Combination of pose and rank consensus in docking-based virtual screening: the best of both worlds

Valeria Scardino[1,2], Mariela Bollini[3], Claudio N. Cavasotto[2,4,5,*]

[1] Meton AI, Inc., Wilmington, DE, 19801
[2] Austral Institute for Applied Artificial Intelligence, Universidad Austral, Pilar, Buenos Aires, Argentina
[3] Centro de Investigaciones en BioNanociencias (CIBION), Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET), Ciudad de Buenos Aires, Argentina
[4] Computational Drug Design and Biomedical Informatics Laboratory, Translational Medicine Research Institute (IIMT), Universidad Austral-CONICET, Pilar, Buenos Aires, Argentina
[5] Facultad de Ciencias Biomédicas, and Facultad de Ingeniería, Universidad Austral, Pilar, Buenos Aires, Argentina

***** Corresponding Author
CCavasotto@austral.edu.ar; cnc@cavasotto-lab.net

Keywords: High-throughput docking; Consensus ranking; Pose consensus; Pose/Ranking Consensus (PRC); Molecular docking

Abstract

The use of high-throughput docking (HTD) in the drug discovery pipeline is today widely established. In spite of methodological improvements in docking accuracy (pose prediction), scoring power, ranking power, and screening power in HTD remain challenging. In fact, pose prediction is of critical importance in view of the pose-dependent scoring process, since incorrect poses will necessarily decrease the ranking power of scoring functions. The combination of results from different docking programs (consensus scoring) has been shown to improve the performance of HTD. Moreover, it has been also shown that a pose consensus approach might also result in database enrichment. We present a new methodology named Pose/Ranking Consensus (PRC) that combines both pose and ranking consensus approaches, to overcome the limitations of each stand-alone strategy. This approach has been developed using four docking programs (ICM, rDock, Auto Dock 4, and PLANTS; the first one is commercial, the other three free). We undertook a thorough analysis for the best way of combining pose and rank strategies, and applied the PRC to a wide range of targets sampling different protein families with a variety of binding site properties. Our approach exhibits an improved systematic performance in terms of enrichment factor and hit rate with respect to either pose consensus or consensus ranking alone strategies at a lower computational cost, while always ensuring the recovery of a suitable number of ligands. An analysis using four free docking programs (replacing ICM by Auto Dock Vina) displayed comparable results.

Introduction

The experimental evaluation of chemical libraries for activity against a target of pharmaceutical interest through high-throughput screening has been long used in the drug discovery pipeline; this is a both time and resource consuming technique[1]. Computational methods are today valuable and established tools in all drug discovery endeavors, saving time, resources, and costs[2-4].

Among *in silico* methods in drug discovery, molecular docking has been widely used during the last three decades[4-6]. In protein-molecule docking, the optimal position, orientation and conformation (pose) of each molecule within the binding site is assessed ("docking stage"), and an estimation of its binding energy calculated. High-throughput docking (HTD) allows the screening of large chemical libraries (from thousands to millions of molecules) to generate a hit-list enriched with potential binders, which will be then prioritize for biochemical and biological evaluation. To be computationally efficient, HTD involves several approximations at different levels[7], and the binding free energy calculation is later replaced by a docking score, which is a measure of the probability that the molecule will bind to the target. Thus, the docking stage is followed in this case by the "scoring stage"[7, 8].

In spite of its undoubted success, HTD is not without challenges, since its performance depends on the energy representation of the system, the degree of target flexibility[4, 9-11], and the consideration of water molecules within the binding site[4, 12, 13]. A recent extensive comparison of docking programs showed that, in agreement with earlier works[14, 15], they perform better in terms of docking accuracy (docking stage) than in terms of scoring power, ranking power, and screening power (scoring stage)[16]. We would like to stress that pose prediction is nevertheless of the utmost importance in molecular docking, since incorrect poses will result in meaningless scores, which would thus reduce the ranking capacity of scoring functions. The performance of HTD using different docking programs has been further evaluated on several systems[17-19], and several inconsistencies have been found, such as different performances across programs, also showing that the effectiveness of each scoring function is system dependent.[18, 20, 21]. Several efforts have been conducted to improve the reliability at the scoring stage, such as machine-learning-based scoring functions[22, 23], and quantum mechanical-base scoring[24-29].

The combination of several docking programs (consensus scoring) has been shown to improve the performance of HTD[20, 30]. It should be highlighted that in consensus scoring (or ranking), it would desirable, for the sake of robustness, that scores for a given molecule be combined when the poses assessed by the different docking programs are similar. In 2013, Houston and Walkinshaw proposed for the first time a consensus docking procedure that used several docking programs to increase the reliability of the predicted poses[31]. Tuccinardi et al. later used ten docking protocols to evaluate pose consensus on database enrichment[32], and later extended their analysis to 36 benchmark targets of the DUD database[33]. They obtained comparable results to Arciniega and Lange's Docking Data Feature Analysis (DDFA), an approach for carrying out virtual screening analysis based on artificial neural networks which was among the best performing methods at the time[34]. To obtain good hit rates with their pose consensus strategy, molecules with at least seven matching poses between programs should be selected; in general, the best results were obtained with ten matching poses, which could represent a high computational cost. However, and more importantly, the number of ligands retrieved in most of those cases was very small, with the risk of being zero in some cases.

We present a new strategy that combines both pose and ranking consensus to overcome the limitations of each strategy when used in a stand-alone fashion, and thus increase the performance of HTD campaigns. This method, named Pose/Ranking Consensus (PRC) is consistent with theory in the sense that scores (or ranks) obtained with different programs are only combined when poses are coincident. We performed an exhaustive search to look for the best way of combining pose and rank requirements, and evaluated this new method over a wide range of targets that correspond to different protein families sampling different binding site properties. Our results show a consistent and improved performance compared to either pose consensus alone, or consensus scoring or ranking alone strategies. This method is simple to use and implement, and simpler than machine learning consensus scoring methods.

