DarwinDock were also performed using the crystal sidechains without alanization using the same range of scoring parameters as in the "best case" tests. As it is not really possible to have crystal sidechains available in a real-world situation, these calculations are provided as a reference, not as guidance for future calculations. Compared to the "best case" predicted sidechains, the crystal sidechains showed a preference for a higher phobic scoring scaling, but surprisingly not 100%. Instead, when 100% of the families were scored, the best result was for 100% polar and 50% phobic, although 40 and 30% phobic provided nearly identical results. At 10% of families scored 100% polar and 30% phobic was the best combination. The shift toward higher phobic scoring was not surprising due to the lack of alanization. The nonpolar residues, which were removed for the predicted sidechains test, interact with ligands predominantly via van der Waals energy. What was surprising about these results is that including the full van der Waals energy was not the best. These results are summarized in Figure 5.

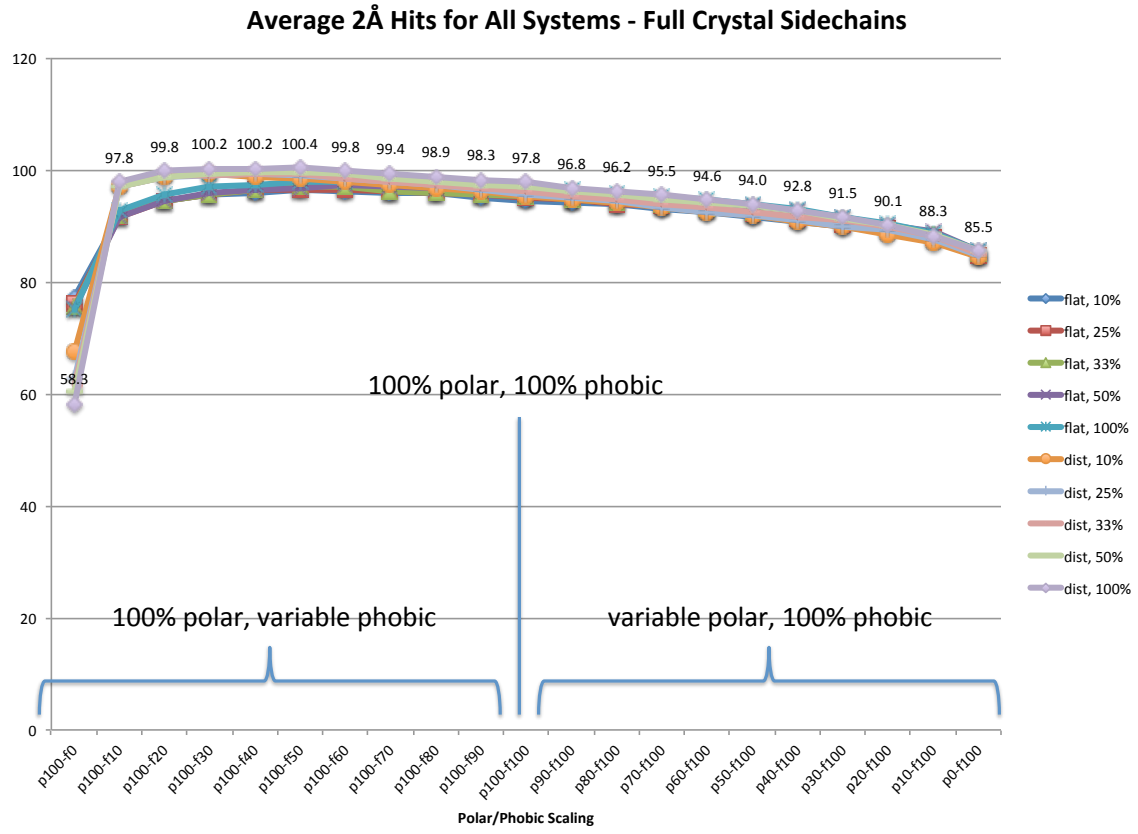

Figure 2-5 – Analysis of docking to the crystal, non-alanized sidechains. Unlike with the "best" predicted sidechains, these results show a preference for a higher phobic content, although not 100%. The best result is for distance-dependent dielectric and 100% polar, 50% phobic when scoring 100% of the families. This drops to 100% polar, 30% phobic when only scoring 10% of the families.

