JOURNAL OF CHEMICAL INFORMATION AND MODELING

Letter

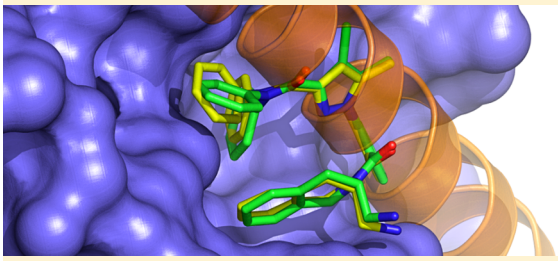
pubs.acs.org/jcim

How Good Are State-of-the-Art Docking Tools in Predicting Ligand Binding Modes in Protein−Protein Interfaces?

Dennis M. Krüger, Gisela Jessen, and Holger Gohlke*

Institute of Pharmaceutical and Medicinal Chemistry, Department of Mathematics and Natural Sciences, Heinrich-Heine-University, Universitätsstr.1, 40225 Düsseldorf, Germany

Supporting Information



ABSTRACT: Protein−protein interfaces (PPIs) are an important class of drug targets. We report on the first large-scale validation study on docking into PPIs. DrugScore-adapted AutoDock3 and Glide showed good success rates with a moderate drop-off compared to docking to "classical targets". An analysis of the binding energetics in a PPI allows identifying those interfaces that are amenable for docking. The results are important for deciding if structure-based design approaches can be applied to a particular PPI.

Molecular docking is one of the widely used approaches for structure-based lead finding and optimization in computational drug design.[1] Predicted complex configurations are used for studying protein−ligand interactions, estimating binding affinities, and, as a final filter step, in virtual screening.[2] During the past 30 years, a plethora of protein−ligand docking tools were developed, mostly aiming at predicting poses of ligands binding to "classical" targets, such as enzymes or receptors.[3] In contrast, much less effort has been devoted to predicting conformations of ligands that bind to protein−protein interfaces, so-called protein−protein interaction modulators (PPIM). Protein−protein interfaces provide an important new class of drug targets because protein−protein interactions are involved in nearly all biological processes.[4,5] Large-scale validation studies on docking into protein−protein interfaces have not yet been reported, despite the fact that protein−protein interfaces provide major challenges for structure-based ligand design approaches,[6,7] for at least two reasons: First, in contrast to "classical" targets, protein−protein interfaces are rather flat and usually lack a distinct binding pocket.[8] Second, due to the often large size of protein−protein interfaces (∼1200 to ∼4660 Å^2),[9] interactions that are favorable for binding can be widely distributed over the interface. Hence, it has remained elusive so far whether state-of-the-art docking tools are generally applicable for protein−protein interfaces. This provided the incentive for us to assess the predictive power of two commonly used protein−ligand docking approaches, AutoDock[10] and Glide,[11] with respect to pose prediction in protein−protein interfaces. For that purpose, we prepared a "PPIM data set" of 22 different target proteins consisting of 80 crystal- and NMR-structures where the ligand binds to an experimentally determined protein−protein interface (Supporting Information (SI) Table S1). The results are finally discussed in terms of structural properties of the protein−protein interfaces and physicochemical properties of the ligands.

AutoDock is the most widely used protein−ligand docking tool[12] and has been applied in multiple studies and drug design projects since its first release in the late 1990s.[12,13] Nowadays, three versions of AutoDock are available, version 3,[10] 4,[14] and AutoDock Vina,[15] with the latter one being a new implementation where the docking procedure differs from the other versions.[15] While AutoDock3 and 4 use a force field-based scoring approach, AutoDock Vina uses a combination of knowledge-based potentials and empirical scoring.[10,14,15] The authors of AutoDock Vina report success rates of 49% and 78% for AutoDock4 and AutoDock Vina, respectively, to identify a near-native conformation (rmsd ≤ 2 Å) on a data set of 192 complexes that has already been used as a training set for the AutoDock3 scoring function.[15] In previous work, AutoDock3 was successfully adapted for use with the knowledge-based pair-potentials of DrugScore[16] as an objective function.[2] Encouragingly, the AutoDock3 results (51%) for identifying good binding geometries (rmsd ≤ 2 Å) could be improved significantly when using DrugScore as an objective function (61%) for a data set of 41 complexes.[2] In another study, we reported a success rate of 77.1% for a data set of 48 complexes.[17] The Glide docking suite is a widely used commercial docking tool.[11] The Glide program comes along with its own physics-based scoring function GlideScore.[11] Glide has been reported to be successful in several studies;[18,19] e.g., Friesner et al. reported a success rate of 83.8% (rmsd ≤ 2 Å) for a data set of 282 protein−ligand complexes.[11]

