JOURNAL OF CHEMICAL INFORMATION AND MODELING

Article

pubs.acs.org/jcim

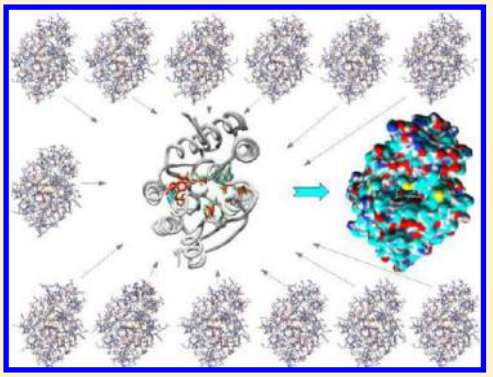
Multiple Structures for Virtual Ligand Screening: Defining Binding Site Properties-Based Criteria to Optimize the Selection of the Query

Nesrine Ben Nasr,[†] Hélène Guillemain,[†] Nathalie Lagarde,[†] Jean-François Zagury, and Matthieu Montes*

Laboratoire Génomique Bioinformatique et Applications, Équipe d'accueil EA 4627, Conservatoire National des Arts et Métiers, 292 rue Saint Martin, 75003 Paris, France

Supporting Information

ABSTRACT: Structure based virtual ligand screening (SBVLS) methods are widely used in drug discovery programs. When several structures of the target are available, protocols based either on single structure docking or on ensemble docking can be used. The performance of the methods depends on the structure(s) used as a reference, whose choice requires retrospective enrichment studies on benchmarking databases which consume additional resources. In the present study, we have identified several trends in the properties of the binding sites of the structures that led to the optimal performance in retrospective SBVLS tests whatever the docking program used (Surflex-dock or ICM). By assessing their hydrophobicity and comparing their volume and opening, we show that the selection of optimal structures should be possible with no requirement of prior retrospective enrichment studies. If the mean binding site volume is lower than 350 A^3, the structure with the smaller volume should be preferred. In the other cases, the structure with the largest binding site should be preferred. These optimal structures may be either selected for a single structure docking strategy or an ensemble docking strategy. When constructing an ensemble, the opening of the site might be an interesting criterion additionaly to its volume as the most closed structures should not be preferred in the large systems. These "binding site properties-based" guidelines could be helpful to optimize future prospective drug discovery protocols when several structures of the target are available.



INTRODUCTION

In silico screening of compound collections has been extensively used to reduce the number of compounds going into high throughput screening procedures.[1] Different strategies can be used, depending on the data available. If the structure of the target is known, a structure-based virtual ligand screening (SBVLS) protocol that generally includes docking methods can be applied.

Conformational changes occur in protein binding sites upon ligand binding. Besides the classical rigid single structure docking methods, several approaches have been developed in order to handle the plasticity of the binding site during docking like induced-fit docking[2] or Monte Carlo simulations.[3] Another way to handle the flexibility of the target is to perform a rigid receptor docking experiment on multiple structures, i.e. ensemble docking. An ensemble can comprise experimental structures[4−7] or theoretical structures derived from simulation studies.[8−11] Compared with the classical single structure docking that consider the binding site as rigid, these approaches are time-consuming and can sometimes increase the rate of false positive results due to scoring errors and/or nonoptimal choice of the structures comprising the ensembles.[4,5,12−14] Despite the previously described advances, most of the available docking methods handle receptor flexibility implicitly by softening the contact repulsive term in the scoring function (i.e., soft docking[13]) or by using Gaussian potentials for contacts.[15]

With the recent developments of high throughput crystallography, several structures of a given target are available in the Protein Data Bank (PDB).[16] Thus, for a prospective virtual screening, two options are possible: (1) choose the most appropriate structure in the PDB or (2) use an ensemble of structures altogether. Although many studies have already been performed on this matter, there is currently no consensus favoring one approach over the other.[5,17,18] Different research groups have concluded that an ensemble should comprise a relatively small number of structures to be optimal (up to five depending on the studies).[4,18−20] Recently, Bottegoni et al.[21] performed an ensemble docking study on 36 targets from the DUD data set with ICM. They concluded that ensemble docking outperforms average single structure runs and suggested that prior to a prospective SBVLS, preliminary docking studies should be performed to identify the best structures. Similar conclusions arose from very recent studies of Korb et al.[22] and Rueda et al.[23] Hence, without preliminary virtual screening studies, it seems, at the moment, very challenging to identify the optimal structure(s) for a "real-life" SBVLS experiment.

Received: September 24, 2012
Published: January 14, 2013

   dx.doi.org/10.1021/ci3004557 | J. Chem. Inf. Model. 2013, 53, 293−311

As shown in the literature,[19,24−26] the choice of the starting conformation for docking studies is critical. We thus decided to search for "binding site properties-based" guidelines to optimize the selection of the conformations that will compose an ensemble when several structures of the target are available. To design our study, we used the targets of the DUD data set for which several experimental structures were available in the PDB (all but PDGFRb). After a careful inspection of the properties of the binding sites of all the conformations available for a given target, we selected up to four "extreme" experimental structures for which the binding site properties in terms of volume and opening were significantly different. For each structure, we evaluated the performance in terms of enrichment of two different SBVLS methods, Surflex-dock (SF) and ICM using the corresponding DUD-own data sets in order to (1) confirm the impact of the choice of the structure on the results of the screening; (2) identify potential trends within the structural properties shared by the structure(s) that resulted into the best enrichments; and (3) reduce the possible bias in the trends that could be observed by using a single docking method. We also assessed the performance of SF and ICM on all the possible ensembles composed by the structures selected for each system in order to optimize the selection of the structures composing the best ensembles and compare their structural properties.

In the present study, we describe our attempt to develop "low-cost" binding site properties-based criteria to identify the optimal structure(s) of the target of a structure-based drug discovery program.

MATERIAL AND METHODS

Directory of Useful Decoys (DUD) Data Set. DUD is a public benchmarking data set designed for docking methods evaluation containing known active compounds for 40 targets, including 36 decoys for each active compound. We selected for each target its corresponding DUD-own data set that comprises only its associated active compounds and decoys. A large random-drug-like data set seeded with known active compounds for each target could have been used to be closer to a prospective VLS, but we decided to use the DUD data set as it is the current standard for fair results comparison between different groups.[27−29] For our study, the DUD release 2 data set was obtained from the Web site http://dud.docking.org.

Selection and Preparation of the Protein Structures. We selected for this study, the 39 targets issued from the DUD for which at least two experimental structures were available; the PDGFR-β target was thus excluded. For each target, all the wild-type experimental holo structures available in the Protein Data Bank (PDB) were downloaded; their binding sites were superimposed and visually inspected. After a careful examination of their corresponding binding site properties, we selected up to four "extreme" structures in order to cover the experimentally available flexibility of the binding sites in terms of volume and opening (i.e.: most open, most closed, smallest volume, largest volume). The area of the opening of the binding sites was assessed using CASTp.[30] The volume of the binding sites was computed with POVME.[31] POVME allows an accurate manual edition of the probes selected for the calculation of the volume. Thus, we edited the points manually to cover, for a given target, only the binding site and not its surroundings (Supporting Information Figure S2). The hydrophobicity of the binding site was defined as the ratio of its apolar solvent accessible surface area (ASASA) over its solvent accessible surface area (SASA). ASASA and SASA were computed with GetArea.[32] Hydrogen atoms were added using Chimera.[33]

Computational Methods. *Surflex-dock.* SF is based on a modified Hammerhead fragmentation/reconstruction algorithm to dock compounds flexibly into the binding site.[34] The query molecule is decomposed into rigid fragments that are superimposed to the Surflex-protomol i.e. molecular fragments covering the entire binding site. The docking poses are evaluated by an empirical scoring function. For each structure, the binding site has been defined at 4 Å around the cocrystallized ligand for the protomol generation step. In this study, Surflex-dock version 2.5 has been used for all calculations with the options +premin, +remin.

ICM. ICM is based on Monte Carlo simulations in internal coordinates to optimize the position of molecules using a stochastic global optimization procedure combined with pseudo-Brownian positional/torsional steps and fast local gradient minimization.[35] The docking poses were evaluated using ICM-VLS empirical scoring function.[36] The binding site defined for docking has been adjusted to be similar to the Surflex protomol. In this study, ICM version 3.6 has been used for all calculations.

Ensemble Docking. For each target, all possible ensembles (up to eleven for a given system) have been constructed from the experimental structures selected in the PDB. Ensemble docking experiments were performed with Surflex-dock and ICM. For a given compound on a given target, the best score obtained within the different structures composing the ensemble has been retained. Thus, for each ensemble, the ensemble docking result consists in a ranked list of the best scoring poses.

Performance Metrics. All graphics were produced with the statistical and graphical tool R (http://www.r-project.org/). The ROCR package was used to plot rate of change (ROC) curves and the Wilcoxon−Mann−Whitney algorithm was used for the ROC area under curce (AUC) calculations.[37] Enrichment factors were computed as follows:

$$\mathrm{EF}_{x\%} = \frac{N^{x\%}_{\mathrm{experimental}}}{N^{x\%}_{\mathrm{actives}}}$$

$N^{x\,\%}_{\mathrm{experimental}}$: number of active compounds retrieved in the top x % of the sorted database. $N^{x\%}_{\mathrm{actives}}$: total number of active compounds in the compound collection.

When comparing the impact of the choice of the structure on enrichment, in the case of similar AUCs, EF1% that reflects early enrichment was considered. Raw tendencies have been compared to an estimation of the random case (Rd) computed as follows:

$$\mathrm{Rd} = \sum \frac{t_n}{n^k}$$

t_n: number of targets for which n conformations are compared. n: number of conformations selected for the study for a given target. k: number of docking softwares used (SF and/or ICM).

RESULTS

Presentation of the 39 Targets—Selection of the Different Experimental Conformations. As shown in the literature,[19,24−26] the choice of the starting conformation for docking studies is of major importance. For each of the 39 targets explored in this study, we examined the different experimental structures available in the PDB (from 2 to 300

Table 1. Physicochemical Properties of the Binding Sites of the Structures Selected for the Study for the (a) Small (i.e., Mean Binding Site Volume under 350 A^3) and "Less-Hydrophobic" (i.e., Mean Percent Hydrophobicity Lower than 70%), (b) Small and "More-Hydrophobic" (i.e., Mean Percent Hydrophobicity over 70%), (c) Large (i.e., Mean Binding Site Volume over 350 A^3) and Less-Hydrophobic, and (d) Large and More-Hydrophobic DUD Targets[a]

