19th June 2019

Protein-Protein Docking for PROTAC discovery

OBJECTIVE:

As explained in my previous post [https://openlabnotebooks.org/2714-2/], my goal is to test whether structure-based approaches can guide the design of PROTACs. As a first step, I am evaluating whether protein-protein docking tools can accurately predict the interface between an E3 ligase and its target, a question recently explore by Drummond and Williams.[1] This step is necessary to define the relative orientation of the chemical moiety binding the E3 ligase and the chemical moiety binding the target protein. Once we have this information, the second step will be to design PROTACs that are compatible with the relative orientation of these 2 chemical moieties. For this first protein-protein docking step, I will compare predicted structures with crystal structures of three complexes: the first bromodomain of BRD4 (BRD4^{BD1}) bound to the E3 ligases CRBN [pdb codes 6boy and 6bnb], the second bromodomain of BRD4 (BRD4^{BD2}) bound to the E3 ligase VHL [pdb code 5t35] and, the bromodomain of SMARCA2^{BD} bound to the E3 ligase VHL [pdb code 6hay]. I will be using three different protein-protein docking tools: HADDOCK[2] Rosetta[3] and ICM[4], which all performed among the best at past CAPRI protein docking competitions (http://www.capri-docking.org/).

METHOD:

In all cases, I first prepared the isolated proteins (structure of the target protein or E3 ligase in complex with small-molecule inhibitor) as follows. I used ICM to protonate the protein and ligand, assign partial charges, build missing side-chains and optimize the rotameric or tautomeric states of Asn, Gln, and His side-chains with the icm command “convertObject a_5ueo. 1==1 yes yes no yes no yes ""+(1==2 ? "water=tight ":"")+(yes ? "tautomer ":"")”. For docking with ICM, I saved the resulting ICM-formatted object and saved as a pdb file for HADDOCK. For Haddock, and Rosetta, I first converted the optimized structure back to PDB format with the ICM commands “strip a_” and “delete a_//*vt*”, deleted hydrogen atoms with “delete a_//h*” and saved the structure as a PDB file.

For the BRD4-CRBN complex, I used for docking the structure of CRBN in complex with lenalidomide (4tz4), and the structure of BRD4^{BD1} in complex with JQ1 (3mxf). Lenalidomide and JQ1 correspond to the chemical handles of the PROTAC co-crystallized in the ternary structures (6boy and 6bnb), which is exactly what I need.

For the BRD4-VHL complex, I used for docking the structure of VHL in complex with an inhibitor (PDB code 4b9k) structurally related to the chemical handle found in the crystallized PROTAC (5t35). For BRD4^{BD2}, I had to use a structure (5ueo) in complex with a ligand that is chemically unrelated to the BRD4-binding moiety of the PROTAC, and manually replace this ligand with the BRD4 inhibitor JQ1 (after superimposing BRD4^{BD2} to the structure of BRD4^{BD1} bound to JQ1).

For the SMARCA2-VHL complex, I used for docking the structure of VHL in complex with an inhibitor (4b9k) structurally related to the chemical handle found in the crystallized PROTAC (6hay). For SMARCA2^{BD}, I used a structure in complex with the inhibitor (6haz) corresponding to the chemical handle found in the crystallized PROTAC.

Here are the docking protocols I used with the three softwares:

1) HADDOCK

HADDOCK (High Ambiguity Driven biomolecular DOCKing) is an information-driven flexible docking approach for the modelling of biomolecular complexes.[2]

The docking takes place between a ligand-bound E3 ligase, and a ligand-bound protein target. I focused the docking simulation on solvent-accessible residues within 5 Å of the small molecule ligands. The corresponding residue numbers are as follows:

-CRBN-BRD4^{BD1} complex

CRBN (4tz4: 351-353, 377, 378, 386, 388,400)
BRD4^{BD1}(3mxf: 78, 79, 81, 85, 92, 94, 97, 139, 140, 145, 146, 149)

-VHL-BRD4^{BD2} complex

VHL (4b9k: 65,67,69,76,107,109-110)
BRD4^{BD2} (5ueo: 374,380-381,385,387,390,429,432,437,438)

-VHL-SMARCA2^{BD} complex

VHL (4b9k: 65,67,69,76,107,109-110)
SMARCA2^{BD} (6haz: 1408,1410,1411,1417,1418,1420-1421,1463,1464,1469,1470)

I used the online server at https://haddock.science.uu.nl/ with the expert-interface. Haddock docking protocol includes three steps: rigid-body energy minimization (it0), semi-flexible refinement in torsion angle space (it1), final refinement in explicit solvent (water). The final output is a list of 40 docking poses clustered in about 9 to 15 groups based on their structural similarity. Next, for each docked pose, I superimpose the E3 ligase structure of the docked complex with the E3 ligase of the crystallized complex and calculated the C-alpha RMSD between the docked protein target and the target from the crystallized complex. According to the CAPRI protein docking competition, a Cα-RMSD $\leq$ 10 Å is considered an acceptable pose in protein-protein docking[5]. I also calculated the RMSD between the ligand in the docked target and the ligand in the target from the crystallized complex.