So far, a general AutoDock4 evaluation study has not been reported, as have not been results obtained for the combination of AutoDock4 and DrugScore. A large-scale comparison between all three AutoDock versions and Glide has not been reported either. Thus, in the present study, we first aimed at evaluating the protein−ligand docking tools with respect to success rates on "classical" targets before evaluating these tools'

Published: October 16, 2012

ACS Publications © 2012 American Chemical Society

dx.doi.org/10.1021/ci3003599 | J. Chem. Inf. Model. 2012, 52, 2807−2811

applicability for docking to protein−protein interfaces. For that purpose, we evaluated AutoDock3, AutoDock4, and AutoDock Vina together with their native scoring functions, DrugScore-adapted versions of AutoDock3 and 4, and GlideSP on a total of 224 protein−ligand complexes retrieved from the CCDC/Astex clean list,[20] a data set that has been established for the purpose of protein−ligand docking studies.[21,22] For detailed results including PDB codes of the protein−ligand complexes and rmsd values see SI Table S2. Best results from redocking experiments on the Astex clean set were obtained with DrugScore-adapted AutoDock3 (success rate 64.3%; Figure 1

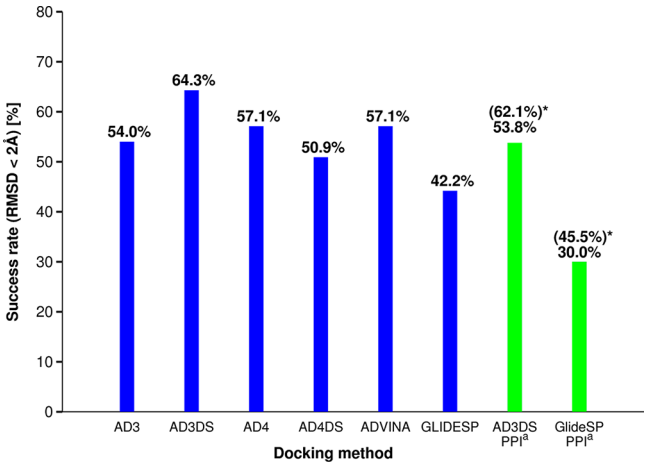


Figure 1. Success rates for redocking 224 complexes of the CCDC/Astex clean set with three versions of AutoDock (3, 4, and Vina), two DrugScore-adapted AutoDock versions (3 and 4), and GlideSP (blue bars) as well as results obtained for redocking 80 complexes from the PPIM data set (green bars). The following abbreviations are used: (AD3) docking was performed with AutoDock3.0.5; (AD4) docking was performed with AutoDock4.2.3; (AD3DS) docking was performed with AutoDock3.0.5 using DrugScore as objective function; (AD4DS) docking was performed with AutoDock4.2.3 using DrugScore as an objective function; (ADVINA) docking was performed with AutoDock Vina1.1.1. For AutoDock3 and 4, the rmsd between the docking pose found on the first scoring rank of the largest cluster and the native solution was chosen. For AutoDock Vina and GlideSP, the rmsd between the docking pose found on the first scoring rank and the native solution was chosen. For details, see SI Tables S1 and S2: (a) performed on a data set of 80 PPIM; (*) neglecting peptides with >20 rotatable bonds.

and S1) for identifying near-native ligand geometries (rmsd ≤ 2 Å). This combination also showed the best convergence behavior, i.e. the most pronounced tendency to create large clusters of similar ligand conformations (see also "Evaluation of docking accuracy" in the SI page S3), with 71.0% of the clusters comprising more than 30 ligand configurations (SI Figure S2). The success rates of AutoDock3 (54.0%) and AutoDock4 (57.1%) with their native scoring functions are almost similar, but in general AutoDock3 runs converged better (SI Figure S2) and were on average 20 min faster (∼1 h per docking run; SI Figure S3). AutoDock Vina was as successful as AutoDock4 (success rate of 57.1%) but 30-fold faster than AutoDock3 (∼2 min per docking run; SI Figure S3). The fastest docking tool was GlideSP with on average ∼80 s per docking run, but its predictive power was lowest (success rate 42.2−46.0% depending on whether the receptor structure was minimized or not upon setup; Figure 1). Note that for all of the above dockings, the preparation of the protein and ligand structures was done *independently*, i.e. after separating protein and ligand. In contrast, a significantly improved success rate is obtained for GlideSP if the protein and ligand structures are prepared *while still being in the complex*, as described by Friesner et al.[11] As such, for a subset of 135 complexes of the Astex clean set for which Friesner et al. had reported a success rate of 72.6%, we obtained a success rate of 40.7−44.4% (depending on whether the receptor structure was minimized or not upon setup) when preparing protein and ligand independently and 75.6% when the protein and ligand were prepared while still being in the complex (SI Figure S4; Table S3). Note that the latter procedure requires knowledge of the complex crystal structure and, hence, precludes the use of the docking tool in a predictive manner. For the remainder of this study, we will thus resort to preparing the protein and ligand structures independently.

