PROTAC virtual screening: A retrospective screening exercise

BACKGROUND

In my previous post [https://openlabnotebooks.org/protein-protein-docking-for-protac-discovery/], I showed that HADDOCK was the best protein-protein docking tool among those I tested to predict how E3 ligases interact with their protein substrates. Here, I ask whether docking virtual libraries of PROTAC candidates to these E3 ligase – substrate protein interfaces can be used to predict which PROTACs are active.

METHOD

The virtual screening method that I am using is as follows (Figure 1):

1- Use HADDOCK to dock the E3 ligase to the protein target and keep the ~40 best solutions (typically grouped in 5 to 9 clusters of structurally related binding poses)
2- For each PROTAC candidate, tether the two chemical handles to their corresponding ligands in the protein-protein complexes generated by HADDOCK, and minimize the energy of the PROTAC while ignoring the proteins. This produces a selection of conformationally acceptable PROTACs: SELECTION 1
3- Start from the conformationally acceptable PROTACs (SELECTION 1) and further minimize their energy, this time accounting for the surrounding proteins (PROTACS and side-chains within 5Å are kept flexible)

Step 2 and 3 are conducted 3 times independently, and all results are merged, so, for 10 PROTAC candidates, we will have 10 molecules x 40 prot-prot-interfaces x 3 repeats = 1200 binding poses.

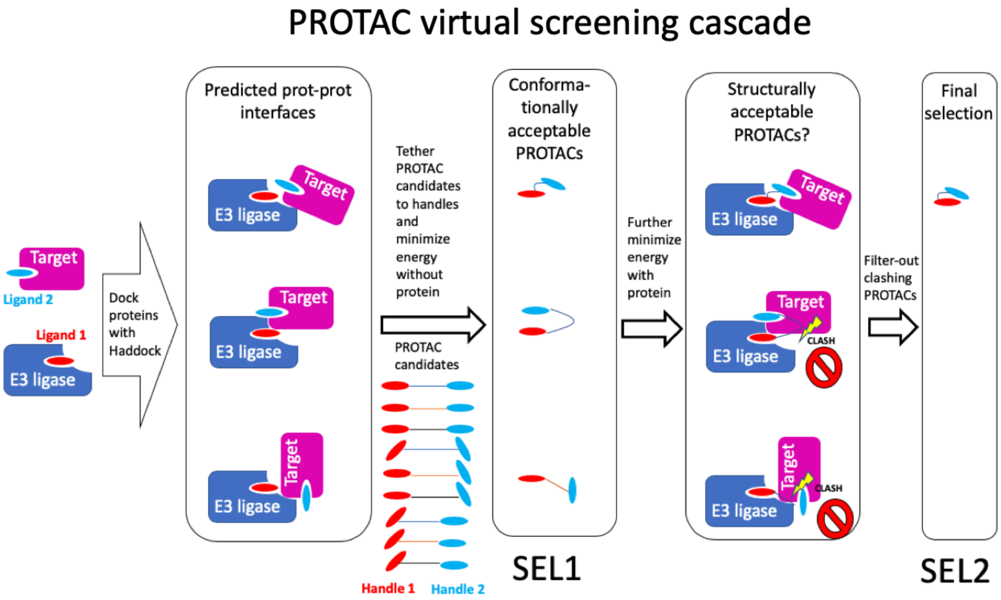

PROTAC virtual screening cascade
Target
Ligand 2
Ligand 1
E3 ligase
Dock proteins with Haddock
Predicted prot-prot interfaces
E3 ligase
Target
E3 ligase
Target
E3 ligase
Target
Tether PROTAC candidates to handles and minimize energy without protein
PROTAC candidates
Handle 1
Handle 2
Conforma-tionally acceptable PROTACs
SEL1
Further minimize energy with protein
Structurally acceptable PROTACs?
E3 ligase
Target
E3 ligase
Target
CLASH
E3 ligase
Target
CLASH
Filter-out clashing PROTACs
Final selection
SEL2

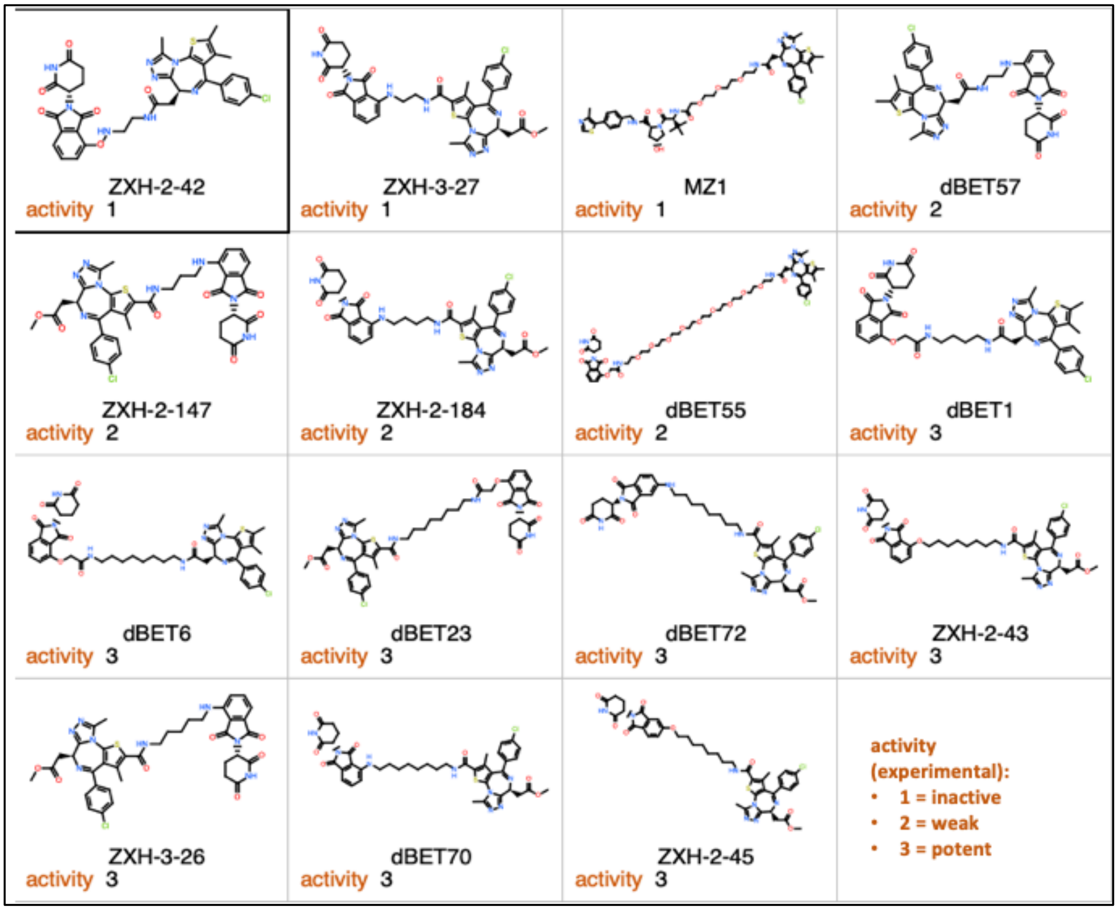
Figure 1: Overview of the virtual PROTAC screen pipeline.

