THE JOURNAL OF PHYSICAL CHEMISTRY B

Article

Cite This: J. Phys. Chem. B XXXX, XXX, XXX−XXX

pubs.acs.org/JPCB

Ensemble Docking in Drug Discovery: How Many Protein Configurations from Molecular Dynamics Simulations are Needed To Reproduce Known Ligand Binding?

Wilfredo Evangelista Falcon,[†,§,‡] Sally R. Ellingson,[‡] Jeremy C. Smith,[*,†,‡] and Jerome Baudry[*,∥]

[†]Department of Biochemistry and Cellular and Molecular Biology, University of Tennessee, Knoxville, Tennessee 37996, United States

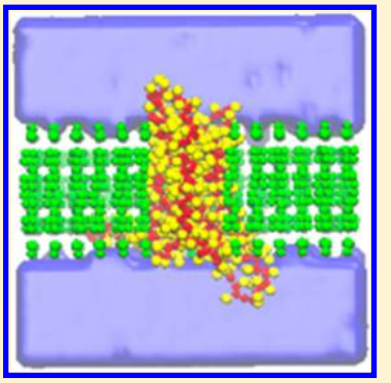
[‡]UT/ORNL Center for Molecular Biophysics, Oak Ridge, Tennessee 37830, United States

[§]College of Medicine, University of Kentucky, Lexington, Kentucky 40506, United States

[∥]Department of Biological Sciences, The University of Alabama in Huntsville, Huntsville, Alabama 35899, United States

Supporting Information

ABSTRACT: Ensemble docking in drug discovery or chemical biology uses dynamical simulations of target proteins to generate binding site conformations for docking campaigns. We show that 600 ns molecular dynamics simulations of four G-protein-coupled receptors in their membrane environments generate ensembles of protein configurations that, collectively, are selected by 70−99% of the known ligands of these proteins. Therefore, the process of ligand recognition by conformational selection can be reproduced by combining molecular dynamics and docking calculations. Clustering of the molecular dynamics trajectories, however, does not necessarily identify the protein conformations that are most often selected by the ligands.



INTRODUCTION

Docking is an important tool for the in silico characterization of protein/ligand interactions. In most cases, docking aims to identify small organic molecules that are most likely to bind to a given protein target, by ranking the strength of binding. This approach is particularly common in an early-stage drug discovery, in which new hits/leads are needed, and potential ligands are predicted using virtual screening and prioritized for experimental validation.[1]

Our laboratories, as well as several others, have increasingly used an ensemble docking approach to perform docking calculations.[2] In this approach, the dynamics of the protein target is characterized, often using molecular dynamics (MD) simulations, and a number of thus-generated conformations of the protein are used in many docking calculations, rather than performing docking to a single (experimental or modeled) structure of the target. The ensemble-based approach aims to reproduce a conformational selection mechanism, in which specific protein conformations are selected by ligand(s) to form a thermodynamically favored protein/ligand complex.

Ensemble-based drug discovery has been successful in our groups in identifying new hits and leads in a variety of biochemical pathways and targets, e.g., in the coagulation pathway;[3] phosphate homeostasis,[4] and antibiotic resistance.[5] In addition to this "hit discovery" role, ensemble-based docking can be used to predict off-target protein interactions of hits/leads leading to adverse effects in preclinical and clinical trials,[6] to repurpose existing drugs[7] and to identify the protein targets of natural products.[8,9]

Ensemble-based docking is a two-step process: (1) generating an ensemble of protein conformations to be used in the docking calculations and (2) the actual docking to the selected protein structures. Taken independently, these two tasks require significant computational power, both for long MD simulations and for docking large databases of compounds. Technical solutions have been developed to allow such demanding calculations on modern hardware, such as graphics processing unit-based MD simulation engines[10] or massively parallelized versions of docking codes.[11] When combined together, the computational time requirements increase combinatorically, as all docking calculations have to be repeated for each selected protein conformation, and the computational requirements can become intractable without supercomputer architectures. This increase can be mitigated by selecting only a subset of protein conformations, for example, by clustering the conformations of the protein (or of a substructure of the protein, such as ligand binding sites) and selecting a representative structure for each cluster. The

Received: November 28, 2018
Revised: January 26, 2019
Published: January 29, 2019

DOI: 10.1021/acs.jpcb.8b11491
J. Phys. Chem. B XXXX, XXX, XXX−XXX

Table 1. Details of the Proteins and Sets of Known Ligands/Decoys Used

			DUD-E/CC−DD				
protein name	gene name	PDB id	known ligands	decoys	clustering RMSD (Å)	number of clusters	number of docking (plus X-ray structure)
adenosine receptor A2A	ADORA2A	3EML	844	10 899	2	33	399 262
β2-adrenergic receptor	ADRB2	2RH1	447	15 255	2	18	298 338
δ-type opioid receptor	OPRD1	4N6H	377	14 703	1.75	32	497 640
κ-type opioid receptor	OPRK1	4DJH	307	11 973	2.25	28	356 120

hypothesis underlying this clustering approach is that each representative structure from a cluster will be representative of the entire cluster in terms of protein/ligand docking performance, i.e., database enrichment would be similar for all conformations within a cluster.