	mean volume ($Å^3$)	mean percent hydrophobicity	no. DUD actives	no. available PDB structures	PDB ID	resolution (Å)	RMSD	volume ($Å^3$)	opening ($Å^2$)	SASA ($Å^2$)	ASASA ($Å^2$)	percent hydrophobicity
					(a) small and less-hydrophobic							
NA	243	45.0	49	5	1A4G	2.20		227	12.1	810.94	356.16	43.9
					1A4Q	1.90	0.471	258	118.3	854.53	394.47	46.2
HSP90	337	60.7	37	62	2CDD	1.90		228	33.7	595.76	326.53	54.8
					3K99	2.10	3.279	233	92.4	565.48	353	62.4
					1UYF	2.00	2.299	549	70.3	775.09	503.33	64.9
PNP	159	60.7	50	7	1B8O	1.50		98	0	398.64	211.62	53.1
					2QPL	2.10	0.579	148	52.8	824.75	513.52	62.3
					1V48	2.20	0.734	230	27.3	886.51	591.93	66.8
TRP	138	63.6	49	300	1BTY	1.50		60	2.9	441.9	272.75	61.7
					3AAU	1.80	0.196	171	12.1	773.09	484.91	62.7
					1V2O	1.62	0.951	183	8.9	810.34	538.76	66.5
VEGFR2	322	63.8	88	3	2XIR	1.50		221	119.1	561.89	387.11	68.9
					1VR2	2.40	2.864	422	253.5	658.96	386.68	58.7
EGFR	279	64.5	475	10	2J5E	3.10		258	140.7	590	381.75	64.7
					2ITX	2.98	0.794	278	299.2	598.66	390.75	65.3
					2RGP	2.00	2.333	282	74.3	466.08	314.1	67.4
					3POZ	1.50	2.404	299	176.3	462.56	281.29	60.8
SAHH	259	66.3	33	2	1A7A	2.80		177	0	522.19	329.13	63.0
					1LI4	2.01	0.349	341	189.2	363.05	252.56	69.6
TK	146	67.4	22	4	1E2P	2.50		114	0	275.68	200.36	72.7
					1E2I	1.90	7.991	128	86.1	467.39	306.21	65.5
					1E2N	2.20	2.568	195	0	510.22	326.11	63.9
					(b) small and more-hydrophobic							
COX2	264	71.0	426	6	3PGH	2.50		248	102.3	757.73	523.27	69.1
					1CX2	3.00	0.832	264	375.6	802.35	562.03	70.1
					1DDX	3.00	1.221	280	371.3	893.57	660.19	73.9
GART	320	71.0	40	6	1C2T	2.10		269	54.6	797.69	561.75	70.4
					1GAR	1.96	2.561	293	39.5	1610.82	1163.58	72.2
					1JKX	1.60	2.524	399	126.4	1095.99	771.79	70.4
PR	302	72.7	27	12	1A28	1.80		207	0	395.43	328.24	83.0
					2OVH	2.00	5.861	396	130.6	793.22	495.17	62.4
COMT	171	73.5	11	12	2CL5	1.60		84	20.5	1046.81	755.33	72.2
					1H1D	2.00	3.456	157	10.7	1028.2	749.55	72.9
					3A7D	2.40	0.905	272	27.0	1132.68	854.69	75.5
AR	280	74.5	79	48	1T7T	1.70		240	0	376.4	284.97	75.7
					3B66	1.65	1.006	320	0	404.38	296.45	73.3
GR	315	75.3	78	7	1M2Z	2.50		228	0.65	387.61	299.76	77.3
					3CLD	2.84	1.325	262	0	413.39	323.81	78.3
					1NHZ	2.30	1.524	317	647.7	703.95	494.52	70.3
					3K22	2.10	1.059	452	5.7	493.64	371.79	75.3
MR	173	76.5	15	3	2A3I	1.95		172	0	622.61	472.22	75.9
					2AA2	1.95	0.798	174	0	713.29	549.85	77.1
COX1	243	77.8	25	22	1PTH	3.40		200	336.9	347.91	269.87	77.6
					3KK6	2.75	5.736	256	370.7	431.26	334.67	77.6
					1PGG	4.50	0.527	272	327.2	400.29	312.88	78.2
HIVRT	182	80.1	43	52	2JLE	2.90		133	1742.4	547.7	410.53	74.9
					1DTQ	2.80	1.937	177	1613.3	549	478.49	87.2
					3LAM	2.76	1.751	236	1910.6	774.32	604.6	78.1
RXR	253	80.1	20	17	1FBY	2.25		228	0	649.47	509.52	78.5
					3DZY	3.10	1.207	278	0	649.62	531.41	81.8
ER ago	259	84.9	67	10	3ERD	2.03		245	0	291.99	248.15	85.0
					2P15	1.94	0.972	272	0	458.98	389.22	84.8
INHA	331	86.4	86	25	2X22	2.10		147	143.2	751.32	653.26	86.9
					1P44	2.60	3.275	515	91.4	895.1	769.33	85.9
					(c) large and less-hydrophobic							
AMPC	664	58.6	21	44	1FSW	1.90		592	92.8	748.23	397.21	53.1
					2HDR	2.20	0.527	672	238.8	866.63	527.39	60.9

dx.doi.org/10.1021/ci3004557 | J. Chem. Inf. Model. 2013, 53, 293−311

Table 1. continued

	mean volume (Å^3)	mean percent hydrophobicity	no. DUD actives	no. available PDB structures	PDB ID	resolution (Å)	RMSD	volume (Å^3)	opening (Å^2)	SASA (Å^2)	ASASA (Å^2)	percent hydrophobicity
					(c) large and less-hydrophobic							
					1XGJ	1.97	2.722	728	226.9	929.97	573.72	61.7
FGFR1	632	59.7	120	4	1FGI	2.40		355	145.7	809.6	462.71	57.2
					2FGI	2.50	0.511	909	213.1	1104.25	685.99	62.1
HIVPR	513	59.9	62	87	1NH0	1.03		451	37.4	1003.31	612.23	61.0
					2PWC	1.78	0.761	465	22.6	880.66	523.98	59.5
					1XL2	1.50	0.322	642	50.7	932.73	552.85	59.3
FXA	471	60.0	146	100	2RA0	2.30		415	10.8	797.61	460.25	57.7
					2D1J	2.20	0.829	422	33.8	827.86	492.57	59.5
					1IQF	3.20	1.054	577	7.0	879.35	551.15	62.7
ACE	595	60.4	49	4	3BKL	2.18		448	86.4	657.23	407.95	62.1
					2OC2	2.25	0.323	664	81.9	697.22	415.67	59.6
					1O86	2.00	0.275	672	135.5	684.48	407.95	59.6
SRC	515	60.6	159	4	2H8H	2.20		278	64.5	842.3	523.48	62.1
					1Y57	1.91	2.017	698	771.2	864.17	510.5	59.1
GPB	366	61.3	52	52	6GPB	2.86		174	24.0	269.17	181.18	67.3
					1A8I	1.78	1.281	293	0	335.56	198.76	59.2
					1K06	1.80	1.772	630	96.3	414	237.28	57.3
HMGR	928.5	67.0	35	8	1DQA	2.00		737	62.7	842.41	576.33	68.4
					1HWJ	2.26	0.507	1120	139.9	1157.2	757.43	65.5
ACHE	516	68.7	107	75	2C5F	2.60		378	16.4	569.06	392.2	68.9
					2CEK	2.20	0.685	558	32.1	717.6	493.22	68.7
					1OCE	2.70	0.711	612	29.8	627.45	430.46	68.6
THR	608	68.7	72	4	1DOJ	1.70		568	0.1	696.97	470.88	67.6
					1JWT	2.50	13.717	648	25.51	782.92	547.52	69.9
PARP	508	69.7	35	6	1EFY	2.20		506	90.2	640.37	450.63	70.4
					4PAX	2.80	0.637	509	31.3	664.49	459.34	69.1
					(d) large and more-hydrophobic							
CDK2	456	70.0	72	143	3PY1	2.05		304	62.2	675.73	497.63	73.6
					2I40	2.8	2.6	504	131.3	999.26	635.47	63.6
					2B55	1.85	1.792	560	81.5	972.25	708.62	72.9
P38	367	71.1	454	96	3BX5	2.4		185	7.5	542.64	373.81	68.9
					3FSF	2.1	1.221	442	388.7	848.18	581.67	68.6
					1KV2	2.8	2.559	473	297.7	974.64	737.97	75.7
ADA	507	72.3	39	12	1NDW	2		360	15.2	544.4	382.81	70.3
					1NDV	2.3	1.357	575	20.9	770.89	570.77	74.0
					2E1W	2.5	1.424	585	20.0	796.15	578.72	72.7
PPAR	474	74.4	85	83	3OSW	2.55		332	57.0	873.94	639.04	73.1
					1WM0	2.9	1.272	380	1.5	716.24	515.46	72.0
					1FM9	2.1	1.698	710	74.6	899.83	702.78	78.1
DHFR	723	74.7	410	7	3DFR	1.7		542	16.2	696.07	544.38	78.2
					1BZF	NMR_19[b]	1.185	904	111.6	761.21	541.65	71.2
PDE5	530	75.3	88	3	1TBF	1.3		519	55.2	1039.01	789.05	75.9
					1XOZ	1.37	0.999	540	30.6	1045.72	780.25	74.6
ALR2	356	76.2	26	5	1AH3	2.3		272	16.7	624.68	493.17	78.9
					1EK0	1.48	0.618	440	34.3	616.48	452.66	73.4
ER antago	421	76.6	39	18	3ERT	1.9		358	81.8	794.89	599.04	75.4
					2IOG	1.6	2.025	483	4.5	785.79	611.62	77.8

[a]Binding site RMSD has been computed using ICM on all heavy atoms towards the coordinates of the smallest conformer. Binding site volumes and opening were computed using respectively POVME and CASTp. ASASA and SASA were calculated with GetArea. Percent hydrophobicity is the ratio of ASASA over SASA. [b]For 1BZF, the conformation selected for the study is the 19th model from the NMR structure.

depending on the system). We selected up to four structures (2 for 18 targets, 3 for 19 targets, and 4 for 2 targets) in order to cover the experimentally available flexibility of the binding sites in terms of volume and opening. For a given structure, the volume of the binding site was calculated with POVME and its opening was computed with CASTp. The mean volume value among the conformers selected for the study, 350 Å^3, was used as a cutoff to classify the targets in two equilibrated groups of 20 small targets and 19 large targets. Similarly, we split the 39 targets in two equilibrated groups of 20 more-hydrophobic targets harboring a cutoff of more than 70% of the solvent accessible surface area of the binding site being apolar, and of 19 less-hydrophobic targets. For four small and more-hydrophobic targets (AR, ER ago, MR, and RXR), the opening of the binding site was very limited and did not vary significantly between the structures available in the PDB. They were considered as closed and invariable in terms of binding site opening. The binding sites properties (volume,