Methods

Target systems preparation

The sixteen targets listed in Table 1 were downloaded from the PDB. Water molecules and cofactors were deleted, except in the following cases (cf. Table 1): A Ca^{2+} was conserved within 8 Å of the native ligand in PA2GA and NRAM; a Zn^{2+} was conserved for LKHA4, and in HDAC2 both a Zn^{2+} and Ca^{2+} were conserved. In the case of water molecules, they were conserved in the following cases: HSP90a, water molecules 2059, 2121, 2123, and 2236; FA7, 2440; FABP4, 303, 623, 634, 665; LKHA4, 1099, 1322. The structure of the Dopamine D_3 receptor was in the antagonist bound conformation, and that of β_2 adrenergic receptor was in the agonist bound conformation. In the case of the HMG-CoA reductase, the binding site is within both protomers (chains a and b), which were both included for docking calculations.

Receptors were prepared with the ICM program[35] (version 3.8-7c; MolSoft, San Diego, CA 2020), in a similar fashion as in other works[25]. Missing residues and hydrogen atoms were added followed by a local energy minimization of the system. Polar and water hydrogens within the binding site were optimized using a Monte Carlo simulation in the torsional space. Glu and Asp side chains were assigned a -1 charge, and Lys and Arg were assigned a +1 charge. Asn and Gln were inspected for possible flipping and adjusted if necessary. Histidine tautomers were assigned according to their most favorable hydrogen bonding pattern.

Table 1: Target proteins used in the molecular docking calculations

Receptor	Receptor code	PDB entry	Resolution (Å)	Co-factors[a]	Water molecules[b]
Thymidine kinase	KITH	2B8T	2.0	-	-
Phospholipase A2	PA2GA	1KVO	2.0	Ca^{2+}	-
Coagulation factor VII	FA7	1W7X	1.8	-	1
Hexokinase type IV	HXK4	3F9M	1.5	-	-
Cyclin-dependent Kinase 2	CDK2	1FVV	2.8	-	-
Cyclooxygenase-1	COX1	2OYU	2.7	-	-
Fatty acid binding protein adipocite	FABP4	2NNQ	1.8	-	4
Heat shock protein 90 α	HSP90a	1UYG	2.0	-	4
Estrogen receptor α	ESR1	3ERT	1.9	-	-
Neuraminidase	NRAM	1B9V	2.3	Ca^{2+}	-

β_2 adrenergic receptor (agonist bound)	ADRB2	4LDO	3.2	-	-
HMG-CoA reductase	HMDH	3CCW	2.1	-	-
Dopamine D_3 receptor (antagonist bound)	DRD3	3PBL	2.8	-	-
Histone deacetylase 2	HDAC2	5IXO	1.7	Zn^{2+},Ca^{2+}	-
Leukocyte adhesion glycoprotein LFA-1α	LFA1	2ICA	1.6	-	-
Leukotriene A4 hydrolase	LKHA4	3CHP	2.1	Zn^{2+}	2

[a]Within 8 Å of the crystallographic ligand
[b]Within 4 Å of the crystallographic ligand

Docking libraries

Docking chemical libraries were prepared for each target by merging a set of actives and their corresponding matching decoys according to similar physico-chemical properties and structural dissimilarity, what it has been shown to ensure unbiased calculations in docking simulations[36, 37]. The number of actives, decoys and sources for each target are shown in Table 2. For all molecules, chirality and protonation states were inherited from the corresponding original databases.

Table 2: Docking chemical libraries

Receptor	Receptor code	Actives	Decoys	Source
Thymidine kinase	KITH	132	2866	DUD-E[38]
Phospholipase A2	PA2GA	127	5215	DUD-E[38]
Coagulation factor VII	FA7	185	6300	DUD-E[38]
Hexokinase type IV	HXK4	127	4802	DUD-E[38]
Cyclin-dependent Kinase 2	CDK2	72	2074	DUD[37]
Cyclooxygenase-1	COX1	210	6955	DUD-E[38]
Fatty acid binding protein adipocite	FABP4	57	2855	DUD-E[38]
Heat shock protein 90 α	HSP90a	125	4942	DUD-E[38]
Estrogen receptor α	ESR1	133	6555	NRLiSt[39]
Neuraminidase	NRAM	222	6227	DUD-E[38]
β_2 adrenergic receptor (agonist bound)	ADRB2	206	8034	GLL/GDD[36]
HMG-CoA reductase	HMDH	299	8884	DUD-E[38]
Dopamine D_3 receptor (antagonist bound)	DRD3	317	12363	GLL/GDD[36]
Histone deacetylase 2	HDAC2	238	10366	DUD-E[38]
Leukocyte adhesion glycoprotein LFA-1α	LFA1	233	8690	DUD-E[38]
Leukotriene A4 hydrolase	LKHA4	244	9477	DUD-E[38]

Docking calculations

For protein-molecule docking, five programs were used in total: ICM[35], Auto Dock 4[40], rDock[41], PLANTS[42], and Auto Dock Vina[43]. The latter was used for the free software evaluation only, replacing ICM. These programs have different search algorithms and scoring functions as described in previous works[30, 42]. For all the HTD runs, the top scored conformation of each molecule was selected. The box center and dimensions were determined with ICM in such a way that all molecules in the chemical library will fit within the binding site, and then used for all programs. In rDock, the docking cavity was automatically built using the reference ligand method, which defines a docking volume of a given size around the binding mode of a known ligand.

Auto Dock Tools utilities[40] were used to prepare the input files for Auto Dock 4, where the Lamarckian genetic algorithm was used for a 20-run search for each compound using 1750 000 steps of energy evaluation. For PLANTS, the ChemPLP scoring function was used and speed1 set as search speed. For rDock, a radius of 8.0 Å ± 2.0 Å from a reference ligand binding mode was used to represent the cavity. For Vina, an exhaustiveness value of 8 was set. For ICM, a thoroughness of 2 was used for the search algorithm. All the other parameters for every software remained at their default values. On average, each program took between 13 seconds and 130 seconds per core, per molecule, with ICM being the fastest and Auto Dock 4 the slowest program.