This paper tests the above clustering hypothesis. Is the docking performance of a single representative conformation of a protein cluster indeed representative of the docking performance of the cluster members? A further question, in the context of conformational selection, is how many conformations of a protein exist that statistically outperform other conformations in terms of docking performance, and to what extent. This article compares docking performance achieved using a very large ensemble of protein conformations with that achieved using a much more limited number of representative protein conformations.

The computational cost of this study was very significant. Four G-protein-coupled receptors (GPCRs), including their membrane environment, were used in this work. For each of these proteins, coarse-grained (CG) molecular dynamics (MD) simulations were run to generate 1 μs trajectories. Docking calculations were then performed on (i) representative conformations for each protein form the 1 μs trajectories (from 18 to 32 conformations per protein, depending on the protein) and on (ii) 3000 structures representing the first 600 ns of the microsecond trajectories, sampled evenly every 0.2 ns, without structural clustering. In terms of docking calculations only, the total number of docking calculations was about 165.5 million, which required about 33.7 million processor hours on a supercomputer.

METHODS

Structural Data Preparation and MD Simulations. Four GPCR structures were downloaded from the RSCB Protein Data Bank. Their gene names and PDB IDs are listed in Table 1.

Ligand structures and known binders and decoys were obtained from the Directory of Useful Decoys-Enhanced (DUD-E)[12] for ADORA2A and ADRB2. Computational Chemistry and Drug Design (CC−DD) database[13] was used to obtain the ligands for OPRD1 and OPRK1, as summarized in Table 1.

The GPCR structures obtained from PDB are products of chimeric protein expression for crystallization, domains that do not belong to the WT GPCR sequence were deleted as well as co-crystallized ligands. Missing loops were modeled and built using MODELER 9.10.[14] Missing loops on the membrane side opposite to that of the ligand binding site and longer than 20 amino acids were not built, as they have little direct influence on the structure of the extramembrane ligand binding site.

The four protein models in their membrane environments were mapped to coarse-grained (CG) models and placed in a bilayer membrane. The plasma membrane lipids were as suggested by Leventis[15] in both inner and outer leaflets: phosphatidylcholine, phosphatidylethanolamine, phosphatidylserine, and cholesterol representing 42, 25, 14, and 19% of the membrane lipids, respectively. Water and ions were added to equilibrate the system using martinize.py v2.5 and insane.py scripts from the Martini repository.[16−19] Previous comparison of CG and full atom description of protein−membrane system showed that CG does reproduce the structural and dynamics aspects of lipid−protein interactions of the atomistic simulations.[20−22] Each of the systems, protein, membrane, ions, and water, was reduced from ∼125 000 full atoms to ∼14 000 CG particles. MD simulations were performed using the Gromacs v5.1.0[23] for 1 μs, saving frames every 200 ps. The parameters for the energy minimization, equilibration, and production time were used as in Stansfeld et al., 2015.[24] To select groups of similar structures from the trajectories, the Gromacs clustering tool, gromos, was used to build clusters of similar protein structures based on the root mean square deviation (RMSD) of the entire protein backbone.[25] The goal of clustering protein structures was to obtain a number of representative structures such that the docking calculations could be achieved in a manageable computing time on the Newton High-Performance Computer cluster of The University of Tennessee, Knoxville. About 25 clusters were obtained for each protein, using RMSD thresholds ranging from 1.75 to 2.25 Å. The representative structure of each cluster was identified as the structure with the most neighbor structures, with an RMSD value beneath the given threshold, in that cluster. These representative structures were extracted from the trajectories using Gromacs tools and back mapped to an all-atom model using Backward v0.1.[26] CG-MD and backmapping were performed on the Moldyn High-Performance cluster at the UT/ORNL Center for Molecular Biophysics, Oak Ridge, Tennessee.

Docking Calculations. We used the program VinaMPI developed in our laboratory for use on supercomputers,[11] a high throughput docking version of the AutodockVina program. VinaMPI requires input files in PDBQT format for both protein and ligands. Scripts from AutoDockTools (ADT) v.1.5.6[27] were used to preprocess the conformations obtained from the MD simulations. This preprocessing includes removing any nonprotein atoms, keeping polar hydrogen atoms on the protein and assigning Kollman charges.[28] Ligand structures were preprocessed, adding hydrogens and charges, and rotamers were set according to the default ADT method.[29] The configuration files for the virtual screening, receptors, and ligands lists were produced with Python scripts developed in the laboratory and described in Ellingson et al., 2013.[11]

The numbers of structures and ligands tested in this phase are listed in Table 2. This phase of the project was performed on the Newton High-Performance cluster of The University of Tennessee, Knoxville. The total number of docking calculations, when using the cluster conformations for the MD

