Rational Prediction of PROTAC-compatible Protein-Protein Interfaces by Molecular Docking

Gilberto P. Pereira[1], Brian Jiménez-García[2], Riccardo Pellarin[1], Guillaume Launay[1], Sangwook Wu[3,4#], Juliette Martin[1,#], Paulo C. T. Souza[1,#]

[1]Molecular Microbiology and Structural Biochemistry (MMSB, UMR 5086), CNRS & University of Lyon, 7 Passage du Vercors, 69007, Lyon, France;

[2]Zymvol Biomodeling, Barcelona, Spain;

[3]PharmCADD, Busan, 48792, Republic of Korea;

[4]Department of Physics, Pukyong National University, Busan, 48513, Republic of Korea.

Authors for correspondence:
Paulo C. T. Souza. Email: paulo.telles-de-souza@ibcp.fr

Juliette Martin Email: juliette.martin@ibcp.fr

Sangwook Wu Email: s.wu@pharmcadd.com

Abstract

Proteolysis targeting chimeras (PROTACS) are ligands that can mediate interaction between a protein target and a E3 ligase, forming a ternary complex bound to the ubiquitination machinery, leading to target degradation. This technology has become an exciting new avenue for therapeutic development, with currently two PROTACS in clinical trials targeting cancer. Nonetheless, several PROTAC drug discovery campaigns still rely on serendipity. Here, we describe a general and efficient protocol which combines restraint-based LightDock, energy-based rescoring and minimal solvent-accessible surface distance filtering to produce PROTAC-compatible PPIs. Benchmarking our protocol using a manually curated dataset of 16 ternary complex crystals, accuracies of 94% and 70% were achieved on the

redocking and realistic docking experiments starting from unbound protein structures, respectively, evaluated using the CAPRI standards and DockQ. Our protocol is both accurate and computationally efficient, with potential to accelerate PROTAC drug design campaigns, particularly when the ternary complex or PROTAC are unknown.

Introduction

Protein-protein Interactions (PPI) are fundamental driving forces behind the function of large macromolecular assemblies, impacting nearly all cellular processes[1,2]. Across the literature, examples of fundamental cellular processes ruled by PPIs showcase their importance[3]. For example, PPIs are critical to the functioning of systems such as ATP synthase[4], ion channels[5] or the ubiquitination machinery[6,7]. Among these, there is a clear distinction between obligate PPIs, i.e those in which the unbound monomers are unstable in vivo, and non-obligate PPIs[8]. Within non-obligate PPIs there exist transient and permanent complexes, with the distinction being ascribed to a matter of complex stability and lifetime[8]. Furthermore, transient PPIs can either be of high affinity or low affinity nature, which is one of key determinants of the lifetime of the complex formed by these proteins[9]. While it is undeniable that PPIs of all types are fundamental in the majority of cellular processes, recently weak, transient, and ligand-mediated PPIs have gained attention. In particular, the field of PPIs mediated by Proteolysis Targeting Chimeras (PROTACS) has attracted significant attention and funding[10]. While the technology is more than 20 years old[11], the base concept remains: a heterobifunctional PPI promoter composed of a ligand that binds to a protein target (warhead), a ligand that binds an E3 ligase (recruiter), and a flexible linker connecting the two ligands, is used to bring into close proximity a target and the ubiquitination machinery[11–13]. By bringing the target close to the ligase, enabling the formation of a ternary complex, the PROTAC molecule facilitates the transfer of ubiquitin to the target, leading to poly-ubiquitination and subsequent target degradation by the proteasome[13,14]. Thus, PROTAC-based approaches have undeniable advantages with respect to small-molecule therapies, such as requiring lower compound concentration, decreased toxicity risks and the ability to target non-druggable proteins[15]