On the basis of the above results, DrugScore-adapted AutoDock3 and GlideSP were selected for docking experiments to target protein−protein interfaces. The ability of the docking tools to predict near-native ligand geometries in protein−protein interfaces was investigated on a data set of 22 different target proteins consisting of 80 crystal- and NMR-structures from the PDB database (SI Table S1); to the best of our knowledge, this data set constitutes the largest PPIM−protein complex data set assembled so far. PPIM complex structures were selected from the literature, and it was required that either the target protein interface has been described or that the corresponding protein−protein complex has been resolved (for references, see SI Table S1). The data set contains 26 (modified) peptide ligands (SI Table S1). Since peptides in general do not match the definition of a small-molecule inhibitor and provide a challenge to docking approaches, we will discuss them separately.[23] We furthermore excluded results from peptides with >20 rotatable bonds (in total 14 structures) because it is known that highly flexible ligands often do not lead to converged docking results.[12] Accordingly, for 85.7% of those peptides, AutoDock did not converge during docking, and no near-native docking solution could be obtained.

In general, docking to protein−protein interfaces was quite successful for AutoDock3 and DrugScore: We obtained success rates of 53.8% (62.1% when excluding all peptides with >20 rotatable bonds; 57.4% when only considering small-molecule PPIM). These results compare favorably with those achieved from protein−ligand docking performed to "classical targets" of the CCDC/Astex clean list (drop-off of the success rate by ∼10%; Figure 1). Surprisingly, when only considering the 12 peptides with ≤20 rotatable bonds, a success rate of 76.9% is observed. In the case of Glide, the success-rate for docking to protein−protein interfaces is 30.0% (45.5% when excluding all peptides with >20 rotatable bonds), thus being ∼12% lower than the rate obtained for docking to classical targets (Figure 1). The AutoDock3 clustering dropped off by almost 30% compared to docking to classical targets when considering the percentage of clusters that consist of more than 10, 30, and 50 conformations, respectively (SI Figure S2). This finding is paralleled by the average runtime increasing by a factor of 7 (to up to 400 min per docking run; SI Figure S3). These results demonstrate that the structural characteristics of a protein−protein interface have a drastic effect on AutoDock's ability to find converged docking solutions: likely, the flatness and solvent accessibility of the protein−protein interface drive the algorithm through the maximum number of energy evaluations. The same was found for docking with GlideSP: The average runtime for GlideSP increased by a factor of 13 to about 10 min

per docking run. Furthermore, no solution could be found by GlideSP for those 7 ligands where more than 50% of the ligand is exposed to the solvent without making any interactions to the protein in the complex crystal structure (SI Table S1 and S4).

We will now analyze in more detail two examples of successful dockings with AutoDock3 and DrugScore to protein−protein interfaces. Figure 2 shows the crystal structure

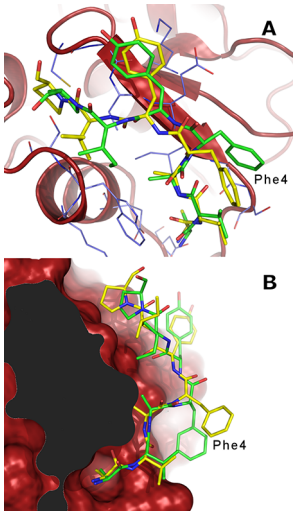


Figure 2. Reaper N-terminal peptide of 10 amino acids length binding to the BIR1 domain of *Drosophila melanogaster* inhibitor of apoptosis protein (DIAP) depicted in red (PDB code: 1SDZ). The crystal structure configuration is depicted in green, the docking solution from the first rank of the largest cluster obtained by AutoDock3 and DrugScore is depicted in yellow (rmsd to the native configuration: 0.99 Å). (A) Top view of the protein−protein interface and its inhibitor. Residues involved in ligand binding are colored in blue. (B) Lateral view of the protein−protein interface. Note the bulge character of the binding site and the flat interface.