DOI: 10.1021/acs.jpcb.8b11491
J. Phys. Chem. B XXXX, XXX, XXX−XXX

Table 2. Total Number of Compounds (Known Ligands, or "Actives"; and Decoys) for Each of the Four Proteins, and Total Number of Known Ligands Identified in Each of the Four Proteins in Different Subsets (Top-Scoring 0.5%; 1, 5, and 10%) of the Total Pool of Compounds Docked[a]

proteins	ligands			average of actives in random selection				actives in crystal structure				unique actives in significant frames in clustered data				unique actives in significant frames in 600 ns of trajectory			
	actives	decoys	total	0.5%	1.0%	5.0%	10.0%	0.5%	1.0%	5.0%	10.0%	0.5%	1.0%	5.0%	10.0%	0.5%	1.0%	5.0%	10.0%
ADORA2A	844	10 899	11 743	4	8	42	84	0	2	39	98	50	99	432	693	450	550	802	836
ADRB2	447	15 255	15 702	2	4	22	45	9	14	53	96	9	12	42	76	56	80	267	392
OPRD1	377	14 703	15 080	2	4	19	38	8	17	53	85	7	13	56	88	79	125	243	276
OPRK1	307	11 973	12 280	2	3	15	31	0	0	4	12	8	0	45	104	58	69	168	247

[a]A: random selection of compounds, B: statistically significant (better than random) docking in the crystal structures, C: statistically significant (better than random) docking using representative structures from clustered MD trajectories, D: docking on all 3000 conformations of 600 ns MD trajectories.

trajectories, was 1 551 360 protein/ligand complexes and generating between 1 and 10 ligand poses per protein/ligand complex. The pose of a specific ligand docked in the binding site exhibiting the most favorable binding free energy was selected for further analysis

A second set of docking calculations was performed on the Titan supercomputer in the Oak Ridge National Laboratory, using 3000 all-atom conformations per target corresponding to the first 600 ns of the 1 μs MD trajectories, with protein conformations saved every 200 ps. Additional frames beyond 600 ns trajectory could not be processed with the computational time available. These structures were prepared the same way as cluster representative structures described above. The total number of docking calculations, when using all of the conformations in the 600 ns trajectory, was 164 million. For each ligand, the ligand binding pose with the lowest energy was selected for further analysis.

Statistical Assessment of Docking Enrichment Performance. The numbers of known ligands and decoys predicted by docking to bind to a given protein structure in the top 0.5%; 1, 5, and 10% of the binding energies were recorded. A protein conformation is defined as "selected" through conformational selection if the number of known ligands in the given subset of the docking results (i.e., top 0.5%; 1, 5, and 10%) is significantly higher than what would be obtained through a random selection of compounds for the same subset of the docking results.

To establish a statistical threshold to decide whether the number of known ligands predicted to bind to a particular protein conformation is statistically significantly better than what would be obtained with a random selection of compounds, an outlier interquartile detection method was used. The interquartile method[30,31] defines the interquartile range (IQR) to set the lower and upper cut-off values $Q_1 - 1.5 \times$ IQR and $Q_3 + 1.5 \times$ IQR, respectively (see Figure 1 left panel). Values below and above these thresholds are defined as outliers. An important feature of this method is that it does not depend on the symmetry of the distribution; interquartiles can be calculated on symmetrical and asymmetrical distributions. Thus, if the number of active ligands bound to a particular frame is higher than the upper cut-off, $Q_3 + 1.5 \times$ IQR, of a random distribution, the docking results on this particular frame will be statistically significant, and this frame will be counted as a "significant frame". Values of this upper limit for each set of ligands belonging to their respective proteins are shown in tables in the Supporting Information section. For example, the chemical library for ADORA2A comprises 844 actives and 10 899 decoys. If 5% of this pool of compounds, e.g., 587 molecules, is randomly selected, then 42 actives and 545 decoys are expected on average, but the actual number of known ligands identified randomly in a specific trial follows a statistical distribution such as illustrated in Figure 1, right. In the case illustrated on Figure 1, the interquartile method yields that docking calculations would have to identify 59 known ligands in 5% of the pool of ligands to be statistically better than random enrichment (Figure 1, right panel, and Supporting Information Tables).

RESULTS AND DISCUSSION

The results show that most of the known ligand to the four GPCRs studied here can be identified from a nonclustered MD trajectory, but that clustering of the MD trajectory leads, in the