Here is a sample script that I use to generate these RMSD values in ICM.

```
errorAction=4
group table t {0.} "N" {""} "name" {""} "cluster" {0.}  “Ca_RMSD" {0.} "ligand_RMSD"
superimpose a_6boy.c a_protein2.m align

for i=1,40
  add t 1
  t.name[1]=Name(a_$i.)[1]
  t.cluster[1]=Split(t.name[1], "_")[1]

#Ca_RMSD
  superimpose a_6boy.b a_$i.a align
  t.Ca_RMSD[1]=Srmsd(a_6boy.c//ca a_$i.b//ca)
#Ligand_RMSD
  superimpose a_$i.b a_protein2_b.m align
  t.ligand_RMSD[1]=Srmsd(a_protein2.jq1 a_protein2_b.jq1 chemical)
  t.N[1]=i
  delete sequence
endfor
```

2) Rosetta

Rosetta is a Monte Carlo based docking approach for protein-protein docking. By default, the docking protocol assumes a fixed backbone and does translation, rotation and sidechain packing. Global docking is used when a pre-defined starting conformation is not available (which is the case here). The smaller protein (ligand) rotates around the larger protein (receptor), using a series of randomized starting positions. Local docking samples the conformational space around a pre-defined starting conformation (which will be in our case the top 10 scoring structures from global docking). In terms of scoring of docked poses, each Rosetta version has different scoring options. Here, I have tested two versions of Rosetta (3.8 and 3.10).

I started with a global docking simulation where I used the inhibitor-bound E3 ligase, and an apo version of the target. I had to delete the target-bound inhibitor, as unlike Haddock or ICM, Rosetta currently does not allow the “ligand” protein to have a bound small-molecule inhibitor. I used Linux clusters available at computecanada.org to run Rosetta (3.8 & 3.10). All jobs were run via MPI mode. Relevant PDB structure coordinates were combined into a single file and prepared for docking using the Rosetta ‘docking_prepack_protocol’ program. The initial global docking simulation was performed using ‘docking_protocol_mpi’ with the following command line options in both versions,

- partners A_B -dock_pert 5 25 -randomize2 -ex1 ex2aro -nstruct 20000

The top 10 scoring solutions from the global docking exercise produced by version 3.8 were used for local perturbations docking with Rosetta ‘docking_protocol_mpi’ as follows:

- partners A_B -dock_pert 8 18 -ex1 ex2aro -nstruct 2000
- partners A_B -dock_pert 3 8 -ex1 ex2aro -nstruct 2000
- partners A_B -dock_pert 1.5 4 -ex1 ex2aro -nstruct 2000

Options for input

Distance: the starting distance separating the E3-ligase and target protein was ~100Å
Docking perturbation: randomly perturb the input structure using a gaussian for translation and rotation with standard deviations (5Å, 25^0)

randomize2: Randomize the orientation of the second docking partner
beta: latest scoring function
spin: Spin a second docking partner around axes from center of mass of the first partner to the second partner
Nstruct: Specify the number of decoys, to generate.

Command for Compute Canada supercomputer:

```
docking options
-s complex.pdb -docking: partners A_B -randomize2 -spin -beta -nstruct 20000 -out:file:pdb XXX.pdb -ex1 -ex2
```

3) ICM

ICM uses a Fast Fourier Transform (FFT) docking approach for protein-protein docking. I used inhibitor bound E3 ligase and inhibitor-bound target protein. As described above for Haddock, docking was focused on solvent-accessible residues within 5 Å of the small-molecule ligands.

A sample command is

```
icm64 -vlscluster fftProtDock.icm Rec_4b9k.ob Lig_5ueo.ob fft_4b9k_5ueo_out_refined.icb recFocus=c/65,67,69,76,107,109-110 ligFocus=a/374,380-381,385,387,390,429,432,437,438
```

RESULTS:

I evaluated how accurately the Haddock, Rosetta and ICM docking protocols detailed above could predict the following complex crystal structures: CRBN-BRD4^{BD1} [6boy, 6bnb], VHL-BRD4^{BD2} [5t35] & VHL-SMARCA2^{BD} [6hay].