Exponential Consensus Ranking

In the Exponential Consensus Ranking (ECR)[30], the consensus rank *ECR(i)* for each molecule *i* is calculated as

$$ECR(i) = \frac{1}{\sigma} \sum_{j} \exp\left[-\frac{r_j(i)}{\sigma} \right] \quad (1)$$

where $r_j(i)$ is the rank of molecule *i* determined using the scoring function of program j, and σ is the expected value of the exponential distribution; while the ECR was found to be quasi-independent on σ,[30] we used σ = 10% of the total number of molecules for each docking library. Since the ECR is based on rank rather than score, it is thus independent on score units, scales and offsets.

Pose Consensus approach

From the HTD campaigns, 4 binding modes were obtained for each molecule in the database, which correspond to the 4 docking programs used. The RMSD between all combinations of these poses was calculated using the ICM software, which allowed for the calculation of the static deviation between molecules. Poses are considered to match if they are within 2.0 Å RMSD. A molecule is considered to have three matching poses (MPs) if the three corresponding combinations of two poses match. For four matching poses, the six corresponding combinations of two poses must be coincident.

Evaluation Metrics

The enrichment factor (*EF*) is defined as

$$EF(x) = \frac{Hits_x}{N_x} \Big/ \frac{Hits_{total}}{N_{total}} \quad (2)$$

where $Hits_x$ represents the number of actives present in a subset *x* of the docked library, N_x the number of molecules in subset *x*, $Hits_{total}$ is the total number of ligands within the entire chemical library, and N_{total} its total number of molecules. EF represents the probability of finding an actual ligand within subset *x* with respect to the probability of finding a ligand at random. Whenever a molecule were represented my multiple states regarding its protonation or chirality, each state was calculated its score, and the lowest score among those was used to build the rank and thus to calculate the *EF*.

The hit rate (HR) is calculated as

$$HR(x) = \frac{Hits_x}{N_x} \quad (3)$$

and is a measure between 0 and 1 which represents the probability of finding an actual ligand within the subset *x*.

Results and Discussion

We ran four HTD campaigns on 16 targets, which represent different protein families, and exhibit different binding site properties, including the presence of co-factors and water molecules (cf. Table M1). The chemical libraries used are described in Methods. Four docking programs were used, AutoDock 4, ICM, rDock and PLANTS, which have different search algorithms and scoring functions. Auto Dock Vina was also evaluated, but we selected only the best four performing programs to develop a method with the lowest computational cost for a future prospective campaign. For each docking program, the pose corresponding to the best score for each molecule was selected, and the ranking was established according to that score. On average, ICM presented the best performance, followed by rDock. There was not a program that performed best over all the systems evaluated.

As starting point, we used the Exponential Consensus Ranking (ECR)[30]. This consensus method combines results from several docking programs using an exponential distribution for each individual rank. In a previous work, it demonstrated a higher performance than other traditional consensus strategies and individual programs. In this work we extended the analysis of ECR to 16 targets using four instead of the original six programs. Our results confirm its better performance when compared to individual programs. On average, it showed at least a 1.5-fold increase (ratio between EF1 for ECR over an individual program) over every program (Table 3).

Table 3: Average enrichment factor at 1% (EF1) for each individual program calculated on the 16 benchmark targets, and the average fold increase of the ECR method over each program.

Average	ICM	rDock	AutoDock4	PLANTS	ECR
EF1	17.3	9.0	4.8	7.1	15.5
Fold increase[a]	1.5	3.1	>100	>100	1.0

[a]Average value calculated as $\frac{1}{N}\sum_{i}^{N} EF1_i^{ECR} / EF1_i^{program}$, where *N* is the number of targets

Pose consensus alone is not enough to guarantee high enrichment

Initially, we evaluated the performance of a pose consensus alone strategy using the four docking programs on the 16 benchmarking targets. Table 4 shows the enrichment factor (EF) for each target, calculated on the subset of molecules that meet the selection criteria according to the number of matching poses (MPs) between programs. Poses are considered to match if they are within 2.0 Å RMSD. Consistent with earlier works[32, 33], in general, the EF increases as the number of coincident poses requested is increased. However, the number of ligands in some cases is already low when considering four coincident poses. For

example, in LKHA4 only 2 ligands are present in the subset of molecules selected, and similar numbers were seen for HDAC2 and ADRB2. Furthermore, it can be seen from these results that a solely pose consensus strategy with four docking programs is not enough to obtain acceptable EFs.

Table 4: EF values for a pose consensus alone strategy of at least two (2 MPs), three (3 MPs) and four matching poses (4 MPs). The best EF for each target is highlighted.

Receptor	2 MPs	3 MPs	4 MPs
KITH	1.4	2.6	4.7
PA2GA	1.7	3.4	4.7
FA7	1.7	3.4	3.2
HXK4	1.3	1.4	1.7
CDK2	1.2	1.9	3.3
COX1	1.1	1.3	1.5
FABP4	1.2	1.5	1.5
HSP90a	1.1	1.3	2.0
ESR1	1.1	1.7	3.6
NRAM	2.0	4.7	5.6
ADRB2	1.4	1.2	0.4
HMDH	1.6	3.2	3.6
DRD3	1.1	1.1	1.0
HDAC2	0.9	1.5	1.2
LFA1	0.6	0.9	1.9
LKHA4	1.0	1.4	0.9
Average	1.3	2.0	2.5