DOI: 10.1021/acs.jpcb.8b11491
J. Phys. Chem. B XXXX, XXX, XXX−XXX

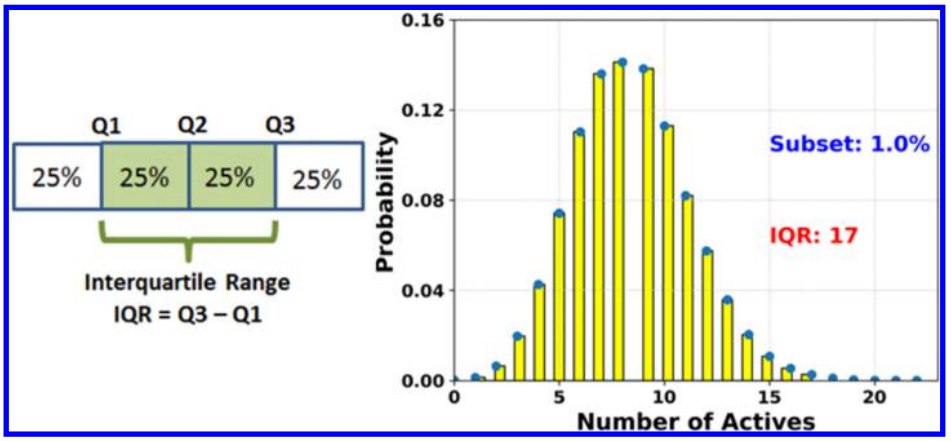


Figure 1. Interquartile definition and probability distribution for random selection at 1.0% for ADORA2A. Left panel: any set of data can be divided in four quartiles, containing 25% of the data each. The interquartile range is defined as IQR = $Q_3 - Q_1$. Right panel: probabilities of obtaining n known ligands 5.0% (587 molecules) of the compound library of ADORA2A (11 743 compounds). The yellow-colored part of the distribution corresponds to a statistical random result, and only a number of actives >59 (Q_3 + 1.5 × IQR) is statistically better than random number of known ligands identified through docking.

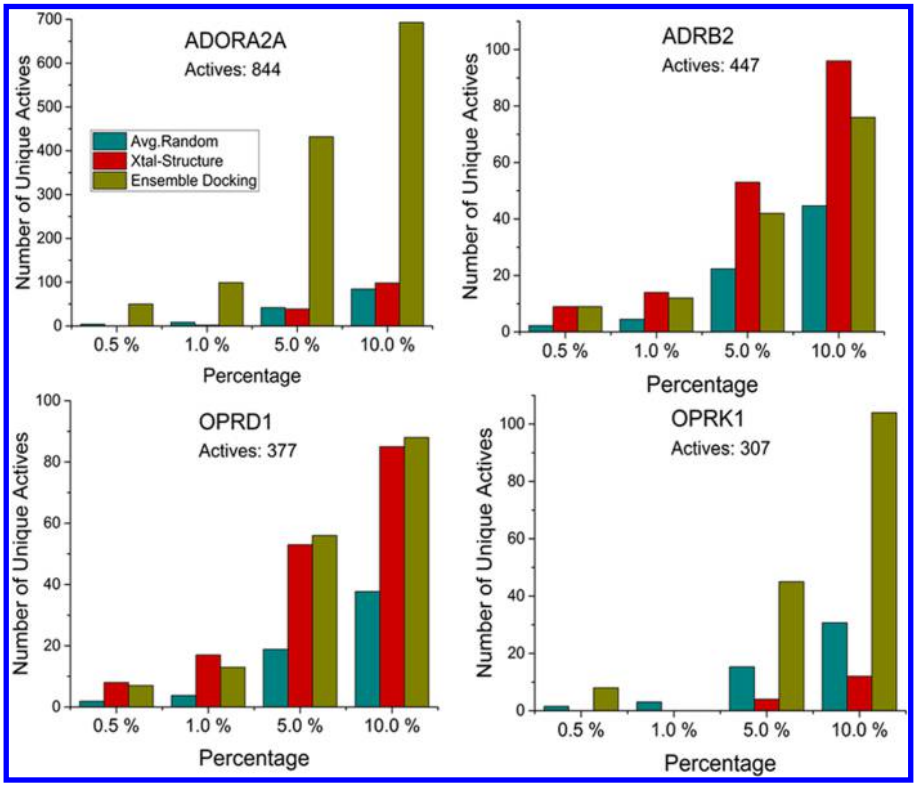


Figure 2. Number of known ligands identified in several subsets of the top-ranking docking score for each of the proteins in statistically significant (better than random) docking using representative structures from clustered MD trajectories.

present cases, to a much-reduced identification rate of protein's known ligands. These results are described in details below.

Table 2 summarizes the number of known ligands identified through random selection and docking calculations using the different docking approaches listed in Methods. Figures 2 and 3 show these results graphically. Table 3 shows the corresponding percentages of known ligands identified in statistically significant results, and Table 4 lists docking details when using all 3000 protein configurations from the first 600 ns of the MD trajectories.

Ensemble Docking Improves Docking Performance when Compared to Using a Single (Crystal) Structure. In the case of ADORA2A and OPRK1, docking using the crystal structures does not significantly identify more known ligands than a random selection of compounds would. Indeed, in most cases, the number of known ligands identified through docking is actually less than what would be obtained by randomly selecting compounds. In contrast, in the cases of ADRB1 and OPRD1, docking using the crystal structure identifies about twice as many known ligands than does selecting compounds randomly. However, in these latter two cases, using representative snapshots from a clustered trajectory does not improve the performance of docking beyond what is achieved using a single-crystal structure. Table