This screening pipeline is implemented in ICM (Molsoft, San Diego), and the corresponding script is attached to this Zenodo post.

RESULTS

TEST CASE 1: CRBN – BRD4^{BD1}

The first test case is a work from Nowak *et al*. where they screened 15 PROTACs for their ability to induce ubiquitination of the first bromodomain of BRD4 by CRBN.[1] I classified the compounds as potent (8 molecules), weak (4), or inactive (3) (Figure 2).



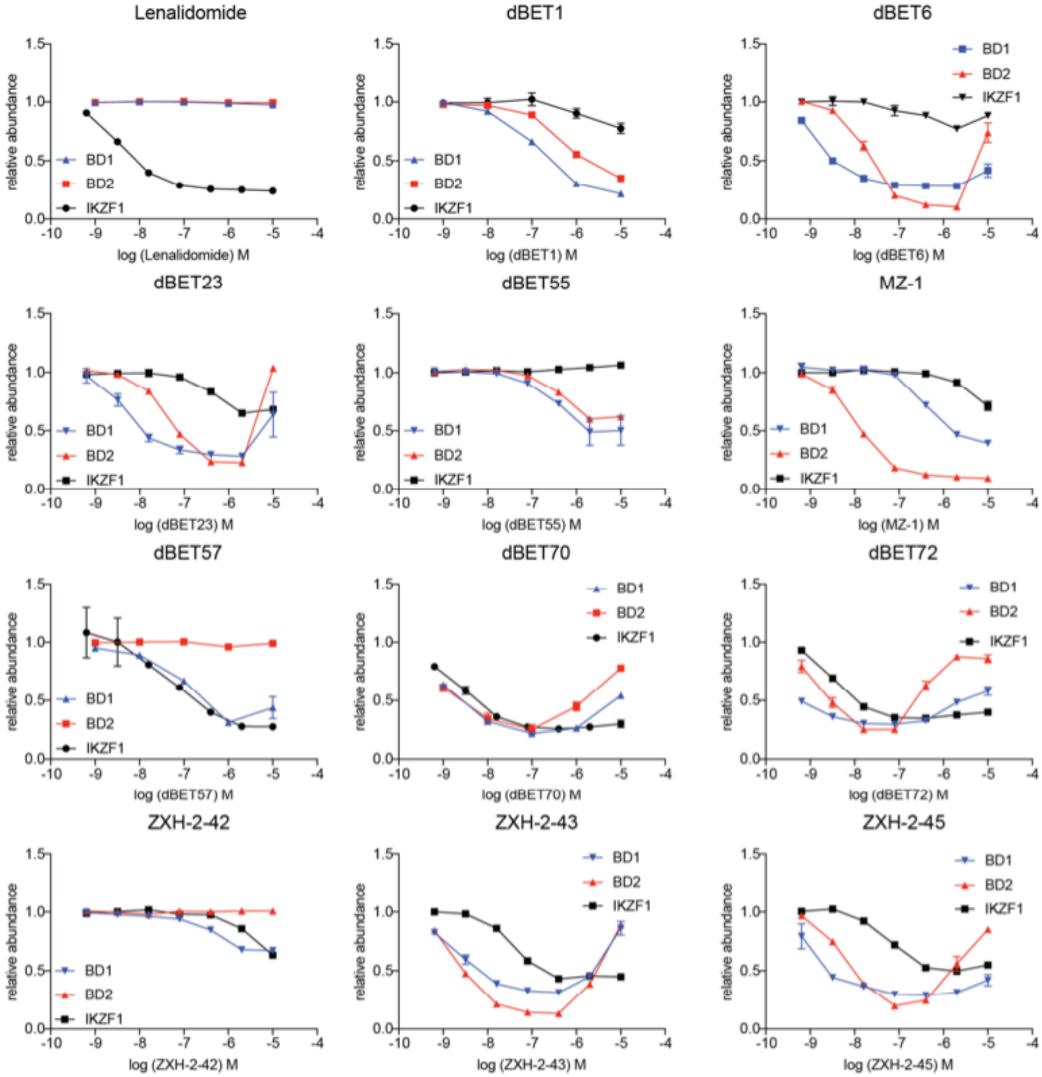
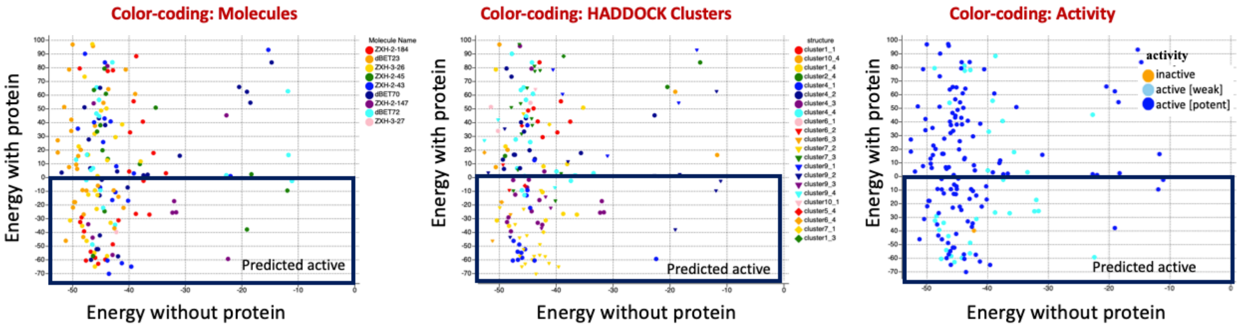