Combining pose and rank consensus outperforms previous strategies

We observed that applying a ranking filter on pose consensus enhanced the performance of the latter. To further explore this fact, various possible combinations of the number of required MPs and ranking thresholds were considered, and three general options were initially explored: A) Pose consensus with at least two programs, selecting among those only molecules with the two corresponding ranks in the top 5, 10, or 20%; B) Pose consensus with at least three programs, selecting among those only molecules with the three corresponding ranks in the top 5, 10, or 20%; C) Pose consensus with the four programs, selecting among those only molecules with the four corresponding ranks in the top 15, 20, or 25%. These three options were evaluated in terms of minimum, maximum, and average EF values for the 16 benchmark targets; among the ones that showed high averages, those with higher minimum values and EFs closer to the average were preferred, in order to prioritize strategies that work well across all targets. Strategies that exhibited the best EFs in those specific targets that displayed low performance in the four programs were also prioritized. For option A (two MPs), the best results were obtained with a 5% rank cutoff. For option B (three MPs), the best results were obtained with a 10% rank cutoff. For option C (four MPs), the best results were obtained with a 20% rank cutoff. Option B marginally showed the best performance among the three options, followed by option C. It was observed, however, that

in some cases there were very few molecules that met the requirements, and in HXK4 no actual ligands could be found. In Table 5 the best performance for each option is presented.

Table 5: EF values and Active/Selected (A/S) molecule rate for Option A (2 MPs – top 5%); Option B (3 MPs – top 10%); and Option C (4 MPs – top 20%). The best option for each target is highlighted.

Receptor	**Option A**		**Option B**		**Option C**	
	A/S[a]	**EF**	**A/S[a]**	**EF**	**A/S[a]**	**EF**
KITH	24/38	14.0	3/4	17.0	1/3	7.6
PA2GA	26/48	22.0	9/10	37.9	3/3	42.1
FA7	84/107	27.5	33/35	33.1	1/1	35.1
HXK4	17/67	9.8	0/10	0.0	0/2	0.0
CDK2	23/56	12.2	17/47	10.8	11/22	14.9
COX1	10/200	1.7	11/134	2.8	9/54	5.7
FABP4	22/65	17.3	14/30	23.8	5/7	36.5
HSP90a	21/82	10.4	5/23	8.8	0/5	0.0
ESR1	39/133	15.4	25/70	18.8	14/24	30.7
NRAM	20/53	11.0	10/12	24.2	0/1	0.0
ADRB2	63/144	17.6	9/38	9.5	1/10	4.0
HMDH	21/66	16.7	4/12	17.5	2/3	34.9
DRD3	11/137	3.2	2/29	2.8	0/6	0.0
HDAC2	19/83	13.0	8/19	23.9	1/5	11.3
LFA1	11/121	5.7	6/46	8.1	2/11	11.4
LKHA4	24/165	8.2	6/43	7.8	1/10	5.6
Average	27/98	12.9	10/35	15.4	3/10	15.0

[a]Number of Actives and Selected molecules for each target

Next, we considered a combination of the three options A, B, and C, in the following fashion: if a molecule has a maximum of two MPs, the corresponding ranks obtained with those two programs should be within the top 5%; with a maximum of three MPs, those corresponding three ranks must be within the top 10%; with four MPs, the four ranks are requested to be in the top 20%. While this strategy (named Option D) showed a slightly less average EF than Option B (15.1 vs 15.4), there were no cases where actual ligands could not be found. Therefore, it was preferred over each individual option. We explored other combinations of ranking thresholds, but 5%, 10% and 20% for two, three and four MPs, respectively, were the best choice (similar results were also obtained with values of 7%, 12% and 15%). If the selected molecules are sorted by ECR, and only those in the top 1.5% are selected (option E), an even better performance is obtained (Table 6). Threshold values between 0.5% and 2% were also evaluated, with 1.5% showing the best results. This last option showed the best performance in almost every target evaluated, with the exception of two cases (NRAM and HMDH) where the difference was minimal.

Table 6: Evolution of the EF values for different strategies as molecular selection criteria are added: pose consensus alone using 4 MPs; Option D (2 MPs top 5% - 3MPs top 10% - 4MPs top 20%; Option E or PRC (Option D with an ECR top1.5% threshold). The hit rate (HR) represents the probability of finding a ligand within the selected molecules.

Receptor	**4 MPs**		**Option D**		**Option E (PRC)**		
	A/S[a]	**EF**	**A/S[a]**	**EF**	**A/S[a]**	**EF**	**HR**
KITH	17/83	4.7	15/23	14.8	13/15	19.7	0.87
PA2GA	9/81	4.7	16/30	22.4	12/16	31.5	0.75
FA7	9/100	3.2	64/73	30.7	44/45	34.3	0.98
HXK4	6/135	1.7	15/45	12.9	9/23	15.2	0.39
CDK2	25/225	3.3	14/33	12.6	11/17	19.3	0.65
COX1	51/114	1.5	11/111	3.4	8/47	5.8	0.17
FABP4	8/270	1.5	20/37	27.6	20/25	40.9	0.80
HSP90a	29/534	2.0	11/37	12.1	8/21	15.4	0.38
ESR1	32/470	3.6	28/77	19.1	28/53	27.8	0.53
NRAM	14/72	5.6	14/29	14.0	9/19	13.8	0.47
ADRB2	3/274	0.4	53/98	21.7	35/60	23.4	0.58
HMDH	4/59	3.6	18/52	18.1	10/30	17.5	0.33
DRD3	10/421	1.0	7/74	3.8	6/48	5.0	0.13
HDAC2	2/93	1.2	18/67	15.2	16/43	21.1	0.37
LFA1	9/290	1.9	5/59	5.3	5/43	7.3	0.12
LKHA4	2/128	0.9	18/129	7.9	10/69	8.1	0.14
Average	14/274	2.5	20/61	15.1	15/36	19.1	0.48

[a]Number of Actives and Selected molecules for each target

Figure 1 shows a schematic representation of this Pose/Rank Consensus (PRC) pipeline. Starting from the binding modes and ranks obtained with the four docking programs, a Pose/Rank filtering approach is carried out. For this, the maximum number of MPs (1-4) is assessed for each molecule, coupled with identifying those programs where the poses matched. Then, we look first for the ones that have four MPs and we filter them according to the 20% rank threshold in the corresponding programs. The same is performed for three MPs (10% rank threshold), and two MPs (5% rank threshold). In parallel, the ECR method is calculated onto the whole database. The molecules that pass the Pose/Rank filters are ordered by their corresponding ECR, previously calculated, and the ones in the top 1.5% are finally selected.