of a reaper N-terminal peptide that binds to the BIR1 domain of *Drosophila melanogaster* inhibitor of apoptosis protein (DIAP) as well as the corresponding docking solution. The binding site shows a rather flat protein−protein interface. Visual inspection revealed that this holds for most of the peptide-binding interfaces in our data set. Except one phenylalanine (Phe4), all residues of the peptide are in contact with the protein, which may be the reason as to why the docking of the peptide was nevertheless successful. During docking the algorithm tries to maximize the number of interactions of the ligand to the protein to optimize the binding score. Unfortunately, this procedure will lead to false predictions if parts of a ligand do not interact with the protein−protein interface.

In contrast, the protein−protein interface of the apoptosis regulator BCL-2 forms a groove to which a phenyl pyrazole ligand binds (Figure 3). The ligand is almost completely buried so that this example resembles docking to a "classical" target. Accordingly, we obtained a redocking solution with an rmsd of 0.71 Å with AutoDock3 and DrugScore (Figure 3). Note that the pocket in the BCL-2 interface is a transient pocket that only exists in a partially open state in the unbound form of the protein (PDB code: 1R2D). Recently, it was shown for the case of interleukin-2 that transient pockets could be identified by constrained geometric simulations starting from the unbound protein.[24] These transient pockets were successfully addressed

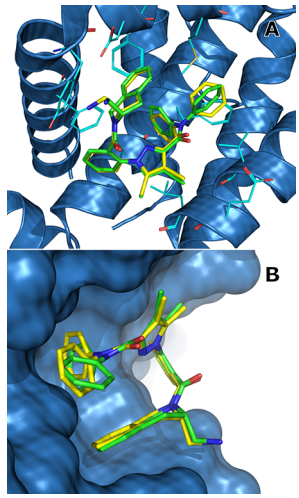


Figure 3. Phenyl pyrazole binding to the apoptosis regulator BCL-2 depicted in blue (PDB code: 2W3L). The crystal structure configuration is depicted in green, the docking solution from the first rank of the largest cluster obtained with by AutoDock3 and DrugScore is depicted in yellow (rmsd: 0.71 Å). (A) Top view of the protein−protein interface and its inhibitor. Residues involved in ligand binding are colored in cyan. (B) Lateral view of the protein−protein interface. Note that the interface is rather well-structured and contains a deep pocket, which makes this case similar to docking to a "classical" target.

by molecular docking, too, and a retrospective virtual screening assessment based on them showed convincing enrichment factors.[24] To further address the question to what extent conformational changes occur in protein−protein interfaces during binding, we calculated interface rmsd values between the ligand-bound and unbound protein conformations (SI Table S1). The values range from about 0.68 Å for HexA to almost 10 Å for Calmodulin. Interestingly, the rmsd values are similar for all protein−ligand complexes of the same protein. Thus, the ligand seems to have little influence on the conformational change; in turn, this finding supports the conformational selection model according to which transient pockets may open in interface regions even in the absence of a ligand.

Finally, we juxtaposed docking success rates and interface properties as well as chemical descriptors of the PPIMs to identify possible reasons as to why docking succeeds or fails. In order to describe the characteristics of a protein−protein interface, an interface was defined as all protein residues within 5 Å distance of the corresponding ligand. The interface residues were classified into four groups based on their physicochemical properties (hydrophobic, aromatic, polar, and charged) and counted to determine the frequency of occurrence of each residue type. In addition, we determined the number of salt bridges that a protein and ligand form by visual inspection and calculated the volume of the interface pocket for each of the receptors (SI Table S5). Unfortunately, none of these interface properties showed any significant relation to docking success or failure neither when using single- nor multiple-linear regression.

We then used the recently developed DrugScore^{PPI} web server[25] to predict hotspots in those protein interfaces where a crystal structure of the protein−protein complex is available (SI Table S5). This revealed that docking was mostly successful on target proteins that provide at least one charged residue (Arg, Glu, or Asp) as a hotspot: if only those targets are considered from the PPIM data set, the success rate raises to 82.8% in the

dx.doi.org/10.1021/ci3003599 | J. Chem. Inf. Model. 2012, 52, 2807−2811

case of docking with AutoDock3 and DrugScore. Such information is valuable if it comes to deciding if a structure-based ligand finding approach is to be used on a novel protein−protein interface.