dx.doi.org/10.1021/ci3004557 | J. Chem. Inf. Model. 2013, 53, 293−311

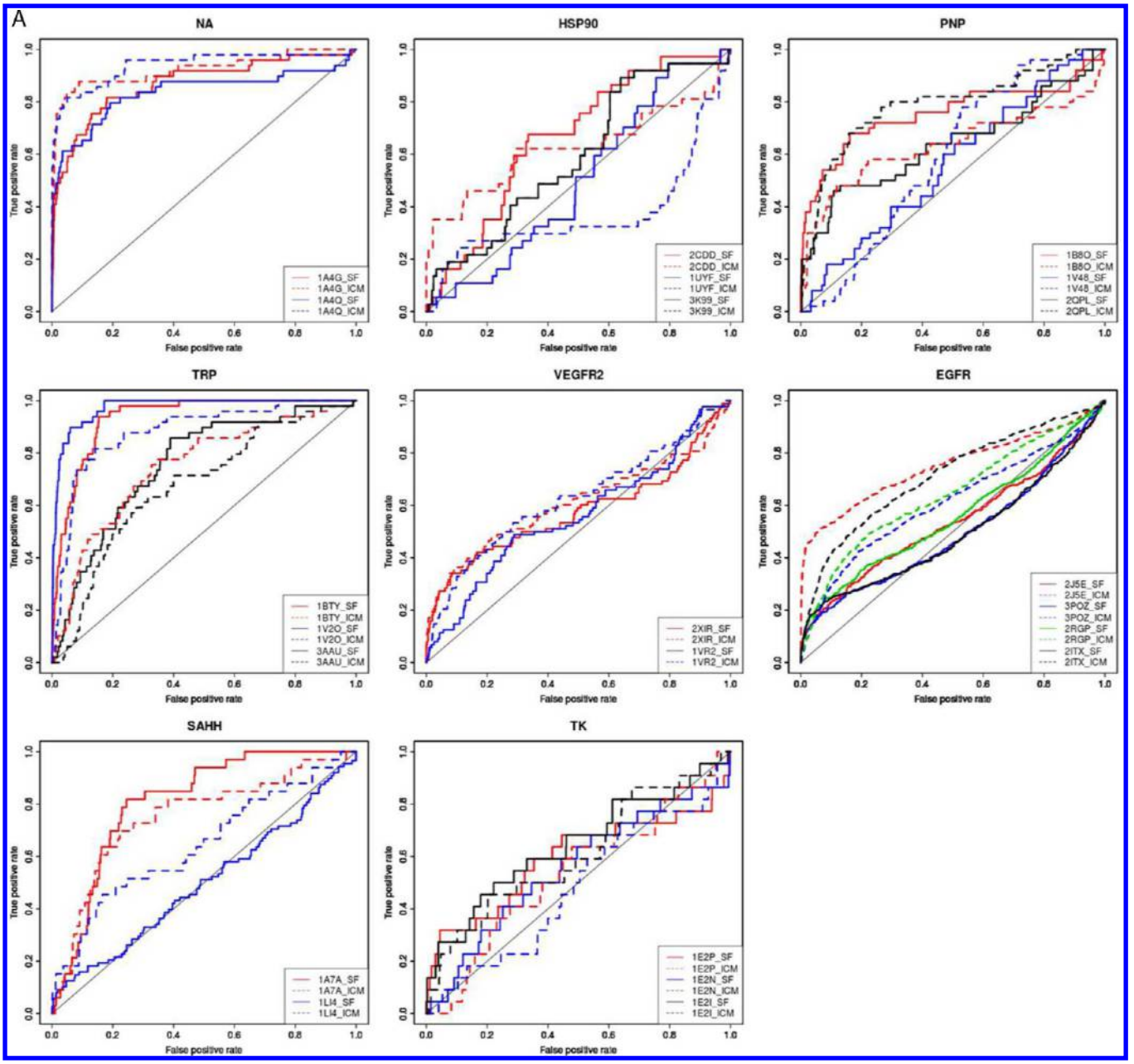


Figure 1. continued

dx.doi.org/10.1021/ci3004557 | J. Chem. Inf. Model. 2013, 53, 293−311

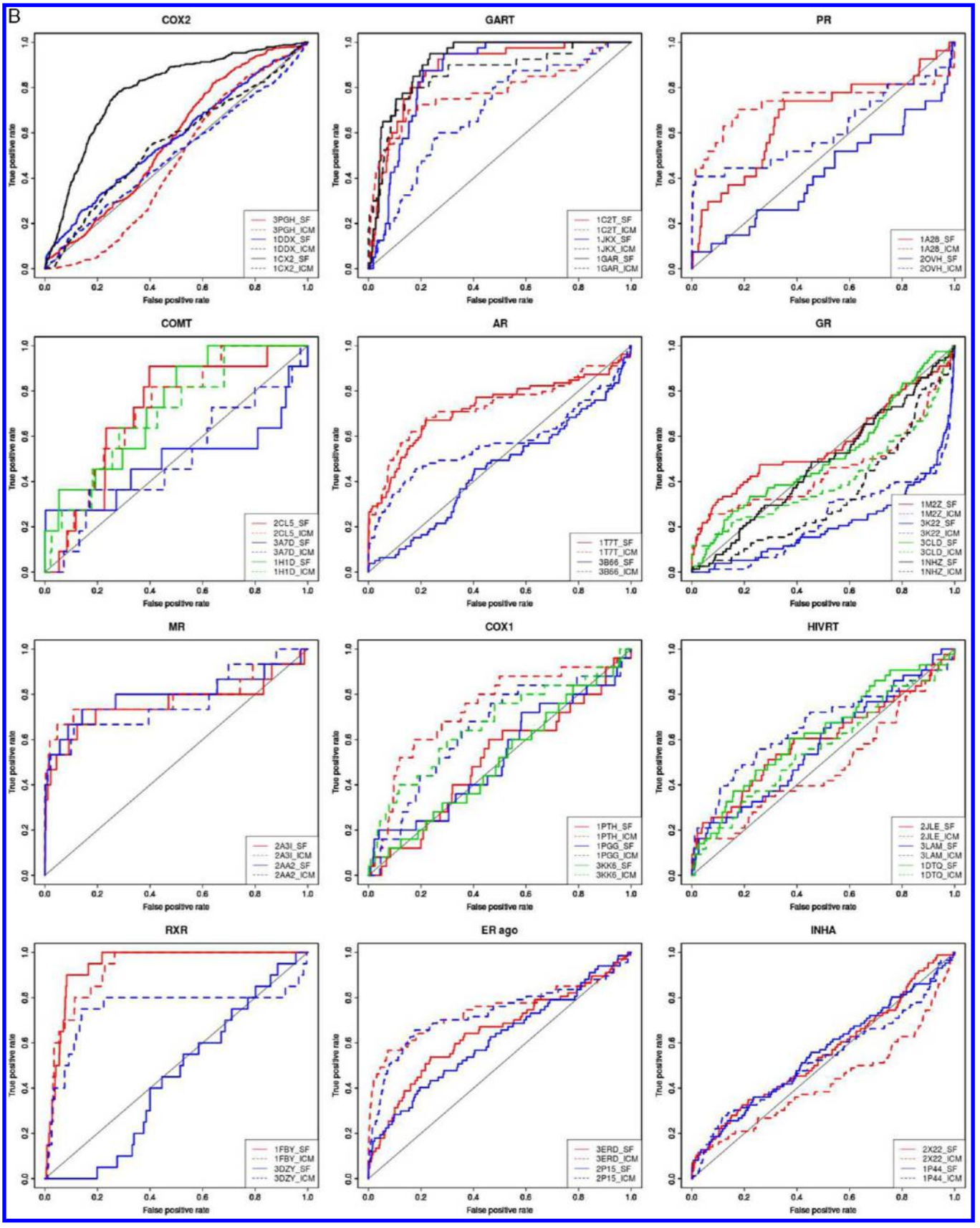


Figure 1. continued

dx.doi.org/10.1021/ci3004557 | J. Chem. Inf. Model. 2013, 53, 293−311

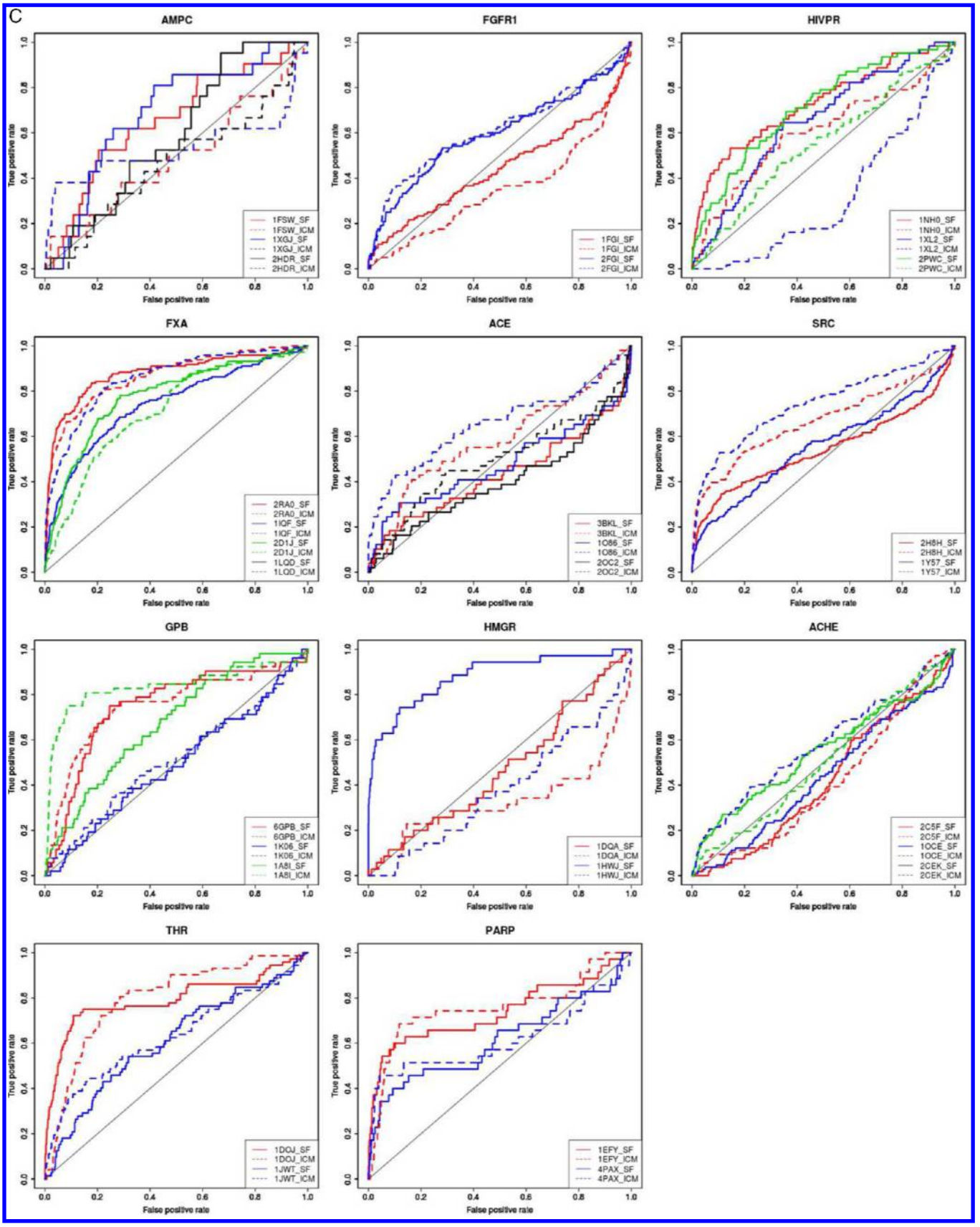


Figure 1. continued

dx.doi.org/10.1021/ci3004557 | J. Chem. Inf. Model. 2013, 53, 293−311

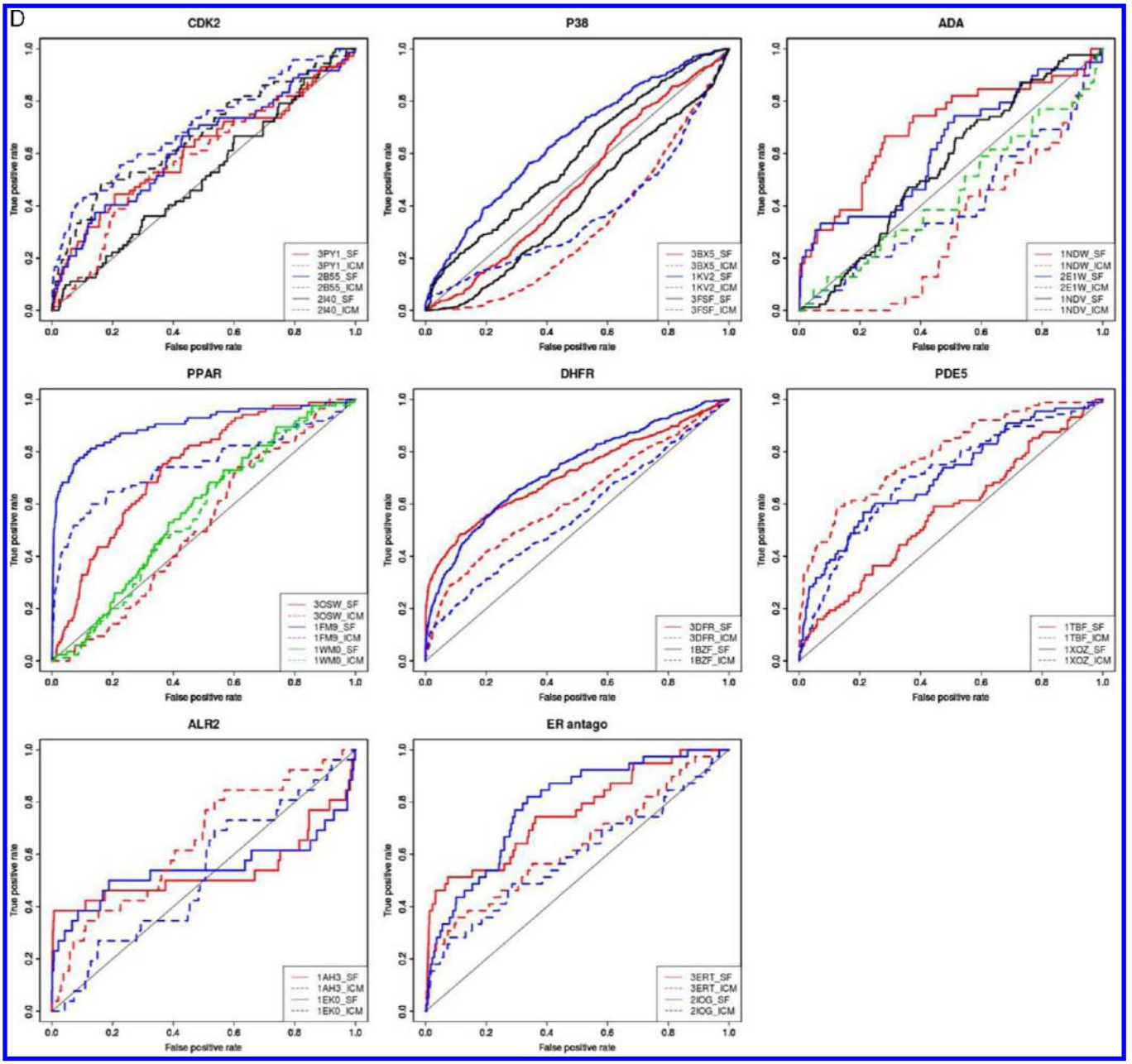


Figure 1. ROC curves with single structure docking protocols on the (A) small and less-hydrophobic, (B) small and more-hydrophobic, (C) large and less-hydrophobic, and (D) large and more-hydrophobic DUD systems using SF (plain lines) and ICM (dotted lines). For a given target, the ROC curves associated to the "extreme" structures in terms of binding site properties are represented in red for the smallest binding site, blue for the largest, green for the most closed and black for the most open. When the structure was considered as extreme in two categories, the volume of the binding site has been priorized for the color of the ROC curve.

opening, and hydrophobicity) of the structures selected for the 39 DUD targets and their corresponding DUD-own sets are summarized in Table 1.