Figure 2: Classification and effect of CRBN-BRD4^{BD1} PROTAC candidates from Nowak *et al*. 1: inactive; 2: weak; 3: potent.

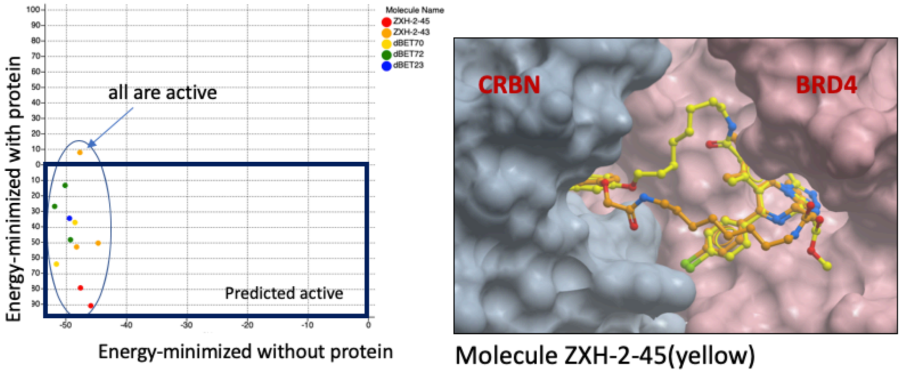
A first observation is that the PROTACs with the lowest energy levels after STEP2 (i.e. energy minimization without protein) are all active (blue) (Figure 3). Plotting the energy with vs energy without protein (STEP3 vs STEP2) shows that many conformations that passed STEP2 are filtered-out at STEP3. In this case, this does not affect the hit rate, as compounds for which conformation X was filtered-out are still selected in STEP3 in another conformational state. The hit rates after STEPS 3 and 4 is the same: 88%. This seems great, but the problem is that there were very few inactive compounds in the source library to begin with, and the enrichment is therefore very low (1.1 fold). This is a recurrent problem in this and other test cases: our top scoring compounds are systematically active, but we don't know for sure whether this is because the virtual screening protocol works great, or just luck: we need more data on inactive compounds!



Method	Number of Unique Cluster	Number of Active PROTAC conformations	Number of Inactive PROTAC conformations	Total Number of Poses	Number of Active PROTAC	Number of Inactive PROTAC	Hit Rate (%)	Enrichment (fold)
Selection 1	8	161	1	162	8	1	88.88	1.11
Selection 2	6	70	1	71	8	1	88.88	1.11
Top 3	6	34	---	34	3	---	100.00	

Figure 3: Energy distribution and summary table of the CRBN-BRD4^{BD1} virtual screening exercise.

In the end, we find that 6 out of 40 protein interfaces produced by HADDOCK are compatible with active PROTACs. Some of these interfaces were close to the crystal structure (PDB code 6boy), but none were perfect match. To see whether the binding conformation of active PROTACs could be retrieved when using the crystal structure of the CRBN-BRD4^{BD1} complex, we repeated the virtual screening procedure, using only the structure 6boy (Figures 4). Again, active PROTACs are enriched after STEP2, while STEP3 makes no difference: the 5 PROTACs that were predicted active are all experimentally active. This is 3 less than when we used HADDOCK binding poses. Maybe this reflects the fact that different PROTACs can induce different CRBN-BRD4^{BD1} binding poses, as shown by Nowak *et al*.



Molecule ZXH-2-45(yellow) prediction with **6boy**:dBET6 (Orange)

Method	Number of Unique Cluster	Number of Active PROTAC conformations	Number of Inactive PROTAC conformations	Total Number of Poses	Number of Active PROTAC	Number of Inactive PROTAC	Hit Rate (%)	Enrichment (fold)
Selection 1	---	11	0	11	5	0	100	1.25
Selection 2	---	10	0	10	5	0	100	1.25
Top 3	---	6	---	6	3	---	100	

Figure 4: Virtual screening results using the crystal structure of the CRBN-BRD4^{BD1} complex (6boy) instead of protein-protein docking interfaces generated by HADDOCK.

The PROTAC crystallized in the 6boy structure (**dBET6**) is not one of our 5 compounds predicted active. Indeed, its docked conformation is different from that observed in the crystal structure (Figure 6). Clearly, there is room for progress here.