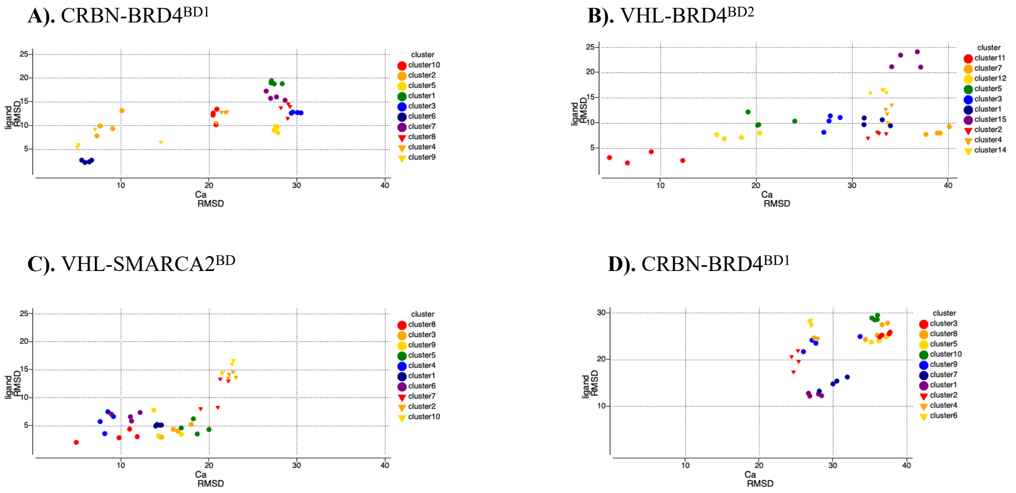
1. HADDOCK

Haddock clustered the top 400 docking poses and produced 40 representative solutions. Two of the 9 CRBN-BRD4^{BD1} clusters had BRD4 Cα-RMSDs between 1 and 10 Å and one had JQ1 (i.e BRD4 ligand) RMSD between 1.5 and 4 Å, which is a rather good result. Similarly, one of the 15 VHL-BRD4^{BD2} clusters had a BRD4 Cα-RMSDs between 1 and 12 Å and JQ1 RMSD between 2 and 5 Å (Fig. 1A, B). The ternary complex CRBN-PROTAC-BRD4^{BD1} was crystallized with two PROTACs leading to two conformational states (6boy and 6bnb). Haddock does reproduce the conformation found in 6boy (Fig. 1A) but fails to find the 6bnb arrangement (Fig 1D).

A major challenge for future steps is that all clusters generated by Haddock are equiprobable: in the absence of the experimental complex structure, how can we identify which of these clusters is populated with accurate docking poses? This is critical, as using the wrong cluster for subsequent PROTAC design is bound to fail (unless the experimental structure captured crystallographically is just one of several acceptable protein-protein interfaces predicted computationally and compatible with PROTAC design).

Results were quite different in the 3rd case-study: VHL-SMARCA2^{BD}. Here a number of diverse clusters positioned the ligand within 4 Å of the crystal structure, even when the SMARCA2 Cα-RMSD is > 10 or even >15 Å (Fig. 1C). If we use exclusively the relative position of the small molecule ligands for subsequent PROTAC design, the ligand RMSD is a more relevant metric. But here again, we don't have a method to distinguish good from poor docking poses.

Fig. 1| Ligand-RMSD vs Cα-RMSD for the 40 clustered poses produced by HADDOCK for **A).** CRBN-BRD4^{BD1}(6boy), **B).** VHL-BRD4^{BD2}, **C).** VHL-SMARCA2^{BD} **D).** CRBN-BRD4^{BD1}(6bnb)

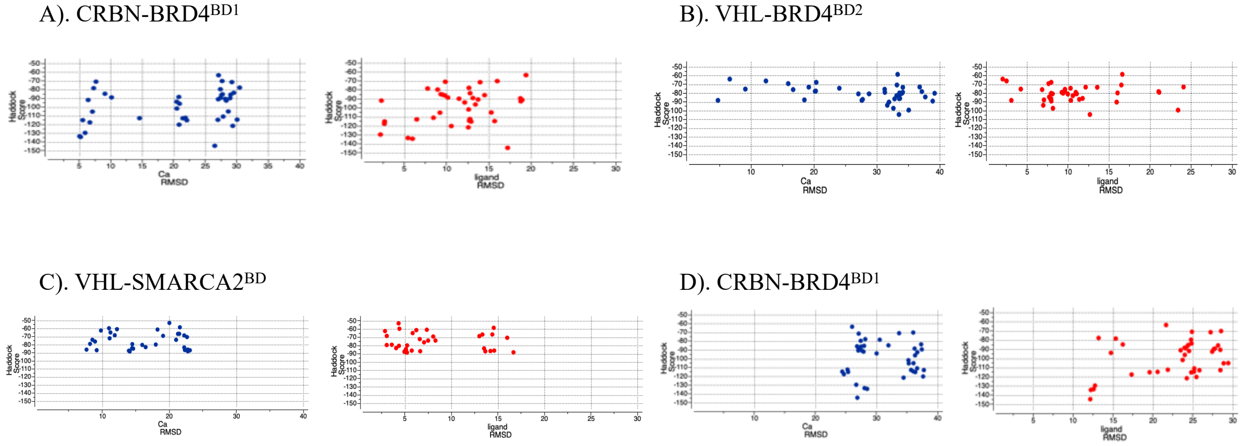

A). CRBN-BRD4^{BD1}

B). VHL-BRD4^{BD2}

C). VHL-SMARCA2^{BD}

D). CRBN-BRD4^{BD1}

Fig. 2| Haddock Score vs Cα-RMSD vs Ligand-RMSD for the 40 clustered poses produced by HADDOCK for **A).** CRBN-BRD4^{BD1}(6boy), **B).** VHL-BRD4^{BD2}, **C).** VHL-SMARCA2^{BD} **D).** CRBN-BRD4^{BD1}(6bnb)