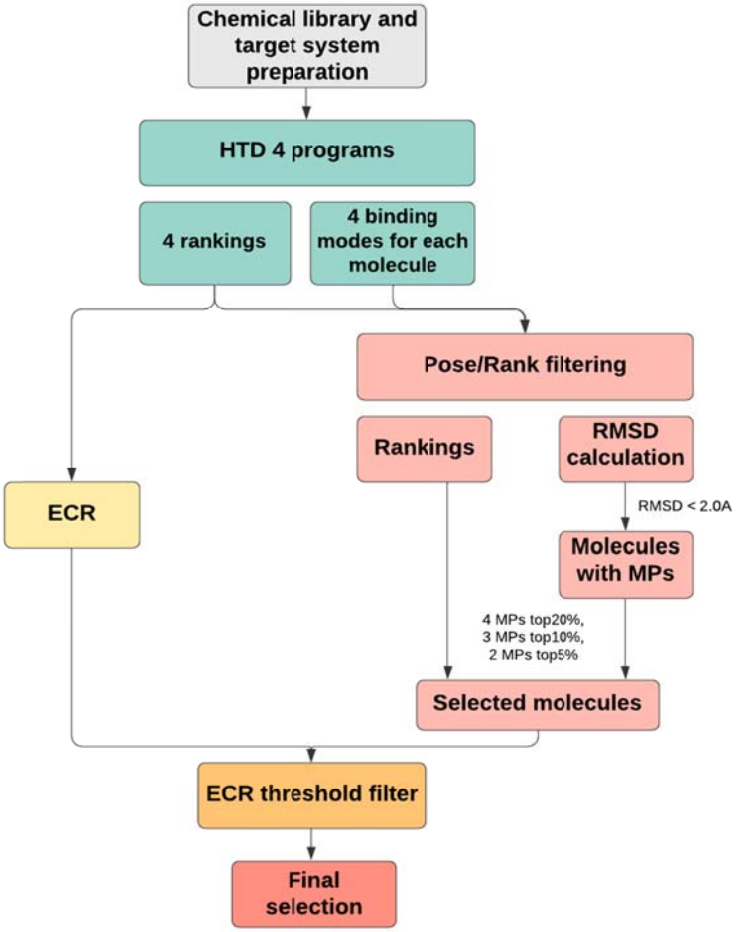

Figure 1: PRC pipeline for high-throughput docking.

This novel method achieves very high EF values, with an average hit rate of 48%, greatly surpassing previous ranking consensus techniques, including the ECR, and pose consensus techniques. The results are especially higher for those targets that have a poor performance in the four docking programs (and ECR), reaching EFs of more than triple the values of EF1 ECR (see below and Table 7).

Performance of the PRC method compared to ECR and Option D in view of traditional metrics

In Table 6, we show the performance of the PRC method in terms of EF and number of actual ligands (actives) retrieved for each target. EF values of pose consensus only and Option D selection strategy are also presented. For pose consensus, the EF of four MPs is shown as it performs best (Table 4). The last column shows the hit rate (probability of finding an actual ligand within the selected pool of molecules) in the PRC selected compounds. It can be readily noticed from these results that both the Pose/Rank filtering and ECR threshold requirements are important to achieve high EF values.

It can be seen from Table 6 that our method allows to obtain very high enrichment values, with an appropriate number of ligands. This could be critical in a prospective scenario, where the number of actual ligands might be scarce. When viewed in terms of probability, an average hit rate of 48% is achieved on the subsets of molecules selected.

The maximum value (98%) was obtained for FA7 where 44 out of 45 selected molecules are ligands. LFA1 showed the lowest hit rate value (12%) and the lowest number of ligands retrieved (5). In 2016, Tuccinardi et al. achieved an average hit rate of 45%, which they demanded to be at the level of the best performing methods[34]. We note, however, that the results they report correspond to the maximum hit rate that can be obtained for each target, which depends on the number of MPs used, and therefore is not directly applicable in a prospective analysis.

To further evaluate the performance of the PRC method we compared the improvement against ECR for every target. Table 7 shows the EFs of the PRC method compared to those of ECR at 1% (EF1). We chose EF1 as it is a standard metric, widely used in virtual screening. The fold increase (the ratio between PRC EF and ECR EF1) is also presented for a clearer comparison of the results. It can be noticed that 11 out of 16 targets showed an increase in the EF. The remaining 5 targets showed almost the same results in both strategies. On average, the PRC method has a 1.57-fold increase over ECR EF1. The improvements are especially noticeable in targets with low EF1 both on individual programs and on ECR; for example, EF values are increased by a factor of three in PRC for Neuraminidase and HSP90a. Regarding LFA1 (the worst performer in PRC), the four docking programs performed poorly on this target, and our method allowed to obtain the same EF as ECR EF1.

We also looked at the ECR EF when selecting the same number of molecules from the top as those returned by the PRC, for each target. It should be noticed that this is not a measure of practical value in prospective HTD, as this threshold is never known beforehand. However, the hit rate of PRC surpassed that of ECR (48% vs 43%, respectively); moreover, our method showed an eight times higher EF in HXK4, and still showed 3-fold increase values in the worst performing targets.

Taking into account that the ECR already implies an improvement of the results over previous consensus strategies and to individual programs, these results show that the PRC method allows for significantly higher hit rates and EF values, with a minimum computational cost, and therefore can reach the best results in future prospective HTD campaigns.