In the last five years, PROTAC-based approaches were pursued to some success. The first PROTAC-containing ternary crystal structure was published by the Ciulli group in 2017, where the MZ1 ligand was shown to promote bromodomain 4 (BRD4) degradation when

bound to the von-Hippel Landau (VHL) E3 ligase, a system implicated in cancer disease[15,16]. Ever since, PROTAC molecules have been shown to induce degradation of several kinases, such as BTK[17] or FAK[18], fusion proteins[19,20] or growth factor proteins[20,21]. Using this technology, Arvinas and collaborators were able to target the androgen receptor and the estrogen receptor, targets for prostate and breast cancer, respectively[22,23]. Currently, two PROTACs (ARV-110 and ARV-471) are undergoing phase 1/2 clinical trials[13]. Additionally, there have been reports of PROTAC molecules being useful in treating immune disorders, viral infections and neurodegenerative disorders[20]. A compilation of bioactive PROTAC molecules, along with their description in terms of physicochemical properties was recently published and is expected to help drive PROTAC drug design campaigns[24].

Regardless of these successes, there are still few experimentally determined structures for E3 Ligase-PROTAC-Target systems deposited in the Protein Data Bank (PDB)[25]. As such, more and more groups are turning to computational methodologies to predict the structure of ternary complexes[26–30]. One of the most popular tools for PPI prediction is protein-protein docking. In molecular docking algorithms, the prediction is typically obtained after two critical steps: sampling of the conformational landscape of the protein-protein complex and scoring of the near-native poses obtained[31]. Due to the large number of degrees of freedom within the system, including all bonds, angles and dihedral for each protein, the sampling stage becomes prohibitively expensive. Hence, most often PPI docking protocols rely on rigid-body approximations, *a priori* disregarding conformational flexibility or approximating it in a rough manner[32–38]. In the particular case of PROTAC-mediated ternary complexes, standard molecular docking approaches alone may require careful set-up and additional post-processing steps. Indeed, recent approaches for PROTAC ternary complex prediction employ molecular docking approaches combined with several steps of refinement to include some degree of flexibility at the level of the side chains and explicit inclusion of PROTAC ligand[28,29]. For example, the study by Weng and co-workers combines a local docking grid

searching in rotational space algorithm (FRODOCK)[39] with filtering steps, full-PROTAC construction, ternary complex rescoring and a final optimization using RosettaDock[40–42].

One alternative route was proposed by Dixon and co-workers[43], where a combination of molecular modeling, SAXS, and HDX-MS experiments led to the determination of physiologically relevant ternary complexes for a VHL system. Harnessing this information is key to the development of new therapeutic avenues using PROTACs. In contrast with the study from Weng and co-workers[29], which aims at reproducing the crystallographic PROTAC-containing ternary complex, one of the main points from this study is that the free energy landscape of these systems is not characterized by a single lower energy conformation but instead by several, interconvertible, low energy minima, highlighting the complexity and flexibility of PROTAC-mediated biomolecular systems[43]. Nonetheless, this procedure would work only if there is experimental data available. Finally, given the immense success that AlphaFold2 (AF2)[44] and other machine learning-based methods have shown in the last CASP14 and CASP15 editions[45,46], some researchers have pointed out that in the absence of experimental structures, future PROTAC-based drug design projects will use AF2 predicted structures[47], which can either be structures of the ligase and target monomers or the ligase-target complex.

In this study, we present an approach to generate ligase-target binding modes compatible with crystallographic-solved PROTAC-mediated ternary complexes. Our final aim is producing a workflow which facilitates the initial stage of PROTAC *in silico* design by generating realistic starting PPI structures. The present approach assumes that the only information available *a priori* are the small molecule ligand-bound monomeric forms of the ligand and target proteins involved in the ternary complex. Our approach is tested on a manually curated dataset of 16 ternary complexes extracted from the PDB and is based on the predictions provided by LightDock[36,38,48], a general molecular docking tool with the ability of harnessing structural information in the form of restraints. LightDock is combined with energy rescoring afforded by VoroMQA[49] and a filtering step based on either the distance