Finally, we assessed the ligand properties molecular weight (MW), polarity, and potency by means of the binding efficiency index (BEI) and the surface efficiency index (SEI) and related them to our docking results (SI Table S4). BEI and SEI have already been applied to PPIMs to describe ligand efficiency in compound assessment.[26,27] Mapping in the SEI-BEI plane the PPIMs of our data set (SI Figures S5 and S6) reveals that the data centers around 12.8 ± 4.4 for BEI and 5.8 ± 3.1 for SEI. These mean values are roughly half (BEI) and one-third (SEI) as large as what has been found for 122 marketed drugs,[26] reflecting a generally lower binding efficiency of the PPIMs. Docking was successful for most of the PPIMs if BEI ≥ 10; such a value represents a ligand with, e.g., pKi = 5 and MW = 0.5 kDa. Likewise, all successful dockings had SEI < 10, reflecting on the ligand side what has been found above when analyzing hot spots in the protein−protein interfaces: PPIMs that participate in (a) polar (hotspot) interaction(s) are easier to dock. Still, we did not find a correlation between the number of polar hotspots and the ligand polar surface area, rendering it impossible to predict only from ligand information whether docking of that ligand to a particular target will be successful.

In summary, the most outstanding result of this study is that both DrugScore-adapted AutoDock3 and Glide showed good success rates when docking PPIM to protein−protein interfaces with only a moderate drop-off (10% and 12%, respectively) compared to docking to "classical targets". This suggests that rational, structure-based ligand finding approaches should be applicable to identifying PPIM. Furthermore, an a priori analysis of the binding energetics of a protein−protein interaction should help identifying those interfaces that are particularly amenable for these approaches: these are the ones with at least one charged hotspot residue in them. In turn, the pocket volume and flatness of the interface turned out to be less decisive for the docking success, as good docking solutions were also obtained for interface regions with small pockets and flat interfaces.

■ ASSOCIATED CONTENT

Supporting Information

Detailed docking results as well as calculated protein interface and ligand properties. This material is available free of charge via the Internet at http://pubs.acs.org.

■ AUTHOR INFORMATION

Corresponding Author

*To whom correspondence should be addressed. Fax: (+49) 211-8113847. E-mail: gohlke@uni-duesseldorf.de.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

We thank the Cambridge Crystallographic Data Center for granting a Relibase+ license to us, OpenEye Scientific Software, Inc. for granting an OEChem license to us, and the "Zentrum für Informations- und Medientechnologie" (ZIM) at Heinrich-Heine-University, Düsseldorf, for computational support.