Table 7: Comparison of the EF at 1% (EF1) for PRC and ECR. The fold increase (PRC EF/ECR EF1) is also displayed in the last column.

Receptor	ECR EF1	PRC EF	Fold Increase
KITH	12.5	19.7	1.57
PA2GA	25.4	31.5	1.24
FA7	34.5	34.3	0.99
HXK4	5.5	15.2	2.74
CDK2	18.5	19.3	1.05
COX1	3.4	5.8	1.73
FABP4	40.5	40.9	1.01
HSP90a	4.9	15.4	3.17
ESR1	31.1	27.8	0.89
NRAM	4.5	13.8	3.03
ADRB2	24.5	23.4	0.96
HMDH	10.0	17.5	1.75
DRD3	3.2	5.0	1.57
HDAC2	13.6	21.1	1.55
LFA1	7.3	7.3	1.00
LKHA4	9.4	8.1	0.87
Average	15.5	19.1	1.57[a]

[a] Average of the fold increase values

Performance of the PRC method using only free available docking programs

In some cases, it may happen that only free docking programs are available. Therefore, we present the results using only free and accessible programs. For this task, we replaced ICM with Auto Dock Vina, which was the other available software. Table 8 shows the results obtained after applying the same PRC Pipeline (Figure 1) using Auto Dock 4, rDock, PLANTS and Auto Dock Vina for the HTD. For 2MPs, we evaluated the possibility of excluding the combination of Auto Dock 4 and Auto Dock Vina, as we saw that there were many molecules that met this requirement. This exclusion allowed better results, and so it was maintained for the free programs procedure. Option D selection strategy and pose consensus with 3 MPs are also displayed in Table 8 as a reference. In this case, 3 MPs are displayed instead of 4 MPs because the former performed best when using free docking programs. The maximum hit rate (77%) was obtained for PA2GA where 10 out of 13 molecules selected were active. For HSP90a, all the individual programs performed poorly, and no actual ligands could be found. On average, a hit rate of 30% was obtained. It can be noted that the best results were also obtained when combining the Pose/Rank filters with the ECR threshold (option E). However, Option D is shown as a good alternative for targets that do not perform well in none of the programs used, as it is the case of HXK4, HSP90a and NRAM.

Table 8: Same as Table 6 in a Free Programs context. 3 MPs are shown instead of 4 MPs because in this case the former performs best.

Receptor	3 MPs		Option D		Option E (PRC)		
	A/S[a]	EF	A/S[a]	EF	A/S[a]	EF	HR
KITH	42/380	2.5	3/17	4.0	3/9	7.6	0.33
PA2GA	24/450	3.9	11/16	28.9	10/13	32.4	0.77
FA7	49/488	3.5	27/40	23.7	22/29	26.6	0.76
HXK4	44/920	1.9	2/31	2.5	1/22	1.8	0.05
CDK2	39/549	2.1	17/34	14.9	12/22	16.3	0.55
COX1	119/2894	1.4	11/236	1.6	5/91	1.9	0.05
FABP4	22/714	1.6	13/64	10.4	12/23	26.7	0.52
HSP90a	53/1457	1.5	2/48	1.7	0/31	0.0	0.00
ESR1	50/1598	1.6	27/144	9.9	24/74	17.1	0.32
NRAM	54/513	3.1	4/33	3.5	2/25	2.3	0.08
ADRB2	55/2419	0.9	23/119	7.8	20/73	11.1	0.27
HMDH	18/432	2.2	12/41	15.3	4/21	10.0	0.19
DRD3	47/1854	1.0	9/151	2.4	6/70	3.4	0.09
HDAC2	17/781	1.2	20/72	15.7	15/38	22.4	0.39
LFA1	18/1136	1.0	6/86	4.4	6/53	7.1	0.11
LKHA4	18/780	1.3	34/181	10.6	23/74	17.5	0.31
Average	42/1085	1.9	14/82	9.8	10/42	12.7	0.30

[a]Number of Actives and Selected molecules for each target

In Table 9 we compare the results of PRC and ECR for free docking programs. Better results were obtained in 13 of the 16 targets with an average 1.75-fold increase of the PRC method over ECR EF1. Of the remaining three, LFA1 achieved almost the same results and HSP90a is the one that shows the highest decrease. In this target, Vina did not manage to perform well, showing zero EF1, and the other three programs also showed a poor performance. Option D achieved a slightly better EF as ECR EF1, and it may be a better selection strategy for cases where individual performances in terms of scoring stage are very poor. Regarding ESR1, while it still shows acceptable EF values, it performed slightly worse than ECR. This was also the case in the previous procedure (Table 7). It should be noted, anyway, that the number of selected molecules (74) is higher than 1% of the database (67). A very noticeable improvement of PRC over ECR can be seen for KITH, HXK4, NRAM and HMDH, where EF values of more than double the ECR EF1 were obtained. The average fold increase was even higher than in the previous case, therefore reaffirming the applicability of PRC method when only free docking programs are available.

Table 9: Same as Table 7 in a Free Programs context.