Single Structure Docking Strategy. The first part of our study was to identify possible trends in the structural properties of the conformations used as a reference for docking that gave the best results in terms of enrichment. We analyzed the performance of two different docking methods, SF and ICM, on all the conformers selected for this study taken separately and compared the structural properties of the structures that led to the best enrichments. The results are presented using ROC curves in Figure 1 and ROC AUC and enrichment factors in Table 2

Overall, for 20 out of the 39 systems used in this study, the conformer that resulted into the best performances was common to SF and ICM (Rd = 6.74/39). Within these 20 systems, in 6 out of the 10 systems for which the pocket volume was inferior to 350 A^3 (small systems) the optimal conformer was the smallest (Rd = 3.5/10). This trend was confirmed within the seven small and more-hydrophobic systems, for which the conformer displaying the best performances was the smallest in five cases (Rd = 3.2/7).

In the systems for which the pocket volume was superior to 350 A^3 (large systems), the conformer that gave the best performances in enrichment was the largest for 7 systems out of 10 (Rd = 4.2/10).

dx.doi.org/10.1021/ci3004557 | J. Chem. Inf. Model. 2013, 53, 293−311

Table 2. Volume, Binding Site Opening, Area under the ROC Curve (AUC), Enrichment Factors at 1% and 10%, and Average Calculation Time Per Compound (CPU) Using Single Structure Docking Protocols with Surflex-dock (SF) or ICM with the Different Structures Selected for the Study on the (a) Small and Less-Hydrophobic, (b) Small and More-Hydrophobic, (c) Large and Less-Hydrophobic, and (d) Large and More-Hydrophobic Data Sets

	SF					ICM					
	PDB ID	AUC	EF1%	EF10%	CPU	PDB ID	AUC	EF1%	EF10%	CPU	volume	opening
(a) small and less-hydrophobic systems												
NA	1A4G	**0.869**	**24.79**	**6.75**	**11.60**	1A4G	0.923	26.85	8.58	17.80	227	12.1
	1A4Q	0.836	30.98	6.34	13.79	1A4Q	**0.938**	**39.24**	**8.18**	**22.73**	258	118.3
HSP90	2CDD	**0.665**	**0.00**	**1.63**	**14.53**	2CDD	0.627	19.22	3.53	14.77	228	33.7
	3K99	0.593	2.75	1.90	17.66	3K99	0.544	13.73	2.99	16.10	233	92.4
	1UYF	0.498	2.75	1.09	13.76	1UYF	**0.631**	**0.00**	**1.90**	**14.69**	549	70.3
PNP	1B8O	**0.755**	**15.20**	**5.43**	**7.60**	1B8O	0.635	8.69	4.02	9.90	98	0
	2QPL	0.638	15.20	3.02	5.64	2QPL	**0.788**	**10.86**	**5.23**	**7.19**	148	52.8
	1V48	0.561	0.00	1.81	7.07	1V48	0.590	0.00	0.40	7.16	230	27.3
TRP	1BTY	0.933	16.45	7.16	27.92	1BTY	0.747	2.06	3.68	47.07	60	2.9
	3AAU	0.755	2.06	3.27	25.81	3AAU	0.675	0.00	2.04	45.25	171	12.1
	1V2O	**0.973**	**26.73**	**9.00**	**26.76**	1V2O	**0.882**	**4.11**	**7.16**	**50.71**	183	8.9
VEGFR2	2XIR	**0.576**	**11.73**	**3.41**	**18.08**	2XIR	0.609	14.08	3.53	15.76	221	119.1
	1VR2	0.560	4.69	1.48	10.00	1VR2	**0.625**	**3.52**	**2.84**	**17.40**	422	253.5
EGFR	2J5E	0.536	8.03	2.23	9.70	2J5E	**0.750**	**26.01**	**5.22**	**11.60**	258	140.7
	2ITX	0.493	9.30	2.44	10.02	2ITX	0.713	7.40	4.00	12.25	278	299.2
	2RGP	**0.558**	**8.46**	**2.48**	**11.88**	2RGP	0.641	5.50	3.14	10.93	282	74.3
	3POZ	0.494	8.88	2.08	11.88	3POZ	0.609	4.02	2.82	13.12	299	176.3
SAHH	1A7A	**0.811**	**0.00**	**3.05**	**8.90**	1A7A	**0.751**	**0.00**	**3.97**	**11.30**	177	0
	1LI4	0.497	5.87	1.59	13.00	1LI4	0.643	9.64	3.05	10.00	341	189.2
TK	1E2P	0.593	9.22	3.19	8.87	1E2P	0.534	0.00	0.46	6.52	114	0
	1E2I	**0.643**	**13.83**	**2.74**	**8.28**	1E2I	**0.607**	**4.61**	**2.74**	**6.80**	128	86.1
	1E2N	0.555	0.00	1.37	7.49	1E2N	0.478	0.00	0.91	6.42	195	0
(b) small and more-hydrophobic systems												
COX2	3PGH	0.573	3.77	1.15	11.56	3PGH	0.530	0.00	0.16	12.58	248	102.3
	1CX2	**0.785**	**3.76**	**3.50**	**12.20**	1CX2	**0.560**	**0.24**	**1.29**	**11.16**	264	375.6
	1DDX	0.574	5.87	1.71	11.80	1DDX	0.500	0.71	1.41	12.21	280	371.3
GART	1C2T	0.880	2.55	5.55	25.60	1C2T	0.780	10.21	5.30	40.15	269	54.6
	1GAR	**0.916**	**2.55**	**6.56**	**21.87**	1GAR	**0.861**	**17.87**	**6.06**	**43.12**	293	39.5
	1JKX	0.861	0.00	4.29	23.31	1JKX	0.692	2.55	2.02	46.91	399	126.4
PR	1A28	**0.667**	**7.91**	**2.99**	**8.80**	1A28	**0.738**	**31.64**	**5.22**	**8.80**	207	0
	2OVH	0.419	7.91	1.12	8.31	2OVH	0.609	27.69	4.10	9.73	396	130.6
COMT	2CL5	0.716	0.00	1.85	4.26	2CL5	**0.712**	**0.00**	**1.85**	**6.33**	84	20.5
	1H1D	**0.733**	**21.77**	**3.71**	**4.50**	1H1D	0.700	0.00	2.78	5.99	157	10.7
	3A7D	0.491	32.66	2.78	6.39	3A7D	0.501	0.00	0.93	5.84	272	27.0
AR	1T7T	**0.726**	**26.88**	**4.18**	**1.43**	1T7T	**0.731**	**21.76**	**4.94**	**7.98**	240	0
	3B66	0.452	3.84	0.63	1.92	3B66	0.572	12.80	3.42	7.93	320	0
GR	1M2Z	**0.564**	**9.05**	**3.08**	**12.60**	1M2Z	0.450	10.34	2.57	11.70	228	0.65
	3CLD	0.511	2.59	1.67	13.51	3CLD	**0.591**	**11.63**	**1.67**	**12.96**	262	0
	1NHZ	0.478	1.29	0.77	11.30	1NHZ	0.340	2.59	0.51	11.72	317	647.7
	3K22	0.209	0.00	0.39	14.10	3K22	0.230	0.00	0.13	11.49	452	5.7
MR	2A3I	0.757	28.93	5.34	12.40	2A3I	**0.788**	**36.17**	**6.68**	**12.90**	172	0
	2AA2	**0.796**	**36.17**	**6.00**	**13.90**	2AA2	0.767	43.40	5.34	13.10	174	0
COX1	1PTH	0.507	0.00	1.21	8.72	1PTH	**0.734**	**0.00**	**3.22**	**8.08**	200	336.9
	3KK6	0.503	4.16	1.21	8.10	3KK6	0.644	0.00	1.61	9.43	256	370.7
	1PGG	**0.521**	**4.16**	**2.01**	**8.53**	1PGG	0.670	4.16	3.22	8.03	272	327.2
HIVRT	2JLE	0.593	9.69	2.56	11.22	2JLE	0.515	9.69	1.63	12.45	133	1742.4
	1DTQ	**0.628**	**7.27**	**1.86**	**10.95**	1DTQ	0.559	0.00	1.86	12.20	177	1613.3
	3LAM	0.578	2.42	2.33	10.45	3LAM	**0.657**	**7.27**	**3.03**	**11.62**	236	1910.6
RXR	1FBY	**0.941**	**5.50**	**8.50**	**26.49**	1FBY	**0.922**	**5.50**	**7.00**	**27.00**	228	0
	3DZY	0.444	0.00	0.00	16.29	3DZY	0.750	0.00	5.00	27.28	278	0
ER ago	3ERD	**0.646**	**7.57**	**2.69**	**8.10**	3ERD	**0.746**	**22.71**	**5.69**	**8.51**	245	0
	2P15	0.613	9.08	2.54	7.00	2P15	0.730	12.11	5.24	7.36	272	0
INHA	2X22	**0.550**	**4.72**	**1.86**	**13.84**	2X22	**0.575**	**8.27**	**1.51**	**11.77**	515	143.2
	1P44	0.550	4.07	2.57	10.90	1P44	0.525	1.16	2.33	12.80	533	91.4