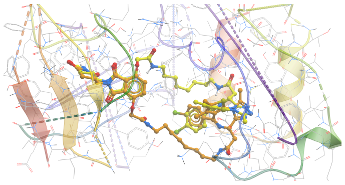


Figure 5: The crystallized PROTAC dBET6 was not selected, and its predicted binding mode (yellow) was different from the crystal structure 6boy (orange).

TEST CASE 2: CRBN – BRD4^{BD2}

The second test case is from the same paper,[1] where CRBN is this time recruited to ubiquitylate the second bromodomain of BRD4. This test case includes more inactive compounds (Figure 6), which is a good thing to evaluate the validity of our virtual screening pipeline.

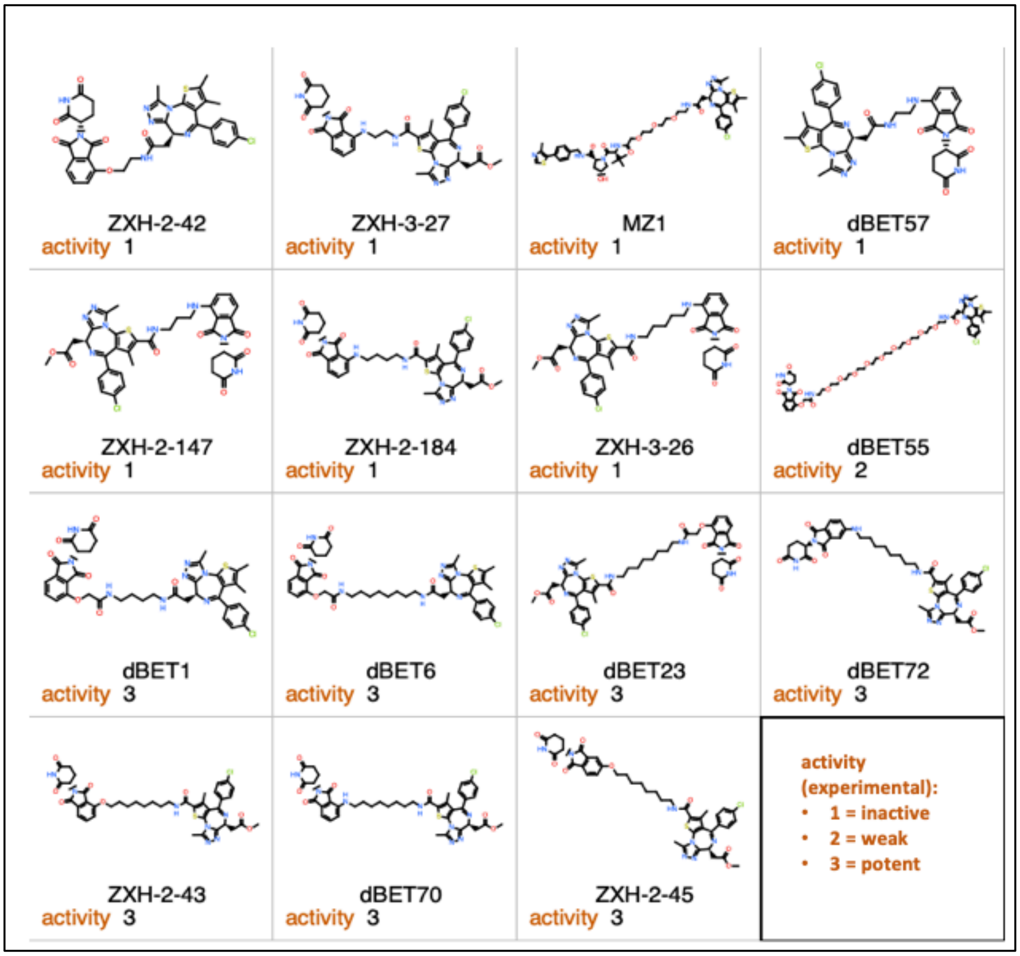
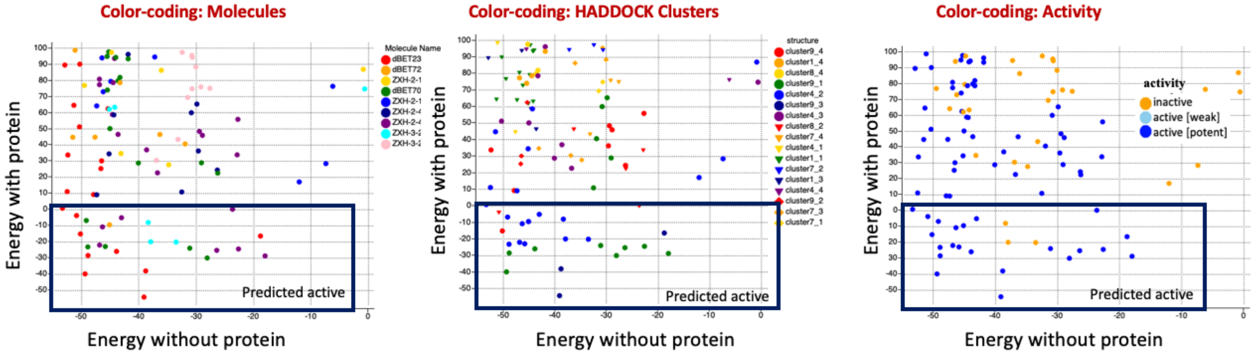


Figure 6: Structure and experimental activity of PROTACs recruiting CRBN and targeting BRD1^{BD2}.

As previously, after STEP2, low-energy compounds are enriched in experimentally active PROTACs, and high-energy ones are enriched in inactive molecules (Figure 8). Good! But there are still a number of false positives among the low-energy molecules, and the hit rate is only 55% (Figure 7). Adding the proteins in the energy-minimization procedure (STEP3) filters-out the majority of false positives, and the enrichment rate goes up to 80%, while the enrichment in actives compared with the source library is 1.5-fold (Figure 7).