Receptor	ECR EF1	PRC EF	Fold Increase
KITH	2.3	7.6	3.22
PA2GA	16.7	32.4	1.94
FA7	24.1	26.6	1.10
HXK4	0.8	1.8	2.23
CDK2	12.8	16.3	1.27
COX1	1.0	1.9	1.95
FABP4	19.4	26.7	1.38
HSP90a	1.6	0.0	0.00
ESR1	23.9	17.1	0.71
NRAM	0.5	2.3	5.12
ADRB2	10.8	11.1	1.03
HMDH	3.5	10.0	2.85
DRD3	3.1	3.4	1.11
HDAC2	13.6	22.4	1.64
LFA1	7.3	7.1	0.97
LKHA4	12.3	17.5	1.42
Average	9.6	12.7	1.75

Conclusions and Perspective

A new method combining both pose and ranking consensus (PRC) is presented and evaluated in 16 diverse protein targets, displaying an improved performance with respect to either pose consensus alone, or consensus scoring alone approaches. Our method is especially robust in the sense that scores (and ranks) are only combined when poses are coincident within a 2 Å threshold. In the PRC method four docking programs to build consensus strategies were used (ICM, rDock, Auto Dock 4, and PLANTS), and we performed a comprehensive analysis for the optimal way of combining pose and rank requirements, which greatly improved the results compared to individual programs and also to previous consensus strategies. It should be noted that high hit rates were obtained with low computational cost, yielding an appropriate number of ligands. It was observed that PRC greatly improves the results even when only free available docking programs are used (replacing ICM by Auto Dock Vina).

In spite of the obvious success, we would like to point out two facts related to this methodology: i) It is still dependent on the performance of the individual programs on the target. If no program managed to perform well, then the PRC method would still improve the results obtained, but in a limited way; ii) Option D (cf. Table 6) is a good alternative in a prospective case when it is suspected that a little number of actual ligands might be present in the query database, or when the target belongs to a family of proteins that does not usually perform well in HTD campaigns, since it will likely retrieve more ligands. While (i) is a common limitation to all consensus strategies, PRC shows itself as a promising tool to solve it. In a follow-up contribution, we will evaluate the dependence of the method on the relationship between the number of ligands and decoys in the database for each target.

Acknowledgments

This work was supported by the National Agency for the Promotion of Science and Technology (ANPCyT) (PICT-2017-3767). CNC thanks Molsoft LLC (San Diego, CA) for providing an academic license for the ICM program. The authors thank the Centro de Cálculo de Alto Desempeño (Universidad Nacional de Córdoba) for granting the use of their computational resources.

References

1. Phatak, S. S.; Stephan, C. C.; Cavasotto, C. N., High-throughput and in silico screenings in drug discovery. *Exp. Opin. Drug Discov.* **2009**, 4, 947-959.
2. Jorgensen, W. L., Efficient drug lead discovery and optimization. *Acc. Chem. Res.* **2009**, 42, 724-33.
3. Schneider, G., Automating drug discovery. *Nat. Rev. Drug Discovery* **2017**, 17, 97-113.
4. Spyrakis, F.; Cavasotto, C. N., Open challenges in structure-based virtual screening: Receptor modeling, target flexibility consideration and active site water molecules description. *Arch. Biochem. Biophys.* **2015**, 583, 105-19.
5. Ciancetta, A.; Moro, S. Protein-Ligand Docking: Virtual Screening and Applications to Drug Discovery. In *In Silico Drug Discovery and Design: Theory, Methods, Challenges, and Applications*, Cavasotto, C. N., Ed.; CRC Press, Taylor & Francis Group: Boca Raton, FL, 2015; Chapter 7, pp 189-213.
6. Sulimov, A.; Kutov, D.; Ilin, I.; Zheltkov, D.; Tyrtyshnikov, E.; Sulimov, V., Supercomputer docking with a large number of degrees of freedom. *SAR QSAR Environ. Res.* **2019**, 30, 733-749.
7. Cavasotto, C. N.; Orry, A. J., Ligand Docking and Structure-based Virtual Screening in Drug Discovery. *Curr. Top. Med. Chem.* **2007**, 7, 1006-1014.
8. Guedes, I. A.; Pereira, F. S. S.; Dardenne, L. E., Empirical Scoring Functions for Structure-Based Virtual Screening: Applications, Critical Aspects, and Challenges. *Front. Pharmacol.* **2018**, 9, 1089.
9. Cavasotto, C. N.; Singh, N., Docking and High Throughput Docking: Successes and the Challenge of Protein Flexibility. *Curr. Comput.-Aided Drug Design* **2008**, 4, 221-234.
10. Cozzini, P.; Kellogg, G. E.; Spyrakis, F.; Abraham, D. J.; Costantino, G.; Emerson, A.; Fanelli, F.; Gohlke, H.; Kuhn, L. A.; Morris, G. M.; Orozco, M.; Pertinhez, T. A.; Rizzi, M.; Sotriffer, C. A., Target flexibility: an emerging consideration in drug discovery and design. *J. Med. Chem.* **2008**, 51, 6237-55.
11. Cavasotto, C. N.; Aucar, M. G.; Adler, N. S., Computational chemistry in drug lead discovery and design. *Int. J. Quantum Chem.* **2019**, 119, e25678.
12. Amadasi, A.; Surface, J. A.; Spyrakis, F.; Cozzini, P.; Mozzarelli, A.; Kellogg, G. E., Robust classification of "relevant" water molecules in putative protein binding sites. *J. Med. Chem.* **2008**, 51, 1063-7.
13. Cozzini, P.; Fornabaio, M.; Mozzarelli, A.; Spyrakis, F.; Kellogg, G. E.; Abraham, D. J., Water: how to evaluate its contribution in protein-ligand interactions. *Int. J. Quantum Chem.* **2006**, 106, 647-651.
14. Cavasotto, C. N.; Abagyan, R. A., Protein flexibility in ligand docking and virtual screening to protein kinases. *J. Mol. Biol.* **2004**, 337, 209-225.
15. Slater, O.; Kontoyianni, M., The compromise of virtual screening and its impact on drug discovery. *Expert Opin. Drug Discov.* **2019**, 14, 619-637.
16. Su, M.; Yang, Q.; Du, Y.; Feng, G.; Liu, Z.; Li, Y.; Wang, R., Comparative Assessment of Scoring Functions: The CASF-2016 Update. *J. Chem. Inf. Model.* **2019**, 59, 895-913.
17. Çınaroğlu, S. S.; Timuçin, E., Comparative Assessment of Seven Docking Programs on a Nonredundant Metalloprotein Subset of the PDBbind Refined. *J. Chem. Inf. Model.* **2019**, 59, 3846-3859.
18. Wang, Z.; Sun, H.; Yao, X.; Li, D.; Xu, L.; Li, Y.; Tian, S.; Hou, T., Comprehensive evaluation of ten docking programs on a diverse set of protein-ligand complexes: the prediction accuracy of sampling power and scoring power. *Phys. Chem. Chem. Phys.* **2016**, 18, 12964-75.
19. Xu, W.; Lucke, A. J.; Fairlie, D. P., Comparing sixteen scoring functions for predicting biological activities of ligands for protein targets. *J. Mol. Graph. Model.* **2015**, 57, 76-88.