dx.doi.org/10.1021/ci3004557 | J. Chem. Inf. Model. 2013, 53, 293−311

Table 2. continued

		SF					ICM					
	PDB ID	AUC	EF1%	EF10%	CPU	PDB ID	AUC	EF1%	EF10%	CPU	volume	opening
(c) large and less-hydrophobic systems												
AMPC	1FSW	0.656	0.00	1.92	6.60	1FSW	0.490	4.80	1.44	6.74	592	92.8
	2HDR	0.593	0.00	0.96	5.72	2HDR	0.530	0.00	0.48	7.22	672	238.8
	1XGJ	**0.688**	**0.00**	**1.44**	**5.70**	1XGJ	**0.530**	**14.41**	**3.84**	**6.62**	728	226.9
FGFR1	1FGI	0.444	5.08	1.58	19.45	1FGI	0.400	2.54	1.33	23.08	355	145.7
	2FGI	**0.598**	**6.77**	**2.67**	**17.96**	2FGI	**0.614**	**3.38**	**3.58**	**26.16**	909	213.1
HIVPR	1NH0	**0.728**	**9.68**	**4.35**	**35.14**	1NH0	0.602	3.23	2.26	55.71	451	37.4
	2PWC	0.713	10.16	2.92	34.09	2PWC	0.550	1.61	1.45	55.21	465	22.6
	1XL2	0.649	3.23	1.61	35.51	1XL2	**0.685**	**0.00**	**0.00**	**58.95**	642	50.7
FXA	2RA0	**0.877**	**18.09**	**6.99**	**21.87**	2RA0	**0.865**	**14.61**	**6.65**	**42.04**	415	10.8
	2D1J	0.778	8.35	4.25	22.04	2D1J	0.714	5.57	2.47	37.80	422	33.8
	1IQF	0.743	9.74	4.18	22.14	1IQF	0.855	22.96	5.41	37.79	577	7.0
ACE	3BKL	0.434	4.19	1.84	15.00	3BKL	0.580	2.09	1.64	18.98	448	86.4
	2OC2	0.402	2.09	1.43	13.39	2OC2	0.502	4.19	1.64	18.84	664	81.9
	1O86	**0.476**	**2.09**	**2.66**	**12.90**	1O86	**0.660**	**16.74**	**4.09**	**17.23**	672	135.5
SRC	2H8H	0.542	8.91	3.15	21.21	2H8H	0.668	14.64	3.97	19.14	278	64.5
	1Y57	**0.555**	**8.91**	**2.33**	**15.94**	1Y57	**0.742**	**12.10**	**4.72**	**19.47**	698	771.2
GPB	6GPB	**0.747**	**2.00**	**3.27**	**11.00**	6GPB	0.752	4.01	4.43	17.56	174	24.0
	1A8I	0.674	4.01	2.12	9.60	1A8I	**0.835**	**8.03**	**7.51**	**13.64**	293	0
	1K06	0.492	0.00	1.15	9.60	1K06	0.513	2.01	1.35	14.44	630	96.3
HMGR	1DQA	0.470	0.00	1.15	25.11	1DQA	0.332	2.89	0.86	36.76	737	62.7
	1HWJ	**0.880**	**34.63**	**6.31**	**20.40**	1HWJ	**0.381**	**0.00**	**0.29**	**35.99**	1120	337.9
ACHE	2C5F	0.424	0.00	0.37	15.66	2C5F	**0.582**	**0.00**	**0.94**	**20.38**	378	16.4
	2CEK	**0.549**	**3.83**	**2.15**	**16.68**	2CEK	0.505	0.96	1.22	21.76	558	32.1
	1OCE	0.436	0.96	0.47	14.96	1OCE	0.581	3.83	2.15	22.85	612	29.8
THR	1DOJ	**0.793**	**15.45**	**6.27**	**31.28**	1DOJ	**0.797**	**1.40**	**4.18**	**41.65**	568	0.1
	1JWT	0.611	1.40	1.95	28.81	1JWT	0.631	7.02	3.62	47.44	648	25.51
PARP	1EFY	**0.738**	**18.28**	**5.74**	**4.70**	1EFY	**0.756**	**3.05**	**5.74**	**5.10**	506	90.2
	4PAX	0.615	9.14	3.73	4.80	4PAX	0.615	12.18	4.59	5.82	509	31.3
(d) large and more-hydophobic systems												
CDK2	3PY1	0.608	2.84	2.65	15.38	3PY1	0.583	1.42	1.25	14.54	304	62.2
	2I40	0.515	0.00	1.11	11.66	2I40	0.669	5.68	3.34	14.48	504	131.3
	2B55	**0.626**	**5.68**	**2.37**	**11.52**	2B55	**0.705**	**14.19**	**4.18**	**14.52**	560	81.5
P38	3BX5	0.488	0.67	0.60	13.75	3BX5	0.318	0.00	0.04	11.38	185	7.5
	3FSF	0.585	1.78	1.90	12.71	3FSF	**0.414**	**0.00**	**0.13**	**11.00**	442	388.7
	1KV2	**0.650**	**5.12**	**2.20**	**11.00**	1KV2	0.367	4.45	0.95	10.80	473	297.7
ADA	1NDW	**0.700**	**16.51**	**3.10**	**8.80**	1NDW	0.320	0.00	0.00	10.77	360	15.2
	1NDV	0.580	5.50	1.81	8.07	1NDV	0.549	0.00	1.29	11.24	575	20.9
	2E1W	0.624	13.76	3.35	7.70	2E1W	**0.595**	**0.00**	**0.77**	**11.05**	585	20.0
PPAR	3OSW	0.737	1.18	3.06	37.98	3OSW	0.516	0.00	0.59	54.50	332	57.0
	1WM0	0.579	0.00	0.71	41.73	1WM0	0.556	1.18	0.59	57.12	380	1.5
	1FM9	**0.901**	**27.16**	**7.65**	**37.90**	1FM9	**0.748**	**16.53**	**5.18**	**51.10**	710	74.6
DHFR	3DFR	0.716	18.45	4.32	12.50	3DFR	**0.620**	**4.19**	**2.94**	**15.36**	542	16.2
	1BZF	**0.731**	**11.32**	**3.59**	**11.01**	1BZF	0.552	0.49	0.71	15.82	904	111.6
PDE5	1TBF	0.565	4.70	1.82	20.56	1TBF	**0.791**	**17.61**	**4.56**	**28.87**	519	55.2
	1XOZ	**0.707**	**5.87**	**3.53**	**23.06**	1XOZ	0.705	4.70	2.62	26.07	540	30.6
ALR2	1AH3	0.537	27.49	3.85	5.20	1AH3	**0.650**	**3.93**	**2.69**	**6.00**	272	16.7
	1EK0	**0.545**	**19.63**	**3.85**	**4.88**	1EK0	0.525	0.00	0.77	6.47	440	34.3
ER antago	3ERT	0.757	16.34	5.15	25.50	3ERT	**0.630**	**13.62**	**3.09**	**32.00**	358	81.8
	2IOG	**0.785**	**5.45**	**3.86**	**25.62**	2IOG	0.591	5.45	2.83	34.53	483	4.5

Overall, in the small systems, the conformer that gave the worst enrichment was the one displaying the largest binding site for 15 out of 20 targets with SF (Rd = 8/20) and for 14 out of 20 targets with ICM (Rd = 8/20).

There was a noticeable difference in the performance of SF and ICM depending on the opening of the conformation chosen for docking. Using SF, the strongest signal was obtained with small targets for which the most closed conformation resulted in the best performance (10 out of 16; Rd = 6/16) whereas the most open conformation resulted in the worst performance (10 out of 16; Rd = 6/16). Using ICM, the most closed conformations of the large targets were associated with the worst performances (11 out of 19; Rd = 7.75/19).

Ensemble Docking Strategy. Ensemble docking can be an alternative when different structures are available. Thus, for each target, we constructed all possible ensembles with the