Docking calculations for crystal sidechains *with* alanization using 100% polar, 10% phobic and distance-dependent dielectric showed many fewer 2Å hits than the crystal, *non*-alanized results, but still more than the "best" predicted sidechains. Two factors explain the lost hits in the crystal, alanized case. First, removing the nonpolar residues obviously removed whatever van der Waals contributions those sidechains make to the binding energy. Second, alanization dramatically increased the size of the binding site for many systems. This allows ligands to make spurious interactions with polar sidechains that wouldn't be possible if they were blocked by the nonpolar residues. It also increased the

number of poses necessary for complete sampling of the binding site, meaning that the final 120 poses represented a smaller percentage of all the poses generated and scored.

Table 2-7 - Number of 2Å hits when using full crystal sidechains.
(Columns are split for compactness.)

system	xtl, full
ache	120
cox1	120
cox2	120
dhfr	120
er_ag	120
gr	120
hivrt	120
inha	120
mr	120
p38	120
pnp	120
ppar	120
pr	120
rxr	120
sahh	120
tk	120
trypsin	120
ace	119
ar	119
alr2	118
gart	118

system	xtl, full
hivpr	117
vegfr2	116
gpb	99
er_ant	98
hmga	92
pde5	92
ada	87
fxa	75
na	64
comt	54
parp	49
fgfr1	45
egfr	37
src	25
hsp90	16
ampc	9

failed	0
avg	97.0
max	120

Conclusions

The results in Table 5 and Table 6 show that DarwinDock is broadly successful when tested against the DUD set. Only two systems completely fail to produce any 2Å hits across all four sidechain predictions (flat- or distance-dependent dielectric, with or without alanization). Several other systems only have small numbers of hits. This is in contrast to

Table 7, which shows that nearly half of the systems have a full set of 120 2Å hits when the full, crystal sidechains are used. This comparison illustrates that the difficulty is not with DarwinDock, but rather the inaccuracy of sidechain predictions. Here the fundamental problem is that in the apo-protein the best sidechain conformations often invade the binding site.

The ideal starting point for a docking calculation would be a crystal structure with a related ligand already bound. In such a situation it should be possible to identify what residues are likely to move and which are not. A crystal structure without a ligand would also provide some insight. Both of those situations would relieve some of the uncertainty of the sidechain positions and yield good docking results with DarwinDock.

Of course a most interesting case is where there is no crystal structure for the protein. Indeed, most applications of DarwinDock have been for cases where the protein structure was predicted. With such *ab initio* starting structures things are more challenging. Clearly the tests discussed above show that some predicted sidechains are reliable and some are not. If one has the time to focus on a single system it may be possible to improve the odds of getting good docking results by trying multiple combinations of sidechain conformations and by using available experimental knowledge of the system.

These results show that DarwinDock is a reliable method for generating docked poses for small molecule ligands. The primary improvements necessary to the docking process are not with DarwinDock itself, but obtaining a good structure to dock to. As such,

DarwinDock is a useful tool for investigating the interactions between proteins and small molecules.