20. Kukol, A., Consensus virtual screening approaches to predict protein ligands. *Eur. J. Med. Chem.* **2011**, 46, 4661-4664.
21. Cross, J. B.; Thompson, D. C.; Rai, B. K.; Baber, J. C.; Fan, K. Y.; Hu, Y.; Humblet, C., Comparison of several molecular docking programs: pose prediction and virtual screening accuracy. *J. Chem. Inf. Model.* **2009**, 49, 1455-74.
22. Ballester, P. J., Selecting machine-learning scoring functions for structure-based virtual screening. *Drug Discovery Today: Technologies* **2019**, 32-33, 81-87.
23. Pereira, J. C.; Caffarena, E. R.; Dos Santos, C. N., Boosting Docking-Based Virtual Screening with Deep Learning. *J. Chem. Inf. Model.* **2016**, 56, 2495-2506.
24. Aucar, M. G.; Cavasotto, C. N., Molecular Docking Using Quantum Mechanical-Based Methods. *Methods Mol. Biol.* **2020**, 2114, 269-284.
25. Cavasotto, C. N.; Aucar, M. G., High-Throughput Docking Using Quantum Mechanical Scoring. *Front. Chem.* **2020**, 8, 246.
26. Eyrilmez, S. M.; Kopruluoglu, C.; Rezac, J.; Hobza, P., Impressive Enrichment of Semiempirical Quantum Mechanics-Based Scoring Function: HSP90 Protein with 4541 Inhibitors and Decoys. *ChemPhysChem* **2019**, 20, 2759-2766.
27. Sulimov, A. V.; Kutov, D. K.; Ilin, I. S.; Sulimov, V. B., [Docking with combined use of a force field and a quantum-chemical method]. *Biomed. Khim.* **2019**, 65, 80-85.
28. Cavasotto, C. N.; Di Filippo, J. I., In silico Drug Repurposing for COVID-19: Targeting SARS-CoV-2 Proteins through Docking and Consensus Ranking. *Mol. Inform.* **2021**, 40, e2000115.
29. Cavasotto, C. N.; Adler, N. S.; Aucar, M. G., Quantum Chemical Approaches in Structure-Based Virtual Screening and Lead Optimization. *Front. Chem.* **2018**, 6, 188.
30. Palacio-Rodriguez, K.; Lans, I.; Cavasotto, C. N.; Cossio, P., Exponential consensus ranking improves the outcome in docking and receptor ensemble docking. *Sci. Rep.* **2019**, 9, 5142.
31. Houston, D. R.; Walkinshaw, M. D., Consensus Docking: Improving the Reliability of Docking in a Virtual Screening Context. *J. Chem. Inf. Model.* **2013**, 53, 384-390.
32. Tuccinardi, T.; Poli, G.; Romboli, V.; Giordano, A.; Martinelli, A., Extensive consensus docking evaluation for ligand pose prediction and virtual screening studies. *J. Chem. Inf. Model.* **2014**, 54, 2980-6.
33. Poli, G.; Martinelli, A.; Tuccinardi, T., Reliability analysis and optimization of the consensus docking approach for the development of virtual screening studies. *J. Enzyme Inhib. Med. Chem.* **2016**, 31, 167-173.
34. Arciniega, M.; Lange, O. F., Improvement of Virtual Screening Results by Docking Data Feature Analysis. *J. Chem. Inf. Model.* **2014**, 54, 1401-1411.
35. Abagyan, R.; Totrov, M.; Kuznetsov, D., ICM - a New Method For Protein Modeling and Design - Applications to Docking and Structure Prediction From the Distorted Native Conformation. *J. Comput. Chem.* **1994**, 15, 488-506.
36. 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.
37. Huang, N.; Shoichet, B. K.; Irwin, J. J., Benchmarking sets for molecular docking. *J. Med. Chem.* **2006**, 49, 6789-801.
38. 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.
39. Lagarde, N.; Ben Nasr, N.; Jeremie, A.; Guillemain, H.; Laville, V.; Labib, T.; Zagury, J. F.; Montes, M., NRLiSt BDB, the manually curated nuclear receptors ligands and structures benchmarking database. *J. Med. Chem.* **2014**, 57, 3117-25.
40. Morris, G. M.; Huey, R.; Lindstrom, W.; Sanner, M. F.; Belew, R. K.; Goodsell, D. S.; Olson, A. J., AutoDock4 and AutoDockTools4: Automated docking with selective receptor flexibility. *J. Comput. Chem.* **2009**, 30, 2785-91.
41. Ruiz-Carmona, S.; Alvarez-Garcia, D.; Foloppe, N.; Garmendia-Doval, A. B.; Juhos, S.; Schmidtke, P.; Barril, X.; Hubbard, R. E.; Morley, S. D., rDock: A Fast, Versatile and Open Source Program for Docking Ligands to Proteins and Nucleic Acids. *PLoS Comput. Biol.* **2014**, 10, e1003571.
42. Korb, O.; Stutzle, T.; Exner, T. E., Empirical scoring functions for advanced protein-ligand docking with PLANTS. *J. Chem. Inf. Model.* **2009**, 49, 84-96.
43. Trott, O.; Olson, A. J., AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *J. Comput. Chem.* **2010**, 31, 455-61.