DOI: 10.1021/acs.jpcb.8b11491
J. Phys. Chem. B XXXX, XXX, XXX−XXX

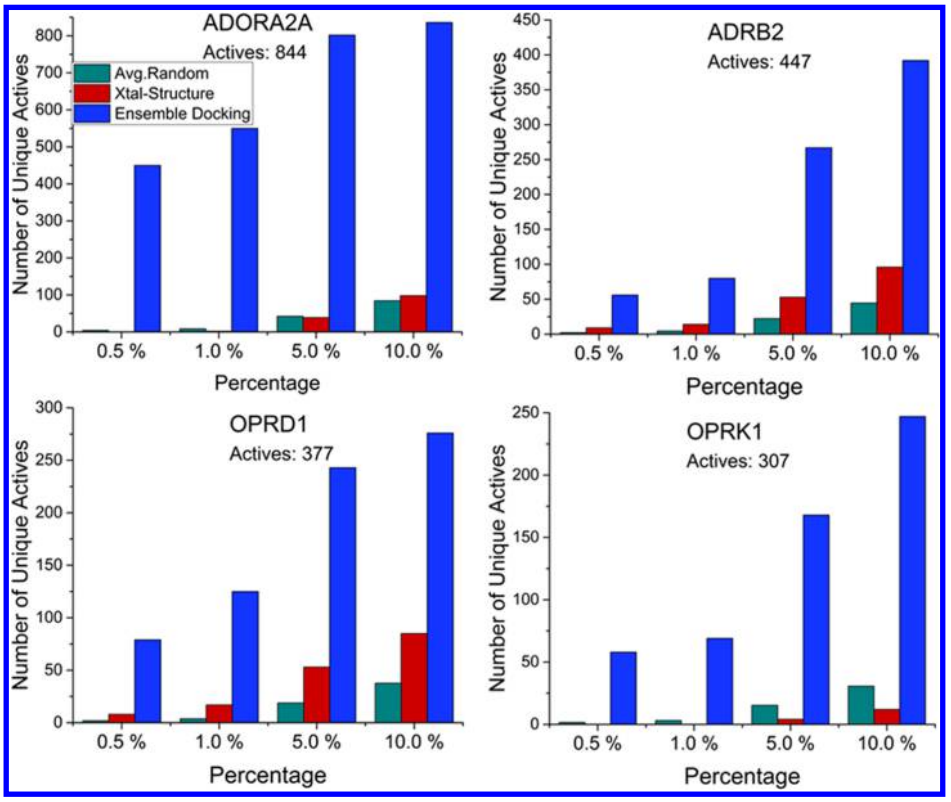


Figure 3. Number of known ligands identified in several subsets of the top-ranking docking score for each of the proteins in statistically significant (better than random) docking using all conformations of the 600 ns MD trajectory.

Table 3. Percentage of Known Ligands Identified by Docking in the Crystal Structures, Clustered Trajectories, and Entire Trajectory; in Different Subsets (Top-Scoring 0.5%; 1, 5, and 10%) of the Total Pool of Compounds Docked[a]

		percentage of actives bound by crystal structure in each subset				percentage of actives bound by cluster in each subset				percentage of actives bound by trajectory in each subset			
protein	total of actives	0.5%	1.0%	5.0%	10.0%	0.5%	1.0%	5.0%	10.0%	0.5%	1.0%	5.0%	10.0%
ADORA2A	844	0.0	0.2	4.6	11.6	5.9	11.7	51.2	82.1	53.3	65.2	95.0	99.1
ADRB2	447	2.0	3.1	11.9	21.5	2.0	2.7	9.4	17.0	12.5	17.9	59.7	87.7
OPRD1	377	2.1	4.5	14.1	22.5	1.9	3.4	14.9	23.3	21.0	33.2	64.5	73.2
OPRK1	307	0.0	0.0	1.3	3.9	2.6	0.0	14.7	33.9	18.9	22.5	54.7	80.5

[a]Only known ligands from statistically significant (better than random) snapshots are shown.

Table 4. Number of Structures Selected by a Number of Known Ligands Statistically Better than a Random Selection of Compounds

	number of significant frames in cluster				number of significant frames in trajectory			
protein	0.5%	1.0%	5.0%	10.0%	0.5%	1.0%	5.0%	10.0%
ADORA2A	3	4	8	16	166	187	508	817
ADRB2	2	1	1	1	20	21	84	128
OPRD1	1	1	1	1	33	34	35	41
OPRK1	2	0	2	3	71	23	40	70

4 shows that only one structure each of from the clustered trajectories for ADRB1 and OPRD1 exhibits a known ligand enrichment that is statistically better than random. Hence, a clustered trajectory approach does not actually correspond to an ensemble docking situation in these ADRB1 and OPRD1 cases. In contrast, in the cases of ADORA2A and OPRK1 (Figure 2), using representative snapshots from a clustered trajectory does very significantly improve the performance of docking, identifying about 7 times (ADORA2A) and 5 times (OPRK1) more known ligands than when docking to a single-crystal structure. Table 4 shows that this increase of performance is obtained from docking in 13 ADORA2 conformations and from docking in 3 OPRK1 conformations. This shows that although ensemble docking is more comprehensive than single-conformation docking in terms of the number of known actives correctly identified, the performance of the calculations in terms of computing time may be still better in the case of a single-conformation docking, as in the case of ADORA2, where a 7-fold improvement in identification of known ligands is obtained through a 13-fold increase in the computational effort.