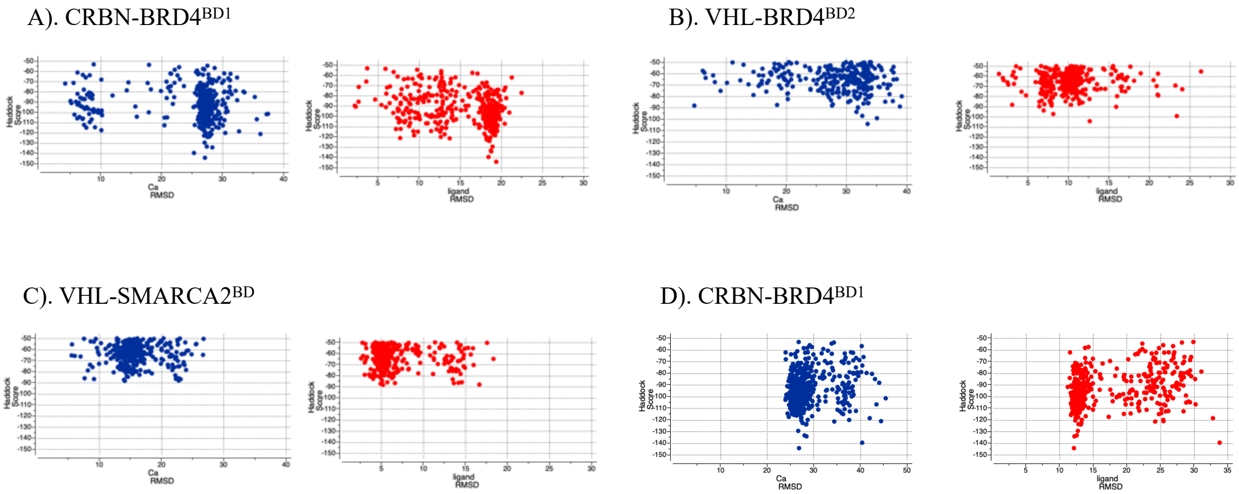

A). CRBN-BRD4^{BD1}

B). VHL-BRD4^{BD2}

C). VHL-SMARCA2^{BD}

D). CRBN-BRD4^{BD1}

Fig. 3| Haddock Score vs Cα-RMSD vs Ligand-RMSD for the 400 poses produced by HADDOCK for **A).** CRBN-BRD4^{BD1}(6boy), **B).** VHL-BRD4^{BD2}, **C).** VHL-SMARCA2^{BD} **D).** CRBN-BRD4^{BD1}(6bnb)


A). CRBN-BRD4^{BD1}

B). VHL-BRD4^{BD2}

C). VHL-SMARCA2^{BD}

D). CRBN-BRD4^{BD1}

I summarize these results in Table 1:

Table 1| Ligand-RMSD and Cα-RMSD for the 40 clustered representative poses and the top 400 parent poses produced by HADDOCK for CRBN-BRD4^{BD1} (6boy), VHL-BRD4^{BD2}, VHL-SMARCA2^{BD}.

1.) Top 40 final representative poses produced by Haddock

	Number of docked poses with C-alpha RMSD < 10Å from crystal structure	Lowest C-alpha RMSD(Å) of docked pose	Number of docked poses with ligand RMSD < 5Å from crystal structure		
			Top 10 scoring poses	Top 40 scoring poses	Lowest ligand RMSD(Å)
CRBN-BRD4^{BD1}	10/40	5.07	0	4/40	2.19
VHL-BRD4^{BD2}	3/40	4.71	1	4/40	2.03
VHL-SMARCA2^{BD}	7/40	4.95	0	15/40	2.02

2.) Total 400 poses produced by Haddock

	Number of docked poses with C-alpha RMSD < 10Å from crystal structure	Lowest C-alpha RMSD(Å) of docked pose	Number of docked poses with ligand RMSD < 5Å from crystal structure			
			Top 10 scoring poses	Top 50 scoring poses	Total scoring poses	Lowest ligand RMSD(Å)
CRBN-BRD4^{BD1}	48	4.20	0	0	8	2.19
VHL-BRD4^{BD2}	8	4.71	1	2	16	1.45
VHL-SMARCA2^{BD}	21	4.95	1	16	115	2.02

2. Rosetta

As mentioned above, Rosetta cannot dock a small molecule-protein complex to another small molecule-protein complex. Therefore, I had to remove the target-bound small molecule before docking the target to the small molecule-E3 ligase complex. Additionally, I could not find an option to focus docking on residues surrounding the small-molecule ligands (which I do when using Haddock or ICM). The first step is therefore a global docking. I first tested different protocols (Table 2) and found that version 3.8 produced better results (Nowak et. al. used version 3.7)[6].

Table 2: Summary of different options tried for finalize Rosetta protocol.

Method No	Protein-Protein Distance	Dock_pert	#Options	#functions	#extra-Options
1	10Å	3,8	randomize2	beta	
2	30Å	3,8	randomize2	beta	
3	100Å	3,8	randomize2	beta	spin randomize1
4	30Å	5,25	randomize2	beta	
5	100Å	5,25	randomize2	beta	spin
6*	100Å	5,25	randomize2		

Method No	Number of docked poses with C-alpha RMSD < 10Å from crystal structure	Number of docked poses with ligand RMSD < 5Å from crystal structure		
		Top 10 scoring poses	Top 50 scoring poses	Top 100 scoring poses
1	76	0	0	1
2	224	0	0	0
3	3	0	0	1
4	224	0	0	0
5	259	0	0	1
6*	284	0	3	4

*Nowak et.al NCB Paper

I plot below the interface score vs Cα-RMSD and vs ligand RMSD for the 3 complexes tested (CRBN-BRD4^{BD1}, VHL-BRD4^{BD2} & VHL-SMARCA2^{BD}) using method #6 above (Fig. 4). Each docking simulation generates an ensemble of 20,000 poses. None of the top 10 scoring poses for any of the 3 complexes had Cα-RMSD < 10 Å or ligand RMSD < 5 Å, but some were close, and local docking may improve the results.