dx.doi.org/10.1021/ci3004557 | J. Chem. Inf. Model. 2013, 53, 293−311

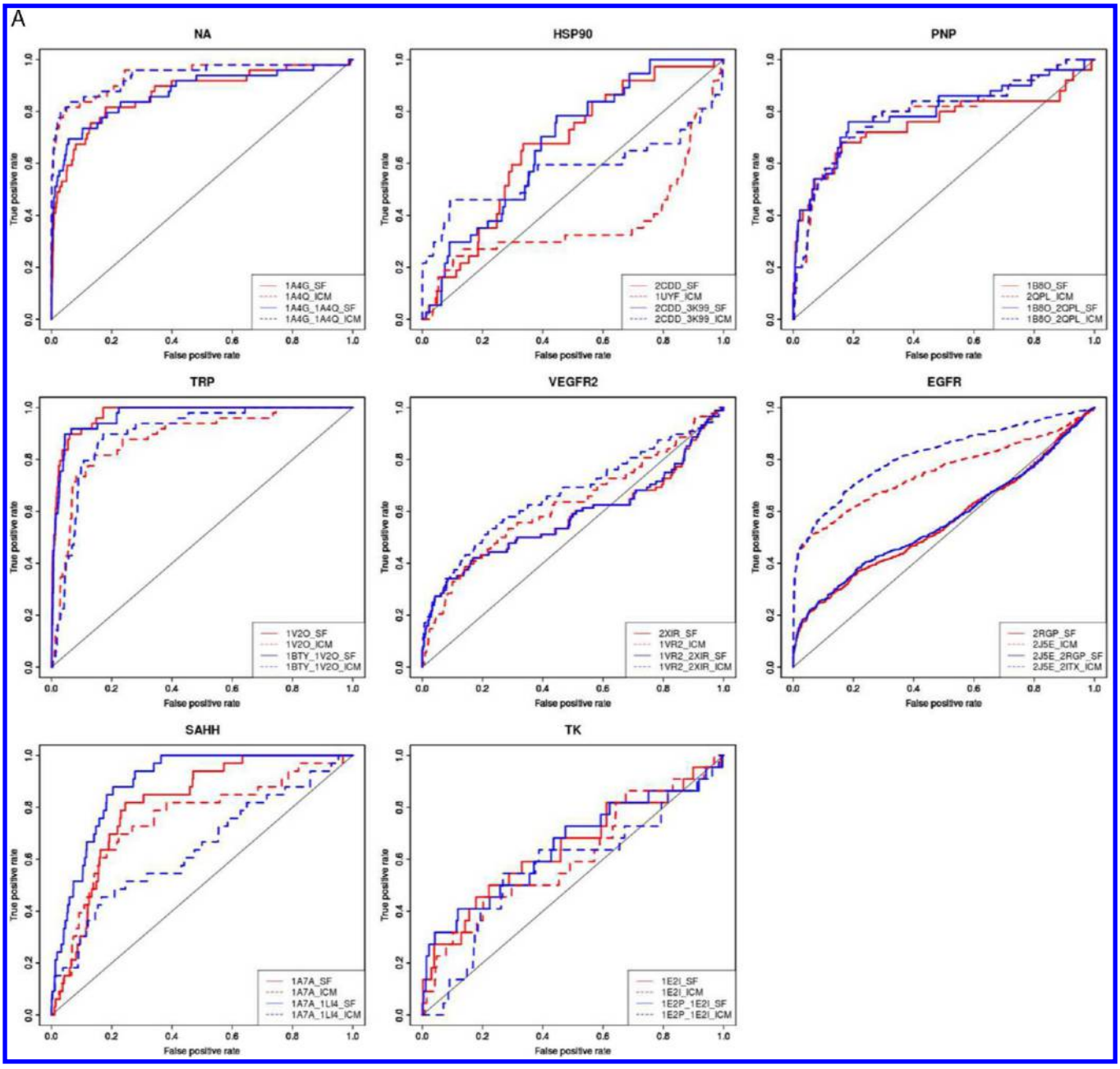


Figure 2. continued

dx.doi.org/10.1021/ci3004557 | *J. Chem. Inf. Model.* 2013, 53, 293−311

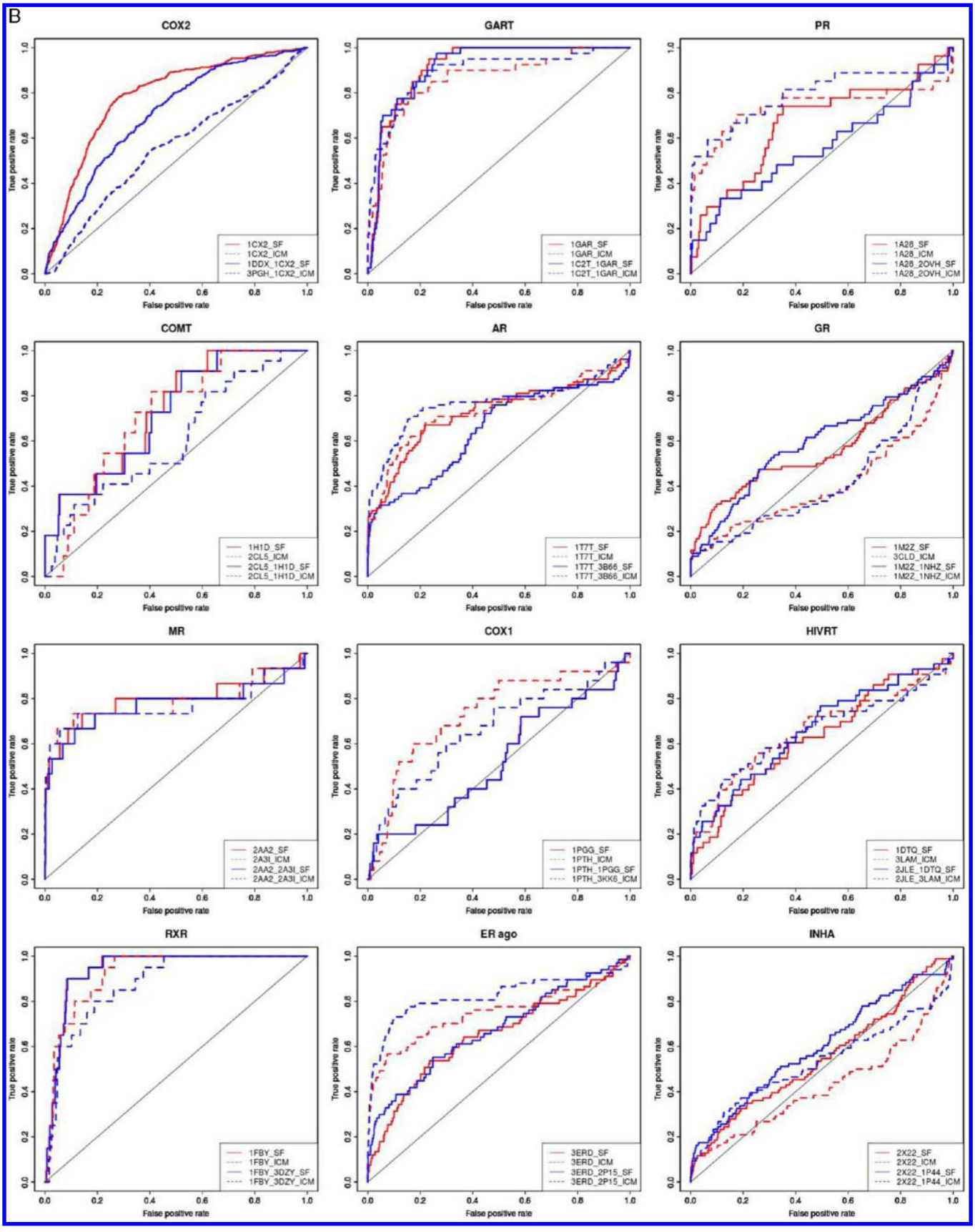


Figure 2. continued

dx.doi.org/10.1021/ci3004557 | J. Chem. Inf. Model. 2013, 53, 293−311

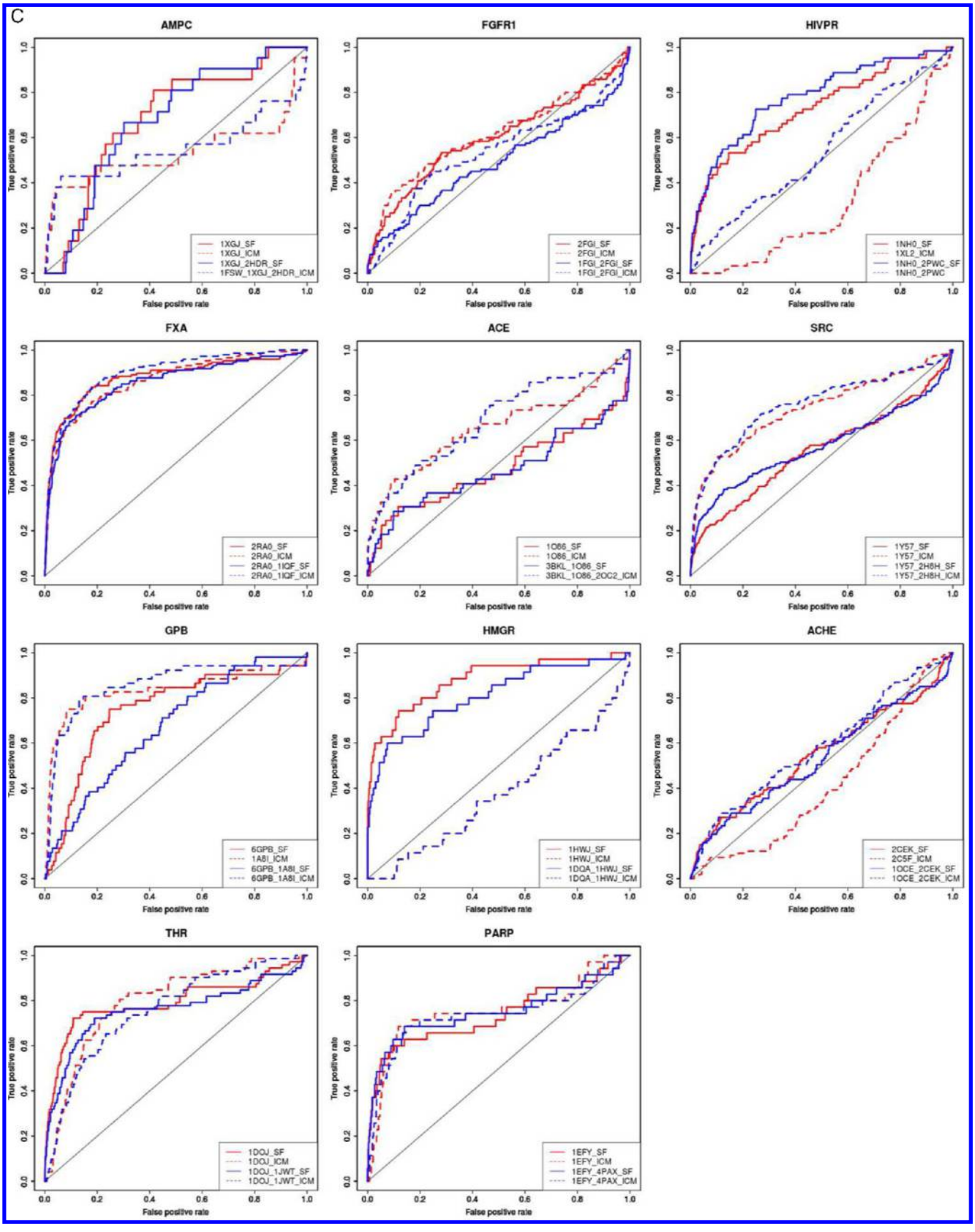


Figure 2. continued

dx.doi.org/10.1021/ci3004557 | J. Chem. Inf. Model. 2013, 53, 293−311

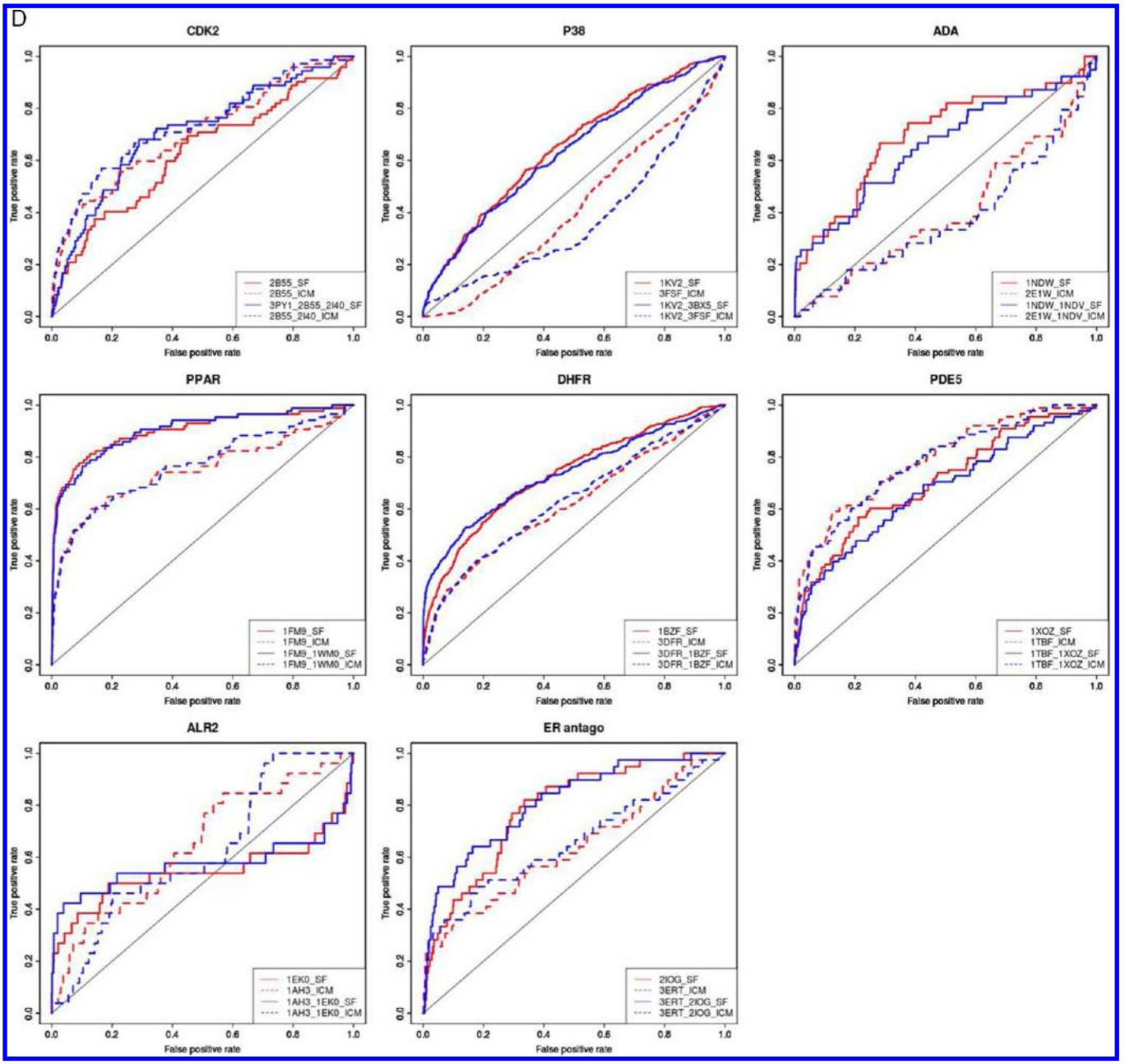


Figure 2. ROC curves with the best single docking (red) and the best ensemble (blue) docking protocols on the (A) small and less-hydrophobic, (B) small and more-hydrophobic, (C) large and less-hydrophobic, and (D) large and more-hydrophobic DUD systems with SF (plain-lines) and ICM (dotted lines).

structures selected for single structure docking. For a given compound on a given target, the best score obtained within the different structures composing the ensemble has been retained. Thus, for each ensemble, the ensemble docking result consists in a ranked list of the best scoring poses. Best ensemble docking ROC curves are presented in Figure 2. Ensemble docking enrichment factors and ROC AUC are presented in supplementary online material in Supporting Information Table S1.

In order to analyze the impact of the choice of the structures composing the ensembles, we excluded the 18 targets for which there was only 1 possible ensemble.

For respectively 20 with SF and 19 with ICM out of the 21 remaining targets, the highest performance was obtained with ensembles comprising only 2 conformations (Rd = 15.3/21). For respectively 21 with SF and 17 with ICM out of the 21 targets, the ensemble that led to the best performances comprised the structure associated with the best performance by single structure docking (Rd = 14.4/21). In respectively 17 with SF and 16 with ICM out of 21 targets, this best ensemble encompassed the conformer displaying the smallest binding site (Rd = 14.4/21). Concerning the worst performances, for respectively 21 with SF and 20 with ICM out of the 21 targets, the structure associated with the worst performance by single structure docking is part of the ensemble providing the worst performances (Rd = 14.4/21). In 17 out of 21 targets with SF, the conformer displaying the largest binding site is always in the worst ensemble (Rd = 14.4/21). Using ICM, no signal was obtained when focusing on the ensemble associated with the worst performance.

dx.doi.org/10.1021/ci3004557 | J. Chem. Inf. Model. 2013, 53, 293−311

For the large systems, the structure with the most closed binding site was in five out of nine cases not present in the best ensemble with SF and ICM (Rd = 2.5/9). For 6 out of the 11 small systems, the most open structure was not part of the best ensemble using SF (Rd = 3/11).