When using frames taken every 200 ps of a nonclustered 600 ns trajectory, the improvement obtained when using ensemble docking is very significant in all protein cases (Figure 3) compared to the results using a single (crystal) structure. In the cases of ADRB1 and OPRD1, where docking using conformations from clustered MD trajectories did not improve upon docking using a single-crystal structure, docking using all of the conformations of the MD trajectory lead to 4 times (ADRB1) and ~2.75 times (OPRD1) more known ligands

DOI: 10.1021/acs.jpcb.8b11491
J. Phys. Chem. B XXXX, XXX, XXX−XXX

identified than when using a single-crystal structure. In the cases of ADORA2A and OPRK1, these improvements are more pronounced, with 8-times (ADORA2A) and ∼20-times (OPRK1) increase in the number of known ligands identified through docking. About 99% of all known ADORA2A ligands are identified in the top 10% ranked using docking to the nonclustered frames, compared to about 10% of known ligands identified using the crystal structure of the same protein (Table 3, enrichment in the top 10% of the pool of compounds). The worst enrichment obtained using nonclustered MD configurations is still a high 70% in the case of OPRD1 (Table 3, enrichment at the top 10% of the pool of compounds). Again, these statistically very significant improvements in identification of known ligands come at the price of a significant increase of computational time: 99% of known ADORA2A known ligands are identified using 817 conformations of the protein, compared to 10% of known ligands identified in a single (crystal) structure. Although the number of known ligands is strongly increased by nonclustered ensemble docking, the performance/effort ratio is not necessarily favorable to ensemble docking: a 9-times increase in performance is obtained through 817-times increase in computational effort.

The known ligands are sometimes calculated to be capable of binding to several conformations, and not all of the conformations are actually needed to identify all known ligands. Figure 4 shows the cumulative percentage of all known

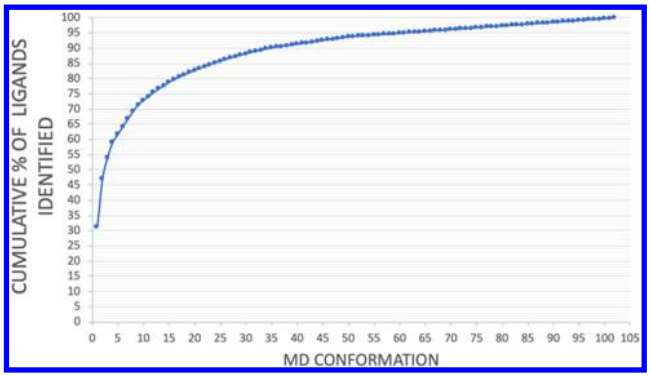


Figure 4. Cumulative percentage of unique ligands identified in the most selected MD conformations for the ADORA2A case.

ligands identified in the nonclustered trajectory of ADORA2A (from Figure 3) as the number of conformations used increases. The first conformation in Figure 4 is that selected by the highest number of known ligands, the second conformation is that identifying the second largest number of additional unique known ligands (i.e., not taking into account ligands that were already identified in the first conformation of Figure 4) etc. Figure 4 shows that in the case of ADORA2A, about 55% of all known ligands are identified in only 3 conformations, and that a total of only 102 conformations (out of the 3000 sampled from the MD trajectory), i.e., 1.7% of the conformations sampled, are needed to identify all unique known ligands of this protein listed in the Zinc database.

Conformational selection is accessible from even relatively short (microsecond time scale) MD trajectories. Notwithstanding the necessity for a large amount of computational time, the performance increase of docking from using nonclustered MD trajectories is significant at the least (ORPD1), or even considerable, such as in the ADORA2A case. Docking using nonclustered MD configurations significantly outperforms docking in clustered MD structures. This indicates that (1) 600 ns MD trajectories generate protein conformations that are selected by 70−99% of the known ligands for these proteins and that (2) clustering of these MD trajectories can miss (in about half the cases in the present work) these conformationally selected protein conformations. Point number 1 is a positive validation of the power and usefulness of molecular dynamics simulations in chemical biology and drug discovery, but point number 2 mitigates this power: the clustering approach that has been needed to limit the computational cost of ensemble docking on non-supercomputing architectures appears to be often missing structures that lead to the optimal identification of known ligands.

CONCLUSIONS

Our results indicate that the protein conformations that are selected by the majority of ligands binding to a protein can be sampled by MD simulations even as short as the sub-microsecond time scale. However, clustering of configurations of such an MD trajectory does not identify the protein structures that are the most selected by its ligands. Clustering of a trajectory could lead to better results than those indicated here, if the clustering focusses on a single binding site. In the case of multiple potential binding site, however, the clustering needs to be performed on a large part of the protein, or on its entirety, and leads to the results shown here. The protein conformers that are selected by the ligands can be identified from a large number of trajectory frames when (as is the case here) known ligands are available to assess the different protein conformations. The current computational cost of such ensemble docking can be very significant, and docking on tens or hundreds of protein structures as described here requires calculations in the millions of hours range on high-performance computers or supercomputers, and this is in addition to the computational power required to enumerate all accessible protein conformations with MD in the first place. It is nonetheless possible to perform such large computations and identify the conformationally selected protein conformations for these proteins for which lists of known ligands are available. As shown in Figure 4, using the 102 conformations of ADORA2A identified as the top selected conformations would identify virtually all known ADORA2A ligands with statistical significance. Such conformationally selected structures, identified in retrospect using known ligands, could potentially be used for the discovery of new drugs and off-target binding predictions with high ligand identification rates, justifying the initial computational investment. However, as of today there are no known metrics that can identify which protein structure will end up being selected by ligands. In other words, even if some of the protein conformations in a MD trajectory may, together, lead to predicting almost all ligands for a given protein, there are no known ways to identify these selected protein structures from the haystack of structures sampled in a MD trajectory for cases where a collection of binding ligands is not already known. There is hence a clear need for an understanding of the physicochemical properties common to conformationally selected protein structures. Our laboratories, and several others, are actively pursuing this goal.