Method	Number of Unique Cluster	Number of Active PROTAC conformations	Number of Inactive PROTAC conformations	Total Number of Poses	Number of Active PROTAC	Number of Inactive PROTAC	Hit Rate (%)	Enrichment (fold)
Selection 1	5	76	24	100	5	4	55.55	1.04
Selection 2	3	20	3	23	4	1	80.00	1.5
Top 3	3	19	---	19	3	---	100.00	

Figure 7: Energy distribution and summary table of the CRBN-BRD4^{BD2} virtual screening exercise.

This is encouraging, but again, we would need more inactive compounds in the source library to verify that enrichment rates can be significantly higher. As previously, our top 3 scoring PROTACs are all active.

TEST CASE 3: VHL – EED

While the E3 ligase in the first two test cases was CRBN, here I am using data from Potjewyd *et al.* on PROTACs recruiting VHL to degrade EED (and associated components of the PRC2 complex).[2] There are only 6 compounds, 4 active and 2 inactive, but this is still valuable information (Figure 8).

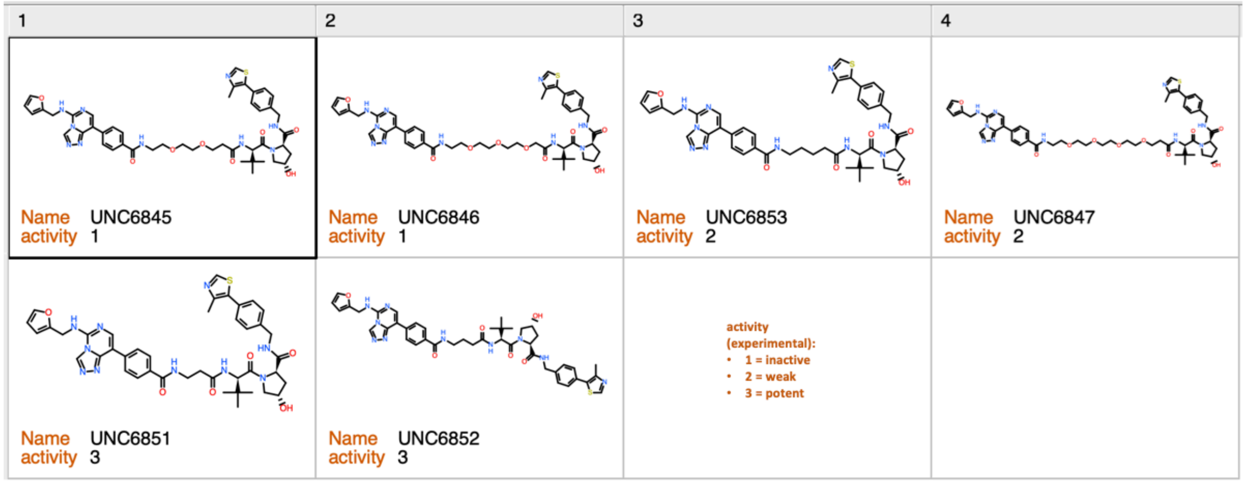
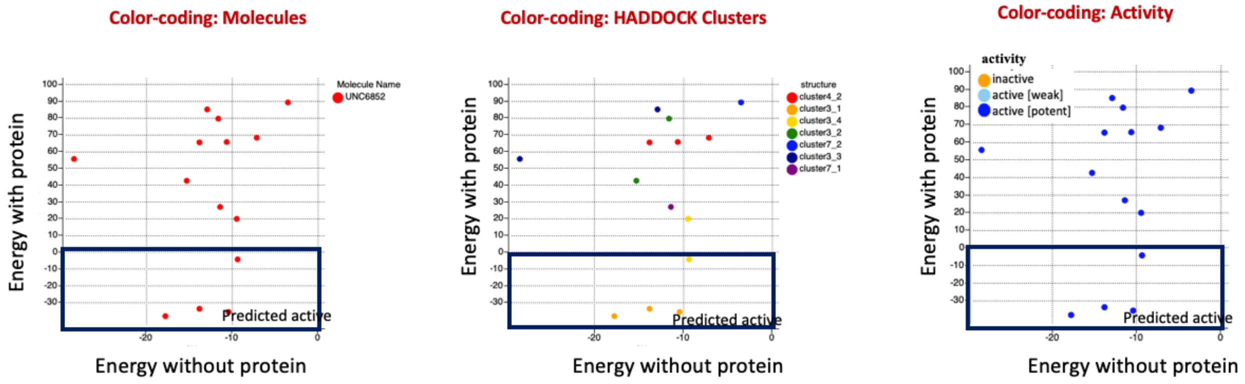


Figure 8: Structure and activity of VHL-EED PROTAC candidates

Once again, I find that the lowest energy binding poses after STEP 2 are all for active molecules (Figure 9), but a closer look reveals that they are all for the same molecule: only one compound (which happens to be experimentally active) is conformationally stable (energy < 0 in the absence of protein when extremities are tethered to positions imposed by docking of EED to VHL). The hit rate is 100% (1 selected – 1 active), but 75% of active molecules were filtered-out. This single hit passes the second filter: its energy level is still negative when incorporating the proteins VHL and EED in the energy minimization simulation, indicating that the molecule can not only be positioned in a way where its functional groups are matching the VHL and EED ligands in the context of the docked VHL-EED binding poses, but that in doing so, neither the chemical handles, nor the linker are clashing with the proteins (Figure 9). There is no experimental structure of the complex at the moment, so, no way to know whether our predicted binding poses are accurate.