References

1. Scott, C. E., Ahn, K. H., Graf, S. T., William A Goddard, I., Kendall, D. A., & Abrol, R. (2016). Computational Prediction and Biochemical Analyses of New Inverse Agonists for the CB1 Receptor. *Journal of Chemical Information and Modeling*, *56*(1), 201–212. http://doi.org/10.1021/acs.jcim.5b00581
2. Li, Q., Kim, S.-K., Goddard, W. A., III, Chen, G., & Tan, H. (2015). Predicted Structures for Kappa Opioid G-Protein Coupled Receptor Bound to Selective Agonists. *Journal of Chemical Information and Modeling*, *55*(3), 614–627. http://doi.org/10.1021/ci500523z
3. Abrol, R., Trzaskowski, B., Goddard, W. A., III, Nesterov, A., Olave, I., & Irons, C. (2014). Ligand- and mutation-induced conformational selection in the CCR5 chemokine G protein-coupled receptor. *Proceedings of the National Academy of Sciences*, *111*(36), 13040–13045. http://doi.org/10.1073/pnas.1413216111
4. Bray, J. K., Abrol, R., Goddard, W. A., III, Trzaskowski, B., & Scott, C. E. (2014). SuperBiHelix method for predicting the pleiotropic ensemble of G-protein–coupled receptor conformations. *Proceedings of the National Academy of Sciences*, *111*(1), E72–E78. http://doi.org/10.1073/pnas.1321233111
5. Kim, S.-K., & Goddard, W. A. (2014). Predicted 3D structures of olfactory receptors with details of odorant binding to OR1G1. *Journal of Computer-Aided Molecular Design*, *28*(12), 1175–1190. http://doi.org/10.1007/s10822-014-9793-4
6. Kim, S.-K., Goddard, W. A., Yi, K. Y., Lee, B. H., Lim, C. J., & Trzaskowski, B. (2014). Predicted Ligands for the Human Urotensin-II G Protein-Coupled Receptor with Some Experimental Validation. *ChemMedChem*, *9*(8), 1732–1743. http://doi.org/10.1002/cmdc.201402087
7. Tan, J., Abrol, R., Trzaskowski, B., & William A Goddard, I. (2012). 3D Structure Prediction of TAS2R38 Bitter Receptors Bound to Agonists Phenylthiocarbamide (PTC) and 6-n-Propylthiouracil (PROP). *Journal of Chemical Information and Modeling*, *52*(7), 1875–1885. http://doi.org/10.1021/ci300133a
8. Kim, S.-K., Fristrup, P., Abrol, R., & William A Goddard, I. (2011a). Structure-Based Prediction of Subtype Selectivity of Histamine H3 Receptor Selective Antagonists in Clinical Trials. *Journal of Chemical Information and Modeling*, *51*(12), 3262–3274. http://doi.org/10.1021/ci200435b
9. Kim, S.-K., Riley, L., Abrol, R., Jacobson, K. A., & Goddard, W. A. (2011b). Predicted structures of agonist and antagonist bound complexes of adenosine A3 receptor. *Proteins: Structure, Function, and Bioinformatics*, *79*(6), 1878–1897. http://doi.org/10.1002/prot.23012
10. Niu Huang, Brian K Shoichet, A., & Irwin, J. J. (2006). Benchmarking Sets for Molecular Docking. *Journal of Medicinal Chemistry*, *49*(23), 6789–6801. http://doi.org/10.1021/jm0608356

11. Allen, W. J., Balius, T. E., Mukherjee, S., Brozell, S. R., Moustakas, D. T., Lang, P. T., et al. (2015). DOCK 6: Impact of new features and current docking performance. *Journal of Computational Chemistry*, *36*(15), 1132–1156. http://doi.org/10.1002/jcc.23905
12. Mayo, S. L., Olafson, B. D., & Goddard, W. A. (2002). DREIDING: a generic force field for molecular simulations. *The Journal of Physical Chemistry*, *94*(26), 8897–8909. http://doi.org/10.1021/j100389a010
13. Lim, K. T., Brunett, S., Iotov, M., & McClurg, R. B. (1997). Molecular dynamics for very large systems on massively parallel computers: the MPSim program. *Journal of Computational Chemistry*, *18*(4), 501-521
14. The PyMOL Molecular Graphics System, Version 1.3 Schrödinger, LLC.
15. Protein Preparation Wizard 2015-2; Epik version 2.4, Schrödinger, LLC, New York, NY, 2015; Impact version 5.9, Schrödinger, LLC, New York, NY, 2015; Prime version 3.2, Schrödinger, LLC, New York, NY, 2015.
16. Olsson, M. H. M., Søndergaard, C. R., Rostkowski, M., & Jensen, J. H. (2011). PROPKA3: Consistent Treatment of Internal and Surface Residues in Empirical pKa Predictions. *Journal of Chemical Theory and Computation*, *7*(2), 525–537. http://doi.org/10.1021/ct100578z
17. Sastry, G. M., Adzhigirey, M., Day, T., Annabhimoju, R., & Sherman, W. (2013). Protein and ligand preparation: parameters, protocols, and influence on virtual screening enrichments. *Journal of Computer-Aided Molecular Design*, *27*(3), 221–234. http://doi.org/10.1007/s10822-013-9644-8
18. A D MacKerell, J., Bashford, D., Bellott, M., R L Dunbrack, J., Evanseck, J. D., Field, M. J., et al. (1998). All-Atom Empirical Potential for Molecular Modeling and Dynamics Studies of Proteins †. *The Journal of Physical Chemistry B*, *102*(18), 3586–3616. http://doi.org/10.1021/jp973084f
19. MacroModel, version 10.9, Schrödinger, LLC, New York, NY, 2015.
20. Jaguar, version 8.9, Schrödinger, LLC, New York, NY, 2015.
21. Kam, V. W. T., & William A Goddard, I. (2008). Flat-Bottom Strategy for Improved Accuracy in Protein Side-Chain Placements. *Journal of Chemical Theory and Computation*, *4*(12), 2160–2169. http://doi.org/10.1021/ct800196k

List of Figures

List of Tables