Comparison of Ensemble Docking and Single Docking Strategies. Using SF and ICM, single docking protocol outperformed the ensemble one for 18 targets whereas the ensemble docking performances are better than the single ones for 19 out of 39 targets. For two targets, no noticeable difference was noted between single and ensemble docking results (Figure 2)

DISCUSSION

Single Structure Docking Strategy. In the present work, we intended to identify possible trends in the structural properties of the conformations used as a reference for docking that led to the best enrichments in active compounds. The best conformation was the same with SF and ICM in 19 out of 39 targets and the worst for 21 out of 39 targets (Rd = 6.74/39). This observation confirms that the conformation of the query impacts the quality of the enrichment and is in agreement with what was described in the literature.[19,25,26] It also shows that the performance of different programs can be dependent on common trends regarding the properties of the structure used as a reference. We thus looked for these common properties shared by the optimal conformations. To our knowledge, no binding site properties-based guidelines have been proposed to identify the optimal conformation within a set of available structures. We thus examined the physicochemical properties of the binding sites of the structures that led to the best enrichments.

For the small systems, the conformers that provided the best enrichments, either using SF or ICM, were in majority the conformers presenting the smallest binding site volume. Interestingly, this trend is strengthened when considering the conformers associated to the worst performances as they were at the opposite the ones that presented the largest binding sites. For the large systems, the conformers that provided the best enrichments were in majority the conformers presenting the largest binding site volume.

No strong trends could be observed when trying to link the quality of the performances to the opening of the binding site using both SF and ICM, even if slight tendencies could be observed with SF. Indeed, we observed that the most closed structures were associated with better performances compared to the most open structures.

Overall, it seemed that the volume of the binding site could be a critical criterion to select the optimal structure when different structures are available for a given target for a docking protocol.

Ensemble Docking. When different structures of the target are available, ensemble docking strategies are feasible, even with limited computational resources. We thus built with the experimental structures selected for each system, all the possible ensembles to conduct ensemble docking experiments using SF and ICM.

The first observation was that, in agreement with the literature,[4,18,20] ensembles limited in size and in particular dual structure ensembles led to better performance than larger sets and thus seemed more appropriate for ensemble docking studies. Indeed with SF and ICM, respectively 20 and 19 out of the 21 best ensembles comprise 2 structures (Rd = 15.3). However to our knowledge, no binding site properties-based guidelines have ever been defined to identify these optimal structures that should be part of the optimal ensemble within a set of known and available structures. An extensive work has been performed by Yoon et al.,[17] Bottegoni et al.,[21] and Rueda[18,23] to remove the structures that should not be comprised in ensembles. They performed systematic preliminary docking and enrichment studies on all the structures available and removed from the ensembles the structures that performed poorly in enrichment with the benchmarking data sets. Despite giving excellent enrichment in known active compounds using the resulting ensembles (even better than each single structure approach), this strategy is very time-consuming as it requires a huge amount of preliminary calculation and analysis before performing the desired virtual screening on the "real-life" target. That is why the aim of our study was to identify "low-cost" guidelines to optimize this selection step and conduct the optimal SBVLS strategy on the best structure(s). In the 21 targets for which there was more than 1 possible ensemble, the best ensembles that we obtained comprised most often the structure that gave the best performance in a single docking strategy, using SF (20/21) and almost always when using ICM (19/21). This explains why the results obtained with ensemble docking protocol are directly correlated with those of single docking strategy. Indeed, both using SF (17/21) and ICM (16/21), the best ensemble comprised the conformer displaying the smallest binding site while; when using SF, the conformers presenting the largest binding site were present in the ensemble associated with the worst performances (17/21). The opening of the binding site could also be an interesting criterion for selecting the structure that should be part of the optimal ensemble as we observed that, in the large systems, the structure with the most closed binding site was in five out of nine cases not present in the best ensemble with SF and ICM (Rd = 2.5/9). For 6 out of the 11 small systems, the most open structure was not part of the best ensemble using SF (Rd = 3/11). This latter result was not replicated using ICM and may be due to the better treatment of solvation in the ICM scoring function compared to SF's scoring function. Indeed, the opening of the binding site could impact its solvent accessibility.

Comparison of Ensemble Docking and Single Docking Strategies. Different opinions are expressed in the literature about the strategy to use when different structures of the target are available. Using prior SBVLS studies on all the structures available to define optimized ensembles, Yoon et al.,[17] Bottegoni et al.,[21] and Rueda et al.[18,23] outperformed single structure approaches. However, our results, similarly to numerous other studies,[4,5,18,38] underline that the best ensemble docking strategies that we used do not systematically outperform the best single structure docking approach (19 systems out of 39; see Figure 2). Since the computational cost of ensemble docking approaches increases linearly with the number of structures composing the ensemble, we estimated, at least for groups that have access to limited computational resources, that ensemble docking approaches should not be systematically preferred as they do not always perform better than single structure approaches.

Caveats within the Benchmarking Database. In addition to the caveats that were recently observed by Schneider et al.[39] in the DUD data set, we identified several points that should be taken into account in future versions of the DUD and that have not been corrected in the very recently released DUD-E.[40]

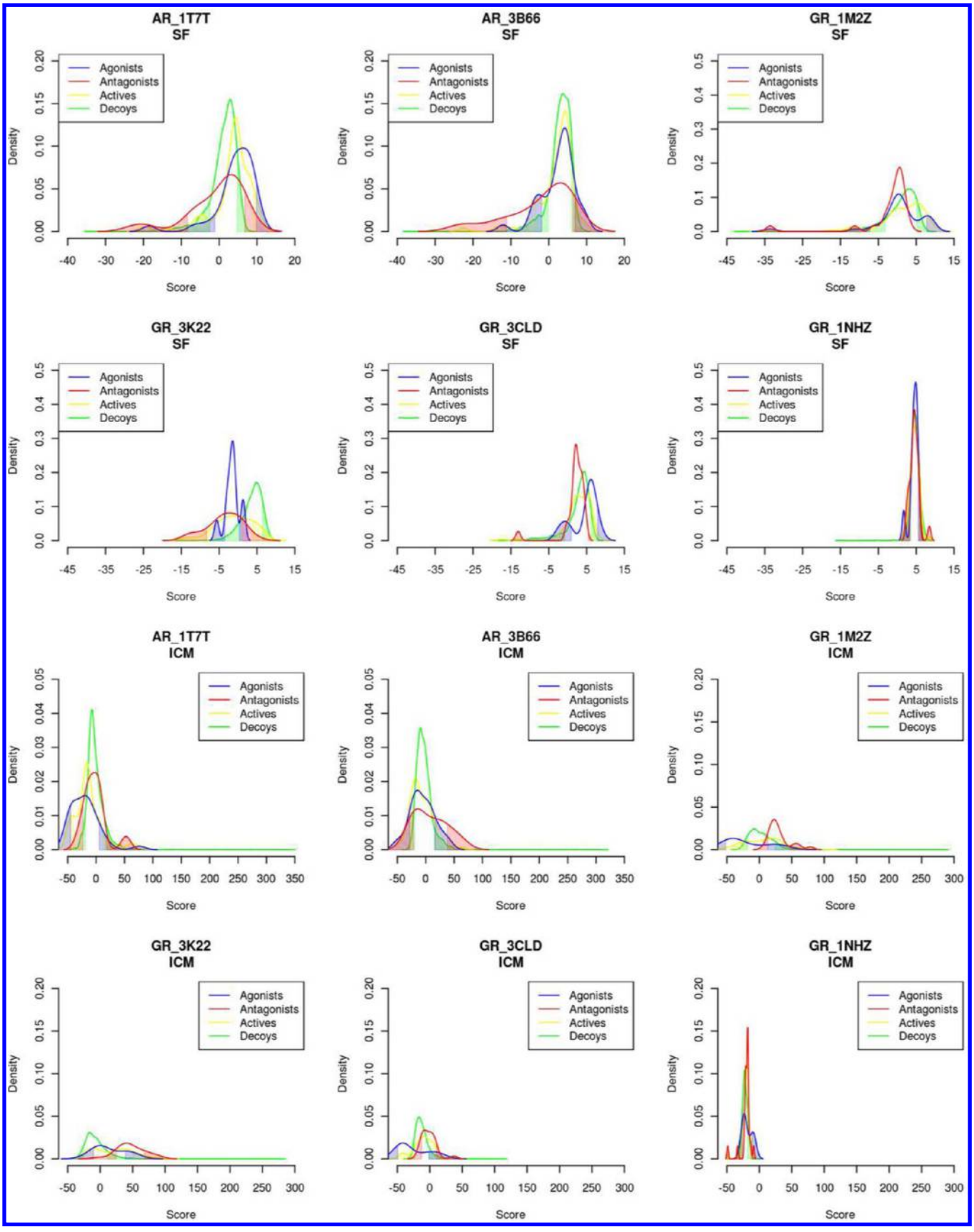


Figure 3. Score distribution curves of the agonist compounds (blue), antagonist compounds (red) within the active compounds (yellow) and their associated decoys (green) of the DUD data sets AR and GR using Surflex-dock or ICM on the different structures of AR and GR used in the study. For AR, 1T7T is agonist-bound; 3B66 is modulator-bound. For GR, 1M2Z, 3K22, and 3CLD are agonist-bound; 1NHZ is antagonist-bound.

dx.doi.org/10.1021/ci3004557 | J. Chem. Inf. Model. 2013, 53, 293−311

Except for ER where the distinction between the pharmacological profile of the ligands (agonist, antagonist, or modulator) has already been made in DUD2, there are 6 other NRs, i.e. AR, GR, MR, PPAR, PR, and RXR, for which this distinction has not been performed. It is known that the pharmacological profile of the bound ligands influences the structure of the target,[41] but this distinction has only been performed for ER in the DUD2. For GR and AR which comprise a sufficient number of agonists, antagonists, or modulators (12 antagonists and 36 agonists for AR; 23 antagonists and 11 agonists for GR), there are significant differences between the score distribution of the agonists and the score distribution of the antagonists when the reference structure is agonist-bound (1T7T for AR, 1M2Z, 3K22, and 3CLD for GR) or antagonist/modulator-bound (3B66 modulator-bound for AR, 1NHZ antagonist-bound for GR). This separation illustrated in Figure 3 is purely based on the pharmacological profile (whatever the docking method used) and demonstrates its influence on the resulting enrichments. Interestingly, this separation for agonists and antagonists of ER that existed in the DUD2 is no longer present in the DUD-E. Our results for GR and AR highlight the need to create separate DUD-own sets for NRs based on the pharmacological profiles of the known active ligands and the bound-structures used as a reference.

CONCLUSION

In the present work, we aimed to define the optimal protocol for a SBVLS experiment when different structures of the target are available. Our main objective was to assess guidelines based on the binding site properties to identify the optimal structure without performing time-consuming preliminary enrichment studies on benchmarking databases. We thus selected up to four structures for each system within the DUD data set into the PDB based on their differences of binding site properties (volume, opening). We evaluated the performance of two docking softwares, ICM and Surflex-dock, on these targets in order to identify the structures that led to optimal enrichments, and we compared their binding site properties. We identified several trends: (1) There was a structure that led to optimal enrichment which is in majority the same whatever the docking program used (in our case SF or ICM). (2) For small systems, the best enrichments are in majority obtained with the conformers presenting the smallest binding site volume whereas the worst performances are associated with the structure with the largest binding sites. Opposite conclusions are driven for the large systems for which the best structures displayed the largest binding sites.

We then constructed, for each system, all the possible ensembles with the structures selected in the previous step, compared the performances of ICM and SF on all these ensembles, and analyzed the composition of the resulting best ensembles. In our results, the best ensembles generally comprised only two structures and always comprised the best structure from single docking. Similarly to the results obtained in single structure docking, in the small systems, the conformers presenting the smallest binding site were in majority present in the best ensemble and the one with the largest binding site in the worst ensemble. For the large systems, the best structures displayed the largest binding sites. Concerning the opening of the binding site, in large systems with both programs, the most closed structures were not part of the best ensembles whereas with SF, in the small systems, the most open structures were not part of the best ensembles. This information we got with two distinct docking programs could constitute helpful guidelines for the docking and screening community as there is a strong need for standardized protocols for selecting the right query especially when few ligands are known.