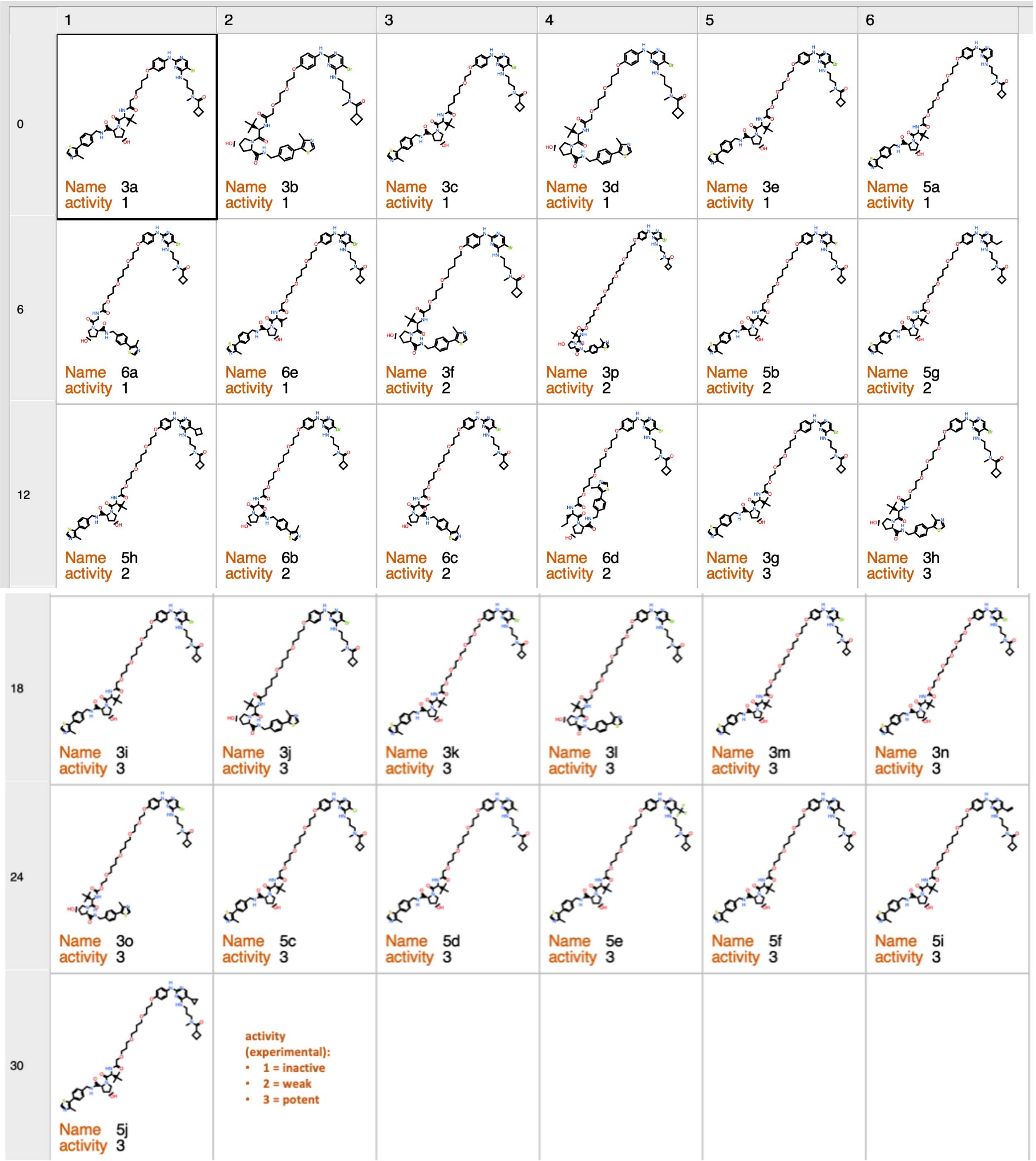


Method	Number of Unique Cluster	Number of Active PROTAC conformations	Number of Inactive PROTAC conformations	Total Number of Poses	Number of Active PROTAC	Number of Inactive PROTAC	Hit Rate (%)	Enrichment (fold)
Selection 1	3	14	0	14	1	0	100	1.33
Selection 2	1	4	0	4	1	0	100	1.33
Top 3	2	4	--	4	1	--	100	

Figure 9: Energy distribution and summary table of the VHL-EED virtual screening exercise

TEST CASE 4: VHL – TBK1

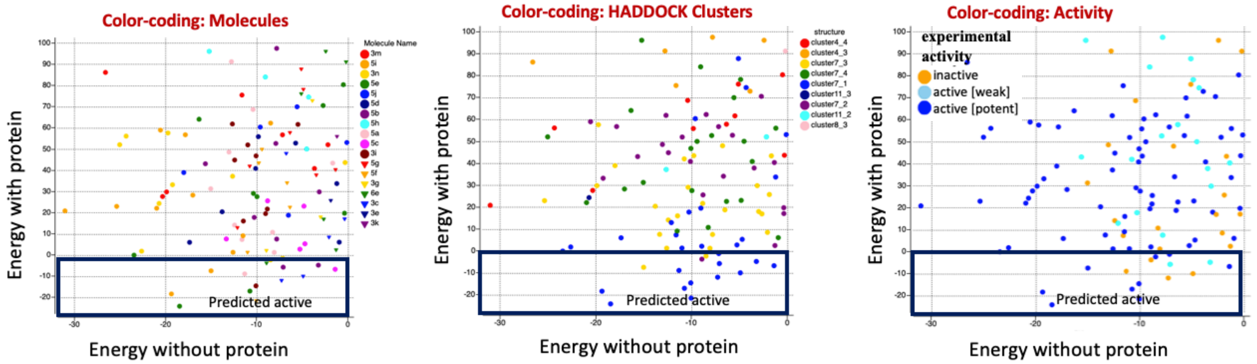
This data is from Crew *et al,*[3] who report the activity of 31 compounds designed to induce the dimerization of VHL and TBK1 (Figure 10). About 25% are inactive.