■ REFERENCES

(1) Jorgensen, W. L. The many roles of computation in drug discovery. *Science* **2004**, *303* (5665), 1813−8.
(2) Sotriffer, C. A.; Gohlke, H.; Klebe, G. Docking into knowledge-based potential fields: A comparative evaluation of DrugScore. *J. Med. Chem.* **2002**, *45* (10), 1967−1970.
(3) Gohlke, H.; Klebe, G. Approaches to the description and prediction of the binding affinity of small-molecule ligands to macromolecular receptors. *Angew. Chem., Int. Ed. Engl.* **2002**, *41* (15), 2644−76.
(4) Wells, J. A.; McClendon, C. L. Reaching for high-hanging fruit in drug discovery at protein-protein interfaces. *Nature* **2007**, *450* (7172), 1001−9.
(5) Wells, J. A.; McClendon, C. L. Reaching for high-hanging fruit in drug discovery at protein-protein interfaces. *Nature* **2007**, *450* (7172), 1001−1009.
(6) Gonzalez-Ruiz, D.; Gohlke, H. Targeting protein-protein interactions with small molecules: challenges and perspectives for computational binding epitope detection and ligand finding. *Curr. Med. Chem.* **2006**, *13* (22), 2607−25.
(7) Metz, A.; Ciglia, E.; Gohlke, H. Modulating Protein-Protein Interactions: From Structural Determinants of Binding to Druggability Prediction to Application. *Curr. Pharm. Des.* **2012**, *18*, 4630−4647.
(8) Jones, S.; Thornton, J. M. Principles of protein-protein interactions. *Proc. Natl. Acad. Sci. U.S.A.* **1996**, *93* (1), 13−20.
(9) Lo Conte, L.; Chothia, C.; Janin, J. The atomic structure of protein-protein recognition sites. *J. Mol. Biol.* **1999**, *285* (5), 2177−98.
(10) Morris, G. M.; Goodsell, D. S.; Halliday, R. S.; Huey, R.; Hart, W. E.; Belew, R. K.; Olson, A. J Automated Docking Using a Lamarckian Genetic Algorithm and and Empirical Binding Free Energy Function. *J. Comput. Chem.* **1998**, *19*, 1639−1662.
(11) Friesner, R. A.; Banks, J. L.; Murphy, R. B.; Halgren, T. A.; Klicic, J. J.; Mainz, D. T.; Repasky, M. P.; Knoll, E. H.; Shelley, M.; Perry, J. K.; Shaw, D. E.; Francis, P.; Shenkin, P. S. Glide: a new approach for rapid, accurate docking and scoring. 1. Method and assessment of docking accuracy. *J. Med. Chem.* **2004**, *47* (7), 1739−49.
(12) Morris, G. M.; Huey, R.; Olson, A. J. Using AutoDock for ligand-receptor docking. *Curr. Protoc. Bioinform.* **2008**, *24*, 8.14.1−8.14.40.
(13) Goodsell, D. S. Computational docking of biomolecular complexes with AutoDock. *Cold Spring Harb. Protoc.* **2009**, *2009* (5), pdb prot5200.
(14) Huey, R.; Morris, G. M.; Olson, A. J.; Goodsell, D. S. A semiempirical free energy force field with charge-based desolvation. *J. Comput. Chem.* **2007**, *28* (6), 1145−52.
(15) Trott, O.; Olson, A. J. AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *J. Comput. Chem.* **2010**, *31* (2), 455−61.
(16) Gohlke, H.; Hendlich, M.; Klebe, G. Knowledge-based scoring function to predict protein-ligand interactions. *J. Mol. Biol.* **2000**, *295* (2), 337−356.
(17) Kazemi, S.; Krüger, D. M.; Sirockin, F.; Gohlke, H. Elastic potential grids: accurate and efficient representation of intermolecular interactions for fully flexible docking. *ChemMedChem* **2009**, *4* (8), 1264−8.
(18) Krüger, D. M.; Evers, A. Comparison of structure- and ligand-based virtual screening protocols considering hit list complementarity and enrichment factors. *ChemMedChem* **2010**, *5* (1), 148−58.
(19) Cross, J. B.; Thompson, D. C.; Rai, B. K.; Baber, J. C.; Fan, K. Y.; Hu, Y.; Humblet, C. Comparison of several molecular docking programs: pose prediction and virtual screening accuracy. *J. Chem. Inf. Model.* **2009**, *49* (6), 1455−74.
(20) Nissink, J. W.; Murray, C.; Hartshorn, M.; Verdonk, M. L.; Cole, J. C.; Taylor, R. A new test set for validating predictions of protein-ligand interaction. *Proteins* **2002**, *49* (4), 457−71.
(21) Korb, O.; Stützle, T.; Exner, T. E. Empirical scoring functions for advanced protein-ligand docking with PLANTS. *J. Chem. Inf. Model.* **2009**, *49* (1), 84−96.

(22) Thompson, D. C.; Humblet, C.; Joseph-McCarthy, D. Investigation of MM-PBSA rescoring of docking poses. *J. Chem. Inf. Model.* **2008**, *48* (5), 1081−91.
(23) Fuller, J. C.; Burgoyne, N. J.; Jackson, R. M. Predicting druggable binding sites at the protein-protein interface. *Drug Discov. Today* **2009**, *14* (3−4), 155−61.
(24) Metz, A.; Pfleger, C.; Kopitz, H.; Pfeiffer-Marek, S.; Baringhaus, K. H.; Gohlke, H. Hot spots and transient pockets: Predicting the determinants of small-molecule binding to a protein-protein interface. *J. Chem. Inf. Model.* **2012**, *52*, 120−133.
(25) Krüger, D. M.; Gohlke, H. DrugScorePPI webserver: fast and accurate in silico alanine scanning for scoring protein-protein interactions. *Nucleic Acids Res.* **2010**, *38* (Web Server issue), W480-6.
(26) Abad-Zapatero, C.; Metz, J. T. Ligand efficiency indices as guideposts for drug discovery. *Drug Discov Today* **2005**, *10* (7), 464−9.
(27) Morelli, X.; Bourgeas, R.; Roche, P. Chemical and structural lessons from recent successes in protein-protein interaction inhibition (2P2I). *Curr. Opin. Chem. Biol.* **2011**, *15* (4), 475−481.