DOI: 10.1021/acs.jpcb.8b11491
J. Phys. Chem. B XXXX, XXX, XXX−XXX

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.jpcb.8b11491.

Thresholds to determine statistical values for for ADORA2A, ADRB2, OPRD1, and OPRK1 in random selection distributions (PDF)

■ AUTHOR INFORMATION

Corresponding Authors

*E-mail: smithjc@ornl.gov (J.C.S.).
*E-mail: jerome.baudry@uah.edu (J.B.).

ORCID

Jeremy C. Smith: 0000-0002-2978-3227
Jerome Baudry: 0000-0002-1969-1679

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

J.B. acknowledges the University of Alabama in Huntsville for support. J.B., J.C.S., and W.E.F. acknowledge support from the University of Tennessee, the LDRD program of ORNL, and SRE, the Cancer Research Informatics Shared Resource Facility of the University of Kentucky Markey Cancer Center (P30CA177558).

■ REFERENCES

(1) Jorgensen, W. L. Efficient Drug Lead Discovery and Optimization. *Acc. Chem. Res.* **2009**, *42*, 724−733.
(2) Amaro, R. E.; Baudry, J.; Chodera, J.; Demir, Ö.; McCammon, J. A.; Miao, Y.; Smith, J. C. Ensemble Docking in Drug Discovery. *Biophys. J.* **2018**, *114*, 2271−2278.
(3) Kapoor, K.; McGill, N.; Peterson, C. B.; Meyers, H. V.; Blackburn, M. N.; Baudry, J. Discovery of Novel Nonactive Site Inhibitors of the Prothrombinase Enzyme Complex. *J. Chem. Inf. Model.* **2016**, *56*, 535−547.
(4) Velazquez, H. A.; Riccardi, D.; Xiao, Z.; Quarles, L. D.; Yates, C. R.; Baudry, J.; Smith, J. C. Ensemble Docking to Difficult Targets in Early-Stage Drug Discovery: Methodology and Application to Fibroblast Growth Factor 23. *Chem. Biol. Drug Des.* **2018**, *91*, 491−504.
(5) Haynes, K. M.; Abdali, N.; Jhawar, V.; Zgurskaya, H. I.; Parks, J. M.; Green, A. T.; Baudry, J.; Rybenkov, V. V.; Smith, J. C.; Walker, J. K. Identification and Structure−Activity Relationships of Novel Compounds That Potentiate the Activities of Antibiotics in *Escherichia coli*. *J. Med. Chem.* **2017**, *60*, 6205−6219.
(6) Evangelista, W.; Weir, R. L.; Ellingson, S. R.; Harris, J. B.; Kapoor, K.; Smith, J. C.; Baudry, J. Ensemble-Based Docking: From Hit Discovery to Metabolism and Toxicity Predictions. *Bioorg. Med. Chem.* **2016**, *24*, 4928−4935.
(7) Zhao, Z.; Martin, C.; Fan, R.; Bourne, P. E.; Xie, L. Drug Repurposing to Target Ebola Virus Replication and Virulence Using Structural Systems Pharmacology. *BMC Bioinf.* **2016**, *17*, 90.
(8) Pi, M.; Kapoor, K.; Ye, R.; Smith, J. C.; Baudry, J.; Quarles, L. D. GPCR6A Is a Molecular Target for the Natural Products Gallate and EGCG in Green Tea. *Mol. Nutr. Food Res.* **2018**, *62*, No. e1700770.
(9) He, H.; Weir, R. L.; Toutounchian, J. J.; Pagadala, J.; Steinle, J. J.; Baudry, J.; Miller, D. D.; Yates, C. R. The Quinic Acid Derivative KZ-41 Prevents Glucose-Induced Caspase-3 Activation in Retinal Endothelial Cells through an IGF-1 Receptor Dependent Mechanism. *PLoS One* **2017**, *12*, No. e0180808.
(10) Abraham, M. J.; Murtola, T.; Schulz, R.; Páll, S.; Smith, J. C.; Hess, B.; Lindah, E. Gromacs: High Performance Molecular Simulations through Multi-Level Parallelism from Laptops to Supercomputers. *SoftwareX* **2015**, *1−2*, 19−25.
(11) Ellingson, S. R.; Smith, J. C.; Baudry, J. VinaMPI: Facilitating Multiple Receptor High-Throughput Virtual Docking on High-Performance Computers. *J. Comput. Chem.* **2013**, *34*, 2212−2221.
(12) Mysinger, M. M.; Carchia, M.; Irwin, J. J.; Shoichet, B. K. Directory of Useful Decoys, Enhanced (DUD-E): Better Ligands and Decoys for Better Benchmarking. *J. Med. Chem.* **2012**, *55*, 6582−6594.