	1	2	3	4	5	6
0	Name 3a activity 1	Name 3b activity 1	Name 3c activity 1	Name 3d activity 1	Name 3e activity 1	Name 5a activity 1
6	Name 6a activity 1	Name 6e activity 1	Name 3f activity 2	Name 3p activity 2	Name 5b activity 2	Name 5g activity 2
12	Name 5h activity 2	Name 6b activity 2	Name 6c activity 2	Name 6d activity 2	Name 3g activity 3	Name 3h activity 3
18	Name 3i activity 3	Name 3j activity 3	Name 3k activity 3	Name 3l activity 3	Name 3m activity 3	Name 3n activity 3
24	Name 3o activity 3	Name 5c activity 3	Name 5d activity 3	Name 5e activity 3	Name 5f activity 3	Name 5i activity 3
30	Name 5j activity 3	activity (experimental): • 1 = inactive • 2 = weak • 3 = potent				

Figure 10: Structure and activity of VHL-TBK1 PROTAC candidates

While the top 3 virtual hits with the lowest energy level after STEP 3 are all experimentally active, the ensemble of compounds with negative energies contains a number of inactive molecules, resulting in a hit rate of 72% and a really bad "enrichment" rate of 0.98 fold between the source library and the final selection. This could be due to bad protein-protein docking poses generated by HADDOCK, or bad PROTAC conformational sampling by ICM. The former seems more likely, as there is less reason that ICM would do worse with these PROTACS than with others, while the accuracy of protein-protein docking is expected to vary significantly from one system to another.



Method	Number of Unique Cluster	Number of Active PROTAC conformations	Number of Inactive PROTAC conformations	Total Number of Poses	Number of Active PROTAC	Number of Inactive PROTAC	Hit Rate (%)	Enrichment (fold)
Selection 1	4	93	22	115	14	4	77.77	1.04
Selection 2	1	12	4	16	8	3	72.72	0.98
Top 3	1	6	---	6	3	---	100.00	

Figure 11: Energy distribution and summary table of the VHL-TBK1 virtual screening exercise

TEST CASE 5: CRBN – CDK6

The last test case for which I found a reasonable number of compounds (including inactive molecules) in the literature is a collection of PROTAC candidates recruiting CRBN to CDK6,[4] with 11 active and 2 inactive PROTACs (Figure 12). Again, the number of inactives is barely sufficient to evaluate the efficiency of my virtual screening exercise.

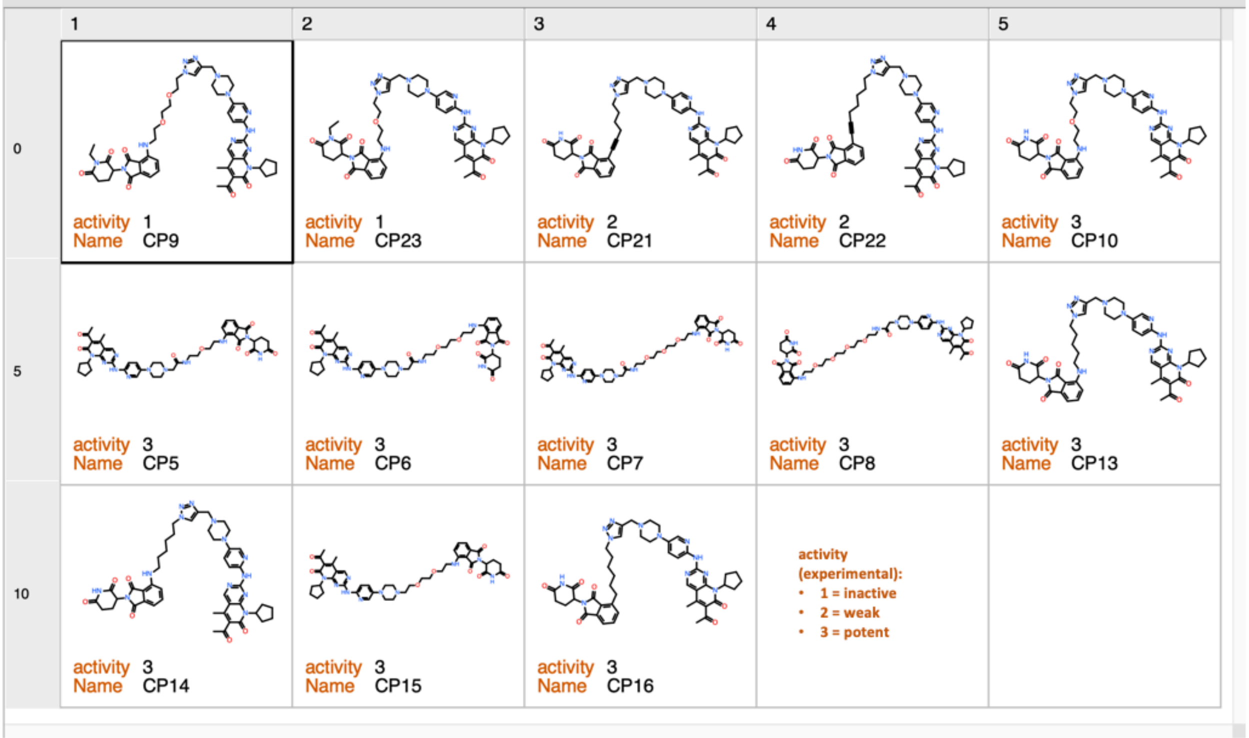
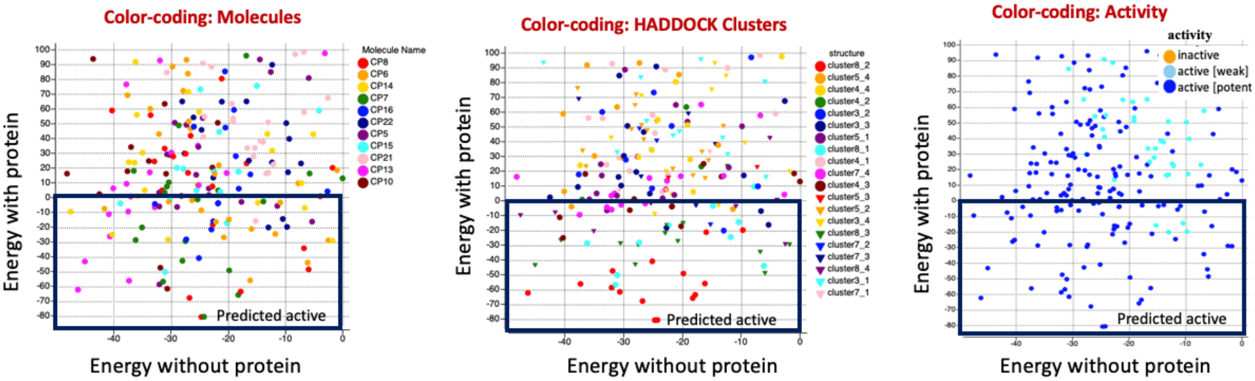