In addition to the issues that have been raised recently by Schneider et al.[39] in the DUD data set, we also identified several points that should be taken into account to help the construction of even more robust data sets.

In conclusion, we have been able to establish useful guidelines for the simple completion of a real-life SBVLS project, based on the binding site properties observed in the structures that have led to the optimal performance in our retrospective SBVLS tests.

ASSOCIATED CONTENT

Supporting Information

Docking accuracy, chemotype enrichments, and detailed ensemble docking results for SF and ICM. Species information for each structure selected for the study and details about the calculation of the volume of the binding sites using POVME. This material is available free of charge via the Internet at http://pubs.acs.org.

AUTHOR INFORMATION

Corresponding Author

*E-mail: matthieu.montes@cnam.fr.

Author Contributions

†These authors contributed equally to this work.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

We thank Prof Jain for generously providing the Surflex package and Molsoft LLC for providing academic licenses for the ICM suite. H.G. and N.L. are recipients of a CIFRE fellowship from ANRT. N.B.N. is the recipient of a MNRT fellowship.

ABBREVIATIONS

SBVLS, structure-based virtual ligand screening; ACE, angiotensin-converting enzyme; ACHE, acetylcholin esterase; ADA, adenosine deaminase; ALR2, aldose reductase; AMPC, AmpC beta lactamase; AR, androgen receptor; CDK2, cyclin dependent kinase 2; COMT, catechol *O*-methyltransferase; COX1, cyclooxygenase-1; COX2, cyclooxygenase-2; DHFR, dihydrofolate reductase; EGFR, epidermal growth factor receptor kinase; ER ago, estrogen receptor agonist; ER antago, estrogen receptor antagonist; FGFR1, fibroblast growth factor receptor kinase; FXA, factor Xa; GART, glycinamide ribonucleotide transformylase; GPB, glycogen phosphorylase beta; GR, glucocorticoid receptor; HIVPR, HIV protease; HIVRT, HIV reverse transcriptase; HMGR, hydroxymethyl-glutaryl-CoA reductase; HSP90, human heat shock protein 90 kinase; INHA, enoyl ACP reductase; MR, mineralocorticoid receptor; NA, neuraminidase; P38, P38 mitogen activated protein kinase; PARP, poly(ADP-ribose) polymerase; PDE5, phosphodiesterase V; PDGFR-β, platelet derived growth factor receptor kinase beta; PNP, purine nucleoside phosphorylase; PPAR, peroxisome proliferator activated receptor gamma; PR, progesterone receptor; RXR, retinoic X receptor alpha; SAHH, *S*-adenosyl-homocystein hydrolase; SRC, tyrosine kinase SRC;

THR, thrombin; TK, thymidine kinase; TRP, trypsin; VEGFR2, vascular endothelial growth factor receptor kinase; NR, nuclear receptors

REFERENCES

(1) Alvarez, J. C. High-throughput docking as a source of novel drug leads. *Curr. Opin. Chem. Biol.* **2004**, *8* (4), 365−370.

(2) Sherman, W.; Beard, H. S.; Farid, R. Use of an induced fit receptor structure in virtual screening. *Chem. Biol. Drug. Des.* **2006**, *67* (1), 83−84.

(3) Cavasotto, C. N.; Abagyan, R. A. Protein flexibility in ligand docking and virtual screening to protein kinases. *J. Mol. Biol.* **2004**, *337* (1), 209−225.

(4) Barril, X.; Morley, S. D. Unveiling the full potential of flexible receptor docking using multiple crystallographic structures. *J. Med. Chem.* **2005**, *48* (13), 4432−4443.

(5) Craig, I. R.; Essex, J. W.; Spiegel, K. Ensemble docking into multiple crystallographically derived protein structures: an evaluation based on the statistical analysis of enrichments. *J. Chem. Inf. Model.* **2010**, *50* (4), 511−524.

(6) Huang, S. Y.; Zou, X. Ensemble docking of multiple protein structures: considering protein structural variations in molecular docking. *Proteins* **2007**, *66* (2), 399−421.

(7) Bolstad, E. S.; Anderson, A. C. In pursuit of virtual lead optimization: pruning ensembles of receptor structures for increased efficiency and accuracy during docking. *Proteins* **2009**, *75* (1), 62−74.

(8) Broughton, H. B. A method for including protein flexibility in protein-ligand docking: improving tools for database mining and virtual screening. *J. Mol. Graph. Model.* **2000**, *18* (3), 247−257 302−304..

(9) Cavasotto, C. N.; Kovacs, J. A.; Abagyan, R. A. Representing receptor flexibility in ligand docking through relevant normal modes. *J. Am. Chem. Soc.* **2005**, *127* (26), 9632−9640.

(10) Frimurer, T. M.; Peters, G. H.; Iversen, L. F.; Andersen, H. S.; Moller, N. P.; Olsen, O. H. Ligand-induced conformational changes: improved predictions of ligand binding conformations and affinities. *Biophys. J.* **2003**, *84* (4), 2273−2281.

(11) Sperandio, O.; Mouawad, L.; Pinto, E.; Villoutreix, B. O.; Perahia, D.; Miteva, M. A. How to choose relevant multiple receptor conformations for virtual screening: a test case of Cdk2 and normal mode analysis. *Eur. Biophys. J.* **2010**, *39* (9), 1365−1372.

(12) Armen, R. S.; Chen, J.; Brooks, C. L. An Evaluation of Explicit Receptor Flexibility in Molecular Docking Using Molecular Dynamics and Torsion Angle Molecular Dynamics. *J. Chem. Theory Comput.* **2009**, *5* (10), 2909−2923.

(13) Ferrari, A. M.; Wei, B. Q.; Costantino, L.; Shoichet, B. K. Soft docking and multiple receptor conformations in virtual screening. *J. Med. Chem.* **2004**, *47* (21), 5076−5084.

(14) Totrov, M.; Abagyan, R. Flexible ligand docking to multiple receptor conformations: a practical alternative. *Curr. Opin. Struct. Biol.* **2008**, *18* (2), 178−184.

(15) McGann, M. R.; Almond, H. R.; Nicholls, A.; Grant, J. A.; Brown, F. K. Gaussian docking functions. *Biopolymers* **2003**, *68* (1), 76−90.

(16) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. The Protein Data Bank. *Nucleic Acids Res.* **2000**, *28* (1), 235−242.

(17) Yoon, S.; Welsh, W. J. Identification of a minimal subset of receptor conformations for improved multiple conformation docking and two-step scoring. *J. Chem. Inf. Comput. Sci.* **2004**, *44* (1), 88−96.

(18) Rueda, M.; Bottegoni, G.; Abagyan, R. Recipes for the selection of experimental protein conformations for virtual screening. *J. Chem. Inf. Model.* **2010**, *50* (1), 186−193.

(19) Thomas, M. P.; McInnes, C.; Fischer, P. M. Protein structures in virtual screening: a case study with CDK2. *J. Med. Chem.* **2006**, *49* (1), 92−104.

(20) Rao, S.; Sanschagrin, P. C.; Greenwood, J. R.; Repasky, M. P.; Sherman, W.; Farid, R. Improving database enrichment through ensemble docking. *J. Comput. Aided Mol. Des.* **2008**, *22* (9), 621−627.

(21) Bottegoni, G.; Rocchia, W.; Rueda, M.; Abagyan, R.; Cavalli, A. Systematic exploitation of multiple receptor conformations for virtual ligand screening. *PLoS One* **2011**, *6* (5), e18845.

(22) Korb, O.; Olsson, T. S.; Bowden, S. J.; Hall, R. J.; Verdonk, M. L.; Liebeschuetz, J. W.; Cole, J. C. Potential and limitations of ensemble docking. *J. Chem. Inf. Model.* **2012**, *52* (5), 1262−1274.

(23) Rueda, M.; Totrov, M.; Abagyan, R. ALiBERO: Evolving a team of complementary pocket conformations rather than a single leader. *J. Chem. Inf. Model.* **2012**, *52* (10), 2705−2714.

(24) Giganti, D.; Guillemain, H.; Spadoni, J. L.; Nilges, M.; Zagury, J. F.; Montes, M. Comparative evaluation of 3D virtual ligand screening methods: impact of the molecular alignment on enrichment. *J. Chem. Inf. Model.* **2010**, *50* (6), 992−1004.

(25) Hawkins, P. C.; Warren, G. L.; Skillman, A. G.; Nicholls, A. How to do an evaluation: pitfalls and traps. *J. Comput. Aided Mol. Des.* **2008**, *22* (3−4), 179−190.

(26) McGovern, S. L.; Shoichet, B. K. Information decay in molecular docking screens against holo, apo, and modeled conformations of enzymes. *J. Med. Chem.* **2003**, *46* (14), 2895−2907.

(27) McGann, M. FRED and HYBRID docking performance on standardized datasets. *J. Comput. Aided. Mol. Des.* **2012**, *26* (8), 897−906.

(28) Spitzer, R.; Jain, A. N. Surflex-Dock: Docking benchmarks and real-world application. *J. Comput. Aided Mol. Des.* **2012**, *26* (6), 687−699.

(29) Brozell, S. R.; Mukherjee, S.; Balius, T. E.; Roe, D. R.; Case, D. A.; Rizzo, R. C. Evaluation of DOCK 6 as a pose generation and database enrichment tool. *J. Comput. Aided Mol. Des.* **2012**, *26* (6), 749−773.

(30) Dundas, J.; Ouyang, Z.; Tseng, J.; Binkowski, A.; Turpaz, Y.; Liang, J. CASTp: computed atlas of surface topography of proteins with structural and topographical mapping of functionally annotated residues. *Nucleic Acids Res.* **2006**, *34* (Web Serverissue), W116−W118.

(31) Durrant, J. D.; de Oliveira, C. A.; McCammon, J. A. POVME: an algorithm for measuring binding-pocket volumes. *J. Mol. Graph. Model.* **2011**, *29* (5), 773−776.

(32) Fraczkiewicz, R.; W, B. Exact and Efficient Analytical Calculation of the Accessible Surface Areas and Their Gradients for Macromolecules. *J. Comput. Chem.* **1998**, *19* (3), 319−333.

(33) Pettersen, E. F.; Goddard, T. D.; Huang, C. C.; Couch, G. S.; Greenblatt, D. M.; Meng, E. C.; Ferrin, T. E. UCSF Chimera–a visualization system for exploratory research and analysis. *J. Comput. Chem.* **2004**, *25* (13), 1605−1612.

(34) Jain, A. N. Surflex: fully automatic flexible molecular docking using a molecular similarity-based search engine. *J. Med. Chem.* **2003**, *46* (4), 499−511.

(35) Abagyan, R.; Totrov, M.; Kusnetsov, D. ICM - a new method for protein modelling and design. Applications to docking and structure prediction from the distorted native conformation. *J. Comput. Chem.* **1994**, *15*, 488−506.

(36) Schapira, M.; Abagyan, R.; Totrov, M. Nuclear hormone receptor targeted virtual screening. *J. Med. Chem.* **2003**, *46* (14), 3045−3059.

(37) Sing, T.; Sander, O.; Beerenwinkel, N.; Lengauer, T. ROCR: visualizing classifier performance in R. *Bioinformatics* **2005**, *21* (20), 3940−3941.

(38) Birch, L.; Murray, C. W.; Hartshorn, M. J.; Tickle, I. J.; Verdonk, M. L. Sensitivity of molecular docking to induced fit effects in influenza virus neuraminidase. *J. Comput. Aided Mol. Des.* **2002**, *16* (12), 855−869.

(39) Schneider, N.; Hindle, S.; Lange, G.; Klein, R.; Albrecht, J.; Briem, H.; Beyer, K.; Claussen, H.; Gastreich, M.; Lemmen, C.; Rarey, M. Substantial improvements in large-scale redocking and screening using the novel HYDE scoring function. *J. Comput. Aided Mol. Des.* **2012**, *26* (6), 701−723.

(40) 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* (14), 6582−6594.
(41) Bourguet, W.; Germain, P.; Gronemeyer, H. Nuclear receptor ligand-binding domains: three-dimensional structures, molecular interactions and pharmacological implications. *Trends Pharmacol. Sci.* **2000**, *21* (10), 381−388.