Fig. 4| Interface Score vs Cα-RMSD and Interface Score vs ligand-RMSD results of Rosetta for A). CRBN-BRD4^{BD1}(6boy), B). VHL-BRD4^{BD2}, C). VHL-SMARCA2^{BD} D). CRBN-BRD4^{BD1}(6bnb)

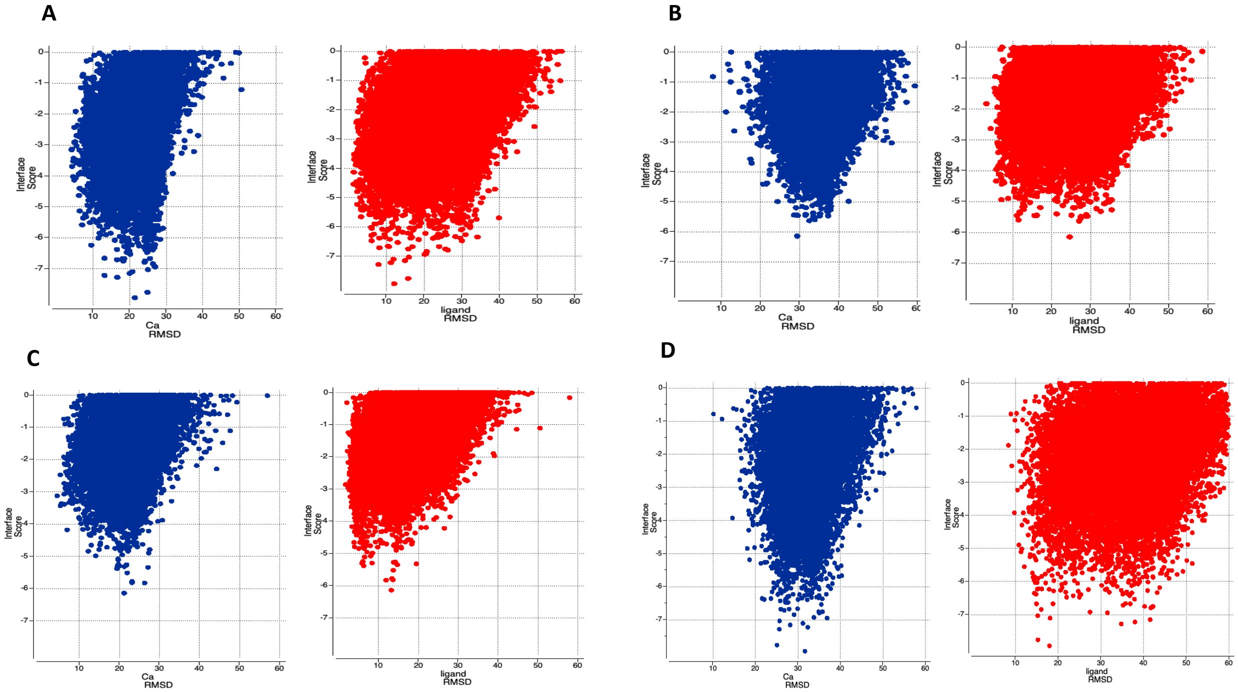


Table 3 summarizes the results obtained with global docking for the 3 complexes.

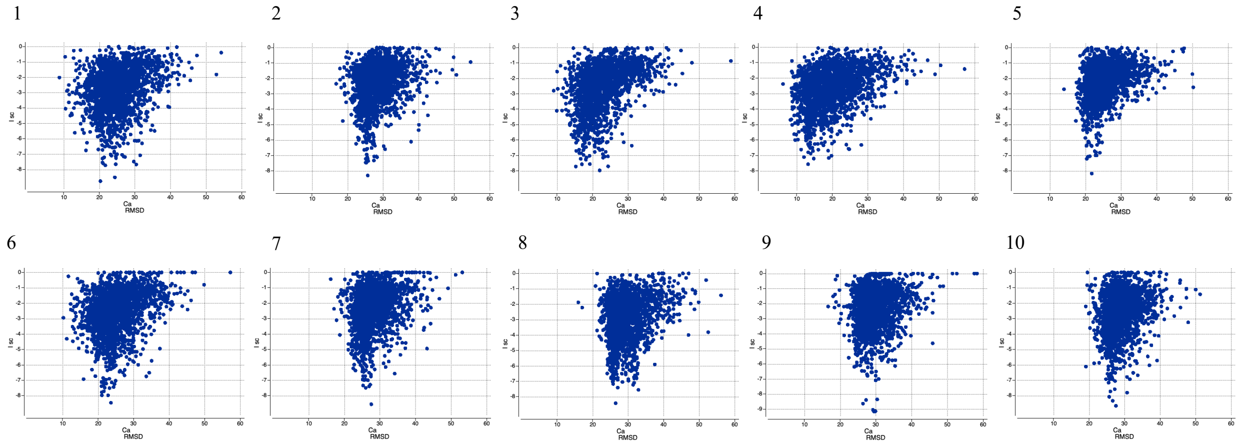
Table 3| Ligand-RMSD and Cα-RMSD for 20000 poses produced by Rosetta for CRBN-BRD4^{BD1}, VHL-BRD4^{BD2}, VHL-SMARCA2^{BD}.