(13) Gatica, E. A.; Cavasotto, C. N. Ligand and Decoy Sets for Docking to G Protein-Coupled Receptors. *J. Chem. Inf. Model.* **2012**, *52*, 1−6.
(14) Fiser, A.; Kinh Gian Do, R.; Sali, A. Modeling Loops in Protein Structures. *Protein Sci.* **2000**, *9*, 1753−1773.
(15) Leventis, P. A.; Grinstein, S. The Distribution and Function of Phosphatidylserine in Cellular Membranes. *Annu. Rev. Biophys.* **2010**, *39*, 407−427.
(16) Ingólfsson, H. I.; Melo, M. N.; Van Eerden, F. J.; Arnarez, C.; Lopez, C. A.; Wassenaar, T. A.; Periole, X.; De Vries, A. H.; Tieleman, D. P.; Marrink, S. J. Lipid Organization of the Plasma Membrane. *J. Am. Chem. Soc.* **2014**, *136*, 14554−14559.
(17) Monticelli, L.; Kandasamy, S. K.; Periole, X.; Larson, R. G.; Tieleman, D. P.; Marrink, S. J. The MARTINI Coarse Grained Force Field: Extension to Proteins. *J. Chem. Theory Comput.* **2008**, *4*, 819−834.
(18) Periole, X.; Marrink, S.-J. The Martini Coarse-Grained Force Field. In *Biomolecular Simulations: Methods and Protocols*; Monticelli, L.; Salonen, E., Eds.; Humana Press: Totowa, NJ, 2013; Vol. *924*, pp 533−565.
(19) Wassenaar, T. A.; Ingólfsson, H. I.; Böckmann, R. A.; Tieleman, D. P.; Marrink, S. J. Computational Lipidomics with Insane: A Versatile Tool for Generating Custom Membranes for Molecular Simulations. *J. Chem. Theory Comput.* **2015**, *11*, 2144−2155.
(20) Bond, P. J.; Sansom, M. S. P. Insertion and Assembly of Membrane Proteins via Simulation. *J. Am. Chem. Soc.* **2006**, *128*, 2697−2704.
(21) Scott, K. A.; Bond, P. J.; Ivetac, A.; Chetwynd, A. P.; Khalid, S.; Sansom, M. S. P. Coarse-Grained MD Simulations of Membrane Protein-Bilayer Self-Assembly. *Structure* **2008**, *16*, 621−630.
(22) Marrink, S. J.; Tieleman, D. P. Perspective on the Martini Model. *Chem. Soc. Rev.* **2013**, *42*, 6801−6822.
(23) Berendsen, H. J. C.; van der Spoel, D.; van Drunen, R. GROMACS: A Message-Passing Parallel Molecular Dynamics Implementation. *Comput. Phys. Commun.* **1995**, *91*, 43−56.
(24) Stansfeld, P. J.; Goose, J. E.; Caffrey, M.; Carpenter, E. P.; Parker, J. L.; Newstead, S.; Sansom, M. S. P. MemProtMD: Automated Insertion of Membrane Protein Structures into Explicit Lipid Membranes. *Structure* **2015**, *23*, 1350−1361.
(25) Daura, X.; Gademann, K.; Jaun, B.; Seebach, D.; Van Gunsteren, W. F.; Mark, A. E. Peptide Folding: When Simulation Meets Experiment. *Angew. Chem., Int. Ed* **1999**, *38*, 236−240.
(26) Wassenaar, T. A.; Pluhackova, K.; Böckmann, R. A.; Marrink, S. J.; Tieleman, D. P. Going Backward: A Flexible Geometric Approach to Reverse Transformation from Coarse Grained to Atomistic Models. *J. Chem. Theory Comput.* **2014**, *10*, 676−690.
(27) Sanner, M. F. Python: A Programming Language for Software Integration and Development. *J. Mol. Graphics Modell.* **1999**, *17*, 57−61.
(28) Case, D. A.; Babin, V.; Berryman, J. T.; Betz, R. M.; Cai, Q.; Cerutti, D. S.; Cheatham, T. E., III; Darden, T. A.; Duke, R. E.; Gohlke, H. et al. *The Amber Molecular Dynamics Package*; University of California: San Francisco, CA, 2014.
(29) Morris, G. M.; Huey, R.; Lindstrom, W.; Sanner, M. F.; Belew, R. K.; Goodsell, D. S.; Olson, A. J. Autodoc4 and AutoDocktools4: Automated Docking with Selective Receptor Flexibility. *J. Comput. Chem.* **2009**, *30*, 2785−2791.
(30) Salgado, C. M.; Azevedo, C.; Proença, H.; Vieira, S. M. Noise Versus Outliers. In *Secondary Analysis of Electronic Health Records*; MIT Critical Data, Springer: Cham, 2016; pp 163−183.
(31) Tukey, J. W. *Exploratory Data Analysis*; Addison-Wesley: Reading, MA, 1977.

DOI: 10.1021/acs.jpcb.8b11491
J. Phys. Chem. B XXXX, XXX, XXX−XXX