Figure 12: Structure and activity of CRBN-CDK6 PROTAC candidates

The 2 inactive compounds were effectively filtered-out by both STEP 2 and STEP 3 and all 11 active PROTACs were selected at the end of STEP 3, so, the virtual screen results were absolutely perfect (Figure 13). The fact that the only 2 experimentally inactive compounds were the only two compounds filtered-out by the virtual screen could be pure luck, but this is still rather encouraging result.



Method	Number of Unique Cluster	Number of Active PROTAC conformations	Number of Inactive PROTAC conformations	Total Number of Poses	Number of Active PROTAC	Number of Inactive PROTAC	Hit Rate (%)	Enrichment (fold)
Selection 1	5	204	0	204	11	0	100	1.18
Selection 2	5	64	0	64	11	0	100	1.18
Top 3	5	23	0	23	3	---	100	

Figure 13: Energy distribution and summary table of the CRBN-CDK6 virtual screening exercise

CONCLUSION

Two general observations can be made on our retrospective attempts to select PROTACs by virtual screening. First: in all test cases, the starting library of compounds for which we have experimental data is highly enriched in experimentally active molecules. Picking compounds at random, one has good chances of selecting active PROTACs. Two: in all 5 test cases, our top 3 scoring PROTACs out of the virtual screening pipeline are systematically active. In four out of 5 test cases, the library selected by virtual screening is enriched in active PROTACs, compared with the starting library, which is better than random. At this point, it is too early to know whether this is pure luck, or whether this reflects a valid virtual screening strategy.

Considering how unreliable virtual screening of conventional small molecules to a crystal structure is (one can expect 95 inactives out of 100 compounds selected virtually when the stars are aligned), it seems almost ridiculous to dock PROTACs (large and floppy molecules) to predicted protein-protein interfaces. And yet, it is not unreasonable to expect that the multiple protein-protein docking poses generated by HADDOCK (or other protein-protein docking tools) while being all suboptimal (after all, these complexes do not form naturally), are structurally sufficiently sound to be stabilized by a PROTAC. Indeed, Nowak *et al.* clearly showed that different PROTACs could induce different binding poses between CRBN and BRD4^{BD11} (Nowak *et al. Nat. Chem. Biol.* 14:706-714). If we accept this hypothesis, the PROTAC docking step should actually be more efficient than a regular virtual screening campaign, as the orientation of the two functional ends of the PROTACs are imposed by the protein complex, which considerably reduces the conformational space available to the PROTAC, and the chances to predict a wrong PROTAC binding pose. There is still room to get things wrong, but not as much as one would expect considering the complexity of the system.

Finally, an important factor I ignored so far is that some PROTAC candidates may induce the formation of a ternary complex in a biophysical assay, but still be inactive because they do not cross cell membranes. I will probably need to incorporate physico-chemical filters to account for this, following-up on work by Maple *et al.*[5] and Edmondson et al.[6] Another factor even more complex is that some compounds may induce the formation of a ternary structure that is not compatible with efficient ubiquitylation of the target.

Two things are sure: we need more experimentally inactive PROTACs, and there is room for improvement.

References:

1. Nowak R., *et al.,* “Plasticity in binding confers selectivity in ligand-induced protein degradation.” *Nat. Chem. Biol.,* **2018,** 14:706-714.
2. Potjewyd F., *et al., “Degradation of Polycomb Repressive Complex 2 with an EED-targeted Bivalent Chemical Degrader.” doi: https://doi.org/10.1101/676965*

3. *Crew AP., et al., "Identification and Characterization of Von Hippel-Lindau-Recruiting Proteolysis Targeting Chimeras (PROTACs) of TANK-Binding Kinase 1" J. Med. Chem.,* **2018,** 61:583-598.
4. Su S., *et al., "Potent and Preferential Degradation of CDK6 via Proteolysis Targeting Chimera Degraders." J. Med. Chem.,* **2019,** 62:7575-7582.
5. Maple HJ., *et al.,* "Developing degraders: principles and perspectives on design and chemical space." *MedChemComm.,* **2019,** DOI: 10.1039/c9md00272c
6. Edmondson SD., *et al.,* "Proteolysis targeting chimeras (PROTACs) in 'beyond rule-of-five' chemical space: Recent progress and future challenges." *Bioorg. Med. Chem. Lett.,* **2019,** 13:1555-1564