	Number of docked poses with C-alpha RMSD < 10Å from crystal structure	Lowest C-alpha RMSD(Å) of docked pose	Number of docked poses with ligand RMSD < 5Å from crystal structure			
			Top 10 scoring poses	Top 50 scoring poses	Total scoring poses	Lowest ligand RMSD(Å)
CRBN-BRD4^{BD1}	235	4.28	0	0	169	1.45
VHL-BRD4^{BD2}	1	7.96	0	0	2	3.38
VHL-SMARCA2^{BD}	160	4.52	0	2	331	1.79

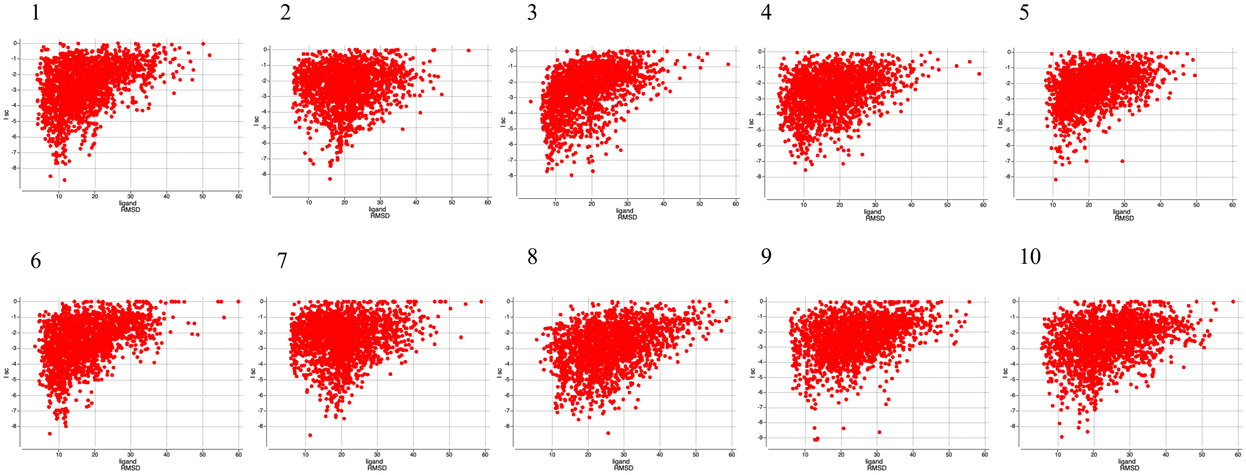
We tested this hypothesis by performing 10 local Rosetta docking using as starting conformation the top 10 scoring poses from the CRBN-BRD4^{BD1} run. The results (Fig. 5) indicate that local docking improves only modestly the accuracy of the predicted model. This is in sharp contrast with results previously published,[5] which I don't understand at the moment. Maybe the versions of Rosetta than I am using (3.8 and 3.10) cannot reproduce results generated with Rosetta 3.7 (not available on computecanada.org)? Any insight readers may have would be welcome.

Fig. 5| A). Interface Score vs Cα-RMSD(6boy), B). vs ligand RMSD(6boy), C). vs Cα-RMSD(6bnb), and D). vs ligand RMSD(6bnb) of CRBN-BRD4^{BD1} local docking simulations, using the top 10 poses from global docking as starting conformations.

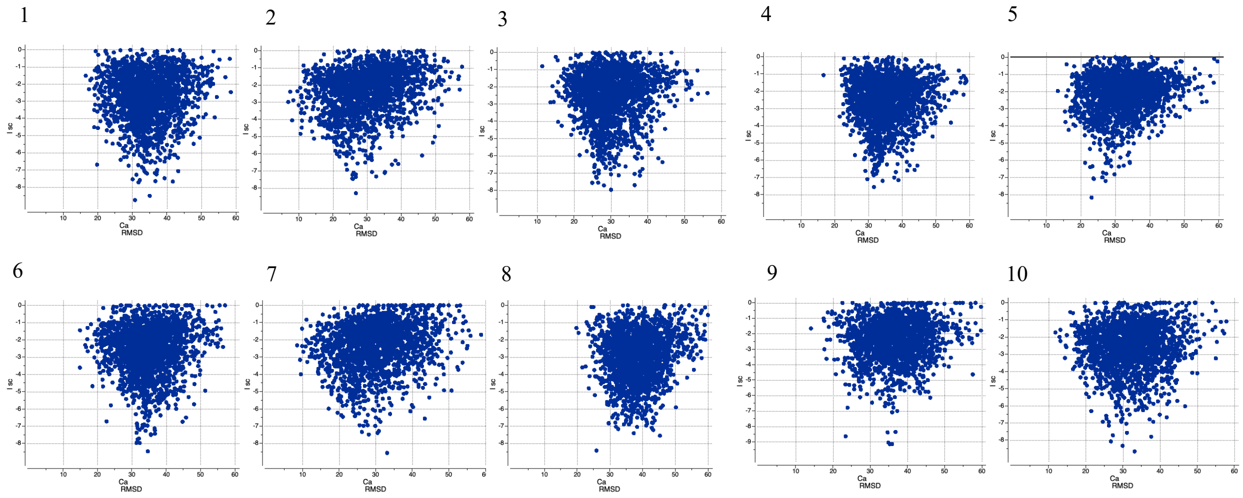
A).


1
2
3
4
5
6
7
8
9
10
I sc
0
-1
-2
-3
-4
-5
-6
-7
-8
-9
10
20
30
40
50
60
Ca
RMSD

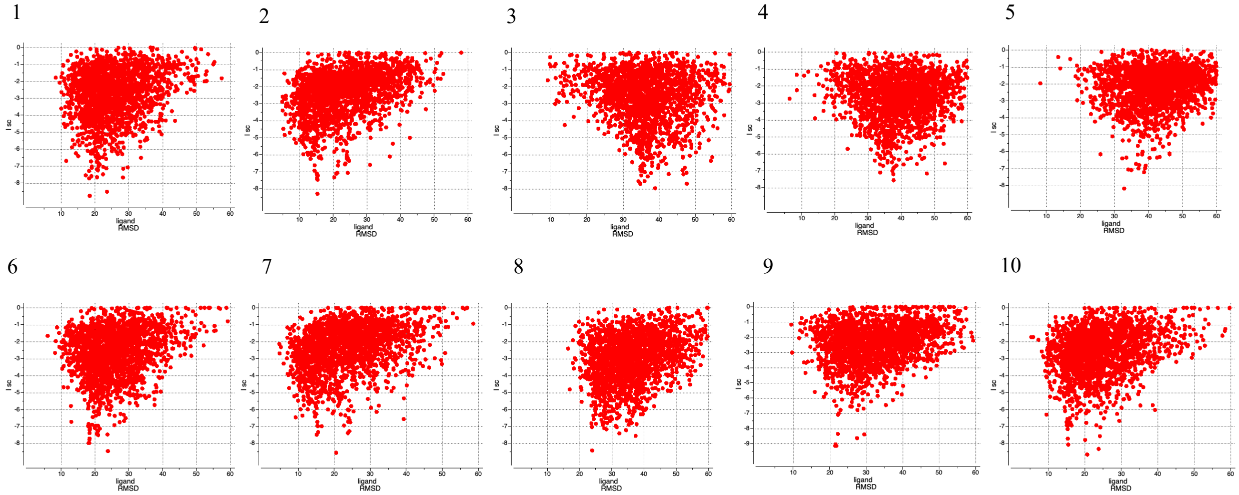
B).


1
2
3
4
5
6
7
8
9
10
I sc
0
-1
-2
-3
-4
-5
-6
-7
-8
-9
10
20
30
40
50
60
ligand
RMSD

C).


1
2
3
4
5
6
7
8
9
10
I sc
0
-1
-2
-3
-4
-5
-6
-7
-8
-9
10
20
30
40
50
60
Ca
RMSD

D).



3. ICM

I used the default version of the ICM FFT docking procedure described in the Methods section, which produces 10000 docking poses per complex. The resulting plots Score vs RMSD are shown Fig. 6, and the results are summarized in Table 4. Overall, results are not as good as those obtained with Haddock.

Fig. 6| Score vs Cα-RMSD and vs ligand-RMSD from ICM docking simulations

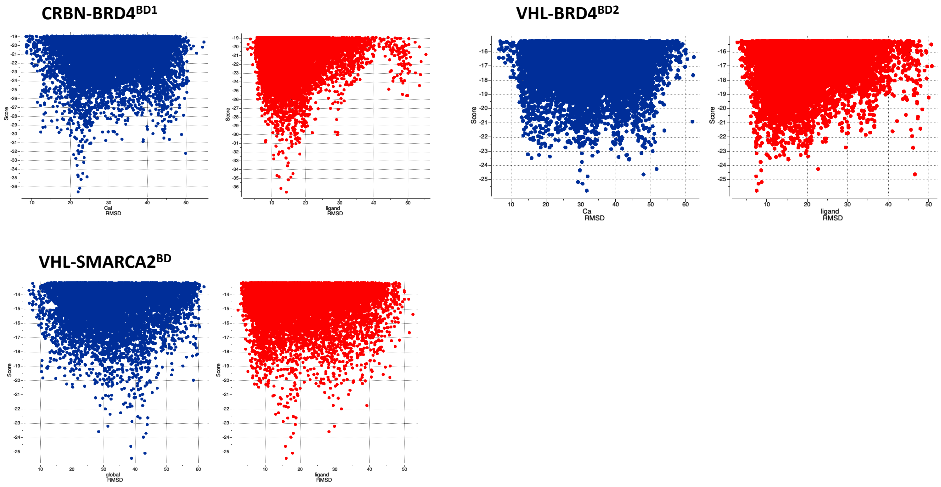


Table 4| Ligand-RMSD and Cα-RMSD for 10000 poses produced by ICM for CRBN-BRD4^{BD1}, VHL-BRD4^{BD2}, VHL-SMARCA2^{BD}.

	Number of docked poses with C-alpha RMSD < 10Å from crystal structure	Lowest C-alpha RMSD(Å) of docked pose	Number of docked poses with ligand RMSD < 5Å from crystal structure			
			Top 10 scoring poses	Top 50 scoring poses	Total scoring poses	Lowest ligand RMSD(Å)
CRBN-BRD4^{BD1}	17	8.42	0	0	8	2.94
VHL-BRD4^{BD2}	23	6.58	0	0	41	2.76
VHL-SMARCA2^{BD}	30	6.42	0	2	122	2.19

SUMMARY:

Haddock outperformed ICM and Rosetta in the three case studies tested here (Fig. 7, Table 5). Surprisingly, in the case of the VHL-SMARCA2 complex, multiple docked poses had > 10 or 15Å Ca-RMSD with the crystal structure, but ligand RMSD < 5 Å, reflecting the fact that different protein-docking poses can position the ligand at near identical (and rather accurate) positions. 5Å RMSD is a bad result for a ligand docking exercise (1.5 or 2Å is probably a better threshold for such exercise), but PROTAC linker design may be compatible with less accurate docking poses. Something that I will need to test in the future.

Fig. 7| Comparison of Haddock, Rosetta and ICM docking results on the 3 protein complexes tested here A) Cα RMSD B) small-molecule ligand RMSD.

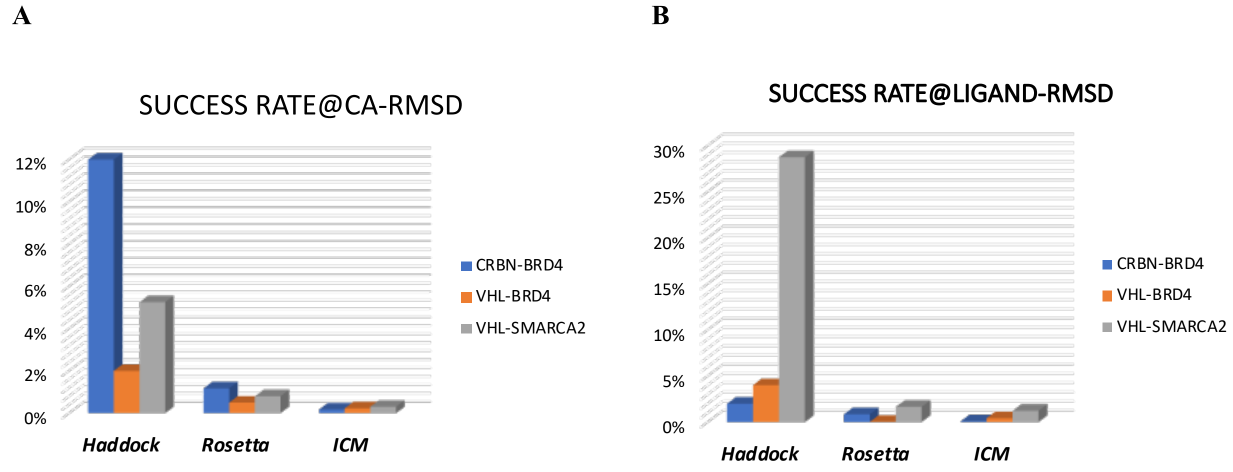


Table 5| Comparison of Haddock, Rosetta and ICM docking results on the 3 protein complexes tested here.

CRBN-BRD4^{BD1}	Number of docked poses with C-alpha RMSD < 10Å from crystal structure	Lowest C-alpha RMSD(Å) of docked pose	Number of docked poses with ligand RMSD < 5Å from crystal structure			
			Top 10 scoring poses	Top 50 scoring poses	Total scoring poses	Lowest ligand RMSD(Å)
Haddock	10/40	5.07	0	4	4/40	2.19
Rosetta	235/20000	4.28	0	0	169/20000	1.45
ICM	17/10000	8.42	0	0	8/10000	2.94

VHL-BRD4^{BD2}	Number of docked poses with C-alpha RMSD < 10Å from crystal structure	Lowest C-alpha RMSD(Å) of docked pose	Number of docked poses with ligand RMSD < 5Å from crystal structure			
			Top 10 scoring poses	Top 50 scoring poses	Total scoring poses	Lowest ligand RMSD(Å)
Haddock	3/40	4.71	1	4	4/40	2.03
Rosetta	1/20000	7.96	0	0	2/20000	3.38
ICM	23/10000	6.58	0	0	41/10000	2.76

VHL-SMARCA2^{BD}	Number of docked poses with C-alpha RMSD < 10Å from crystal structure	Lowest C-alpha RMSD(Å) of docked pose	Number of docked poses with ligand RMSD < 5Å from crystal structure			
			Top 10 scoring poses	Top 50 scoring poses	Total scoring poses	Lowest ligand RMSD(Å)
Haddock	7/40	4.95	0	15	15/40	2.02
Rosetta	160/20000	4.52	0	2	331/20000	1.79
ICM	30/10000	6.42	0	2	122/10000	2.19

CONCLUSION:

In summary, I find that all 3 methods generate a number of predicted structures, including the experimental one, but none of the methods accurately rank the experimental structure at the top (though Haddock results are more enriched in experimental structures). If the experimental structure is the only acceptable E3-target interface, this is a problem, as using the wrong E3-target complex for subsequent PROTAC design is bound to fail. However, Nowak et. al. has shown that different PROTACs can lead to difference protein-protein interfaces for the same E3-target pair.[6] It is therefore possible that the best scoring poses (even if different from experimental poses) are indeed useful for PROTAC design. Something I will test in the next step.

References

1. Drummond et al. *In Silico* modeling of PROTAC-medicated ternary complexes: Validation and application. (2019) *J. Chem. Inf. Model.* 59:1634-1644
2. Van Zundert et al. "The HADDOCK2.2 webserver: User-friendly integrative modeling of biomolecular complexes." (2016) *J. Mol. Biol.* 428:720-725
3. Gray et al. Protein-protein docking with simultaneous optimization of rigid body displacement and side-chain conformations. (2003) *J. Mol. Biol.* 331:281–299
4. Totrov et al. Detailed ab initio prediction of lysozyme-antibody complex with 1.6 Å accuracy. (1994) *Nat. Struct. Biol.* 1:259–263
5. Mendez et al. Assessment of blind predictions of protein-protein interactions: Current status of docking methods. (2003) *Proteins: Structure, Function, and Genetics* 52:51–67
6. Nowak et al. Plasticity in binding confers selectivity in ligand-induced protein degradation. (2018) *Nat. Chem. Biol.* 14:706–714