between ligands center-of-geometry (COG) or the minimal solvent-accessible surface distance (SASD) between the anchor atoms of the ligands bound to the ligase and target proteins in predicted PPI poses, computed using Jwalk[50] and calibrated on the benchmark dataset. The filtering step is necessary to remove physically unreasonable poses in which either one or both small-molecule ligand binding sites are inaccessible, acting as an implicit "PROTACability" measure for the ligase-target interface. The method is general as it only requires the user to have the ligand-bound monomeric forms of the ligase and the target, since contact residue restraints are automatically generated based on the warhead and recruiters' COG positions within their binding sites, notably without the need of knowing a suitable PROTAC molecule. We show that restrained-LightDock simulations produce PPIs which are compatible with crystallographic ternary complex binding modes for nearly all systems under scrutiny. We further compare the performance of our method with AlphaFold2-Multimer[44,51], highlighting critical limitations in the machine-learning based method for this kind of application.

Experimental section

Dataset selection and curation

Sixteen PROTAC ternary complex crystal structures were extracted from the PDB[25], selected based on three criteria: (1) that the crystal structure contained an E3 ligase, a ligand molecule, and a target; (2) that the crystal structure resolution was below 4 Å and (3) that the ligand was a PROTAC and not a molecular glue. The dataset is composed of three types of E3 ligases (9 VHL[16,52–57], 5 Cereblon[58–61] and 2 cIAP[62]), multiple different PROTAC molecules and a diverse set of protein targets, from kinases to bromodomains or proteins implicated in DNA repair (as WRD5). For the realistic docking case scenario, unbound monomers were selected to be different from the initial 16 crystal structures in terms of interface side-chain packing[52,62–69]. All protein termini were capped using the psfgen tool in VMD[70], residue names were converted to standard names, and hydrogens were removed prior to molecular docking experiments. A listing of the structures used in the redocking and realistic experiments is given on **Tables S1** and **S2**.

Molecular Docking and refinement protocol

Docking experiments were carried out using version 0.9.2 (https://github.com/lightdock/lightdock/releases/tag/0.9.2) of the LightDock (LD) framework[36,38]. LightDock is a multiscale framework for the 3D determination of binary macromolecular complexes based on the Glowworm Swarm Optimization (GSO) algorithm[71]. The sampling in LightDock is driven by the *swarm* approach, which relies on the metaphor that, in nature, *glowworms* (which represent ligand, or protein target, poses) feel attracted to each other depending on the amount of emitted light (scoring, energetic value of a docking pose). In this way, the docking poses, which constitute the *swarm* of *glowworms* in LightDock, are optimized towards the energetically more favorable ones through the

translational, rotational, and, if the option is selected, some degree of protein flexibility is considered through an optimization of anisotropic network model (ANM)[36,38]. Every swarm represents an independent docking simulation, and the number of swarms is automatically determined depending on the receptor surface nature, ensuring full sampling coverage of the protein complex surface. If contact residue-restraints are provided, swarms are filtered according to their compatibility with these restraints, focusing sampling to the region of interest described by the restraints[36,38]. The number of glowworms per swarm was set to 200 (which is the default in the protocol), and 50 docking iterations were carried out using the DFIRE scoring function[38,72,73].

In the docking simulations, we define as receptor the E3 ligase and as ligand the protein target to be degraded. The recruiter and warhead molecules were only used to compute the residues acting as contact residue restraints. Thus, these small molecules are not explicitly present in the docking simulation. The experiments were carried out for two cases: (1) redocking (or bound) and (2) realistic (unbound) docking, starting from the monomers within crystal ternary complexes and from independent monomeric forms of the proteins found in the PDB, respectively. Furthermore, in both redocking and realistic docking experiments, we explored the enforcing of contact residue restraints to help drive the docking towards the PPI. Contact residue restraints were defined by identifying residues in the ligand binding site whose C-alpha atoms were within a given distance (4, 6 or 8 Å) from any heavy atom of the ligands bound either to the E3 ligase or the target protein. These residues were determined using MDAnalysis[74,75]. We excluded residues whose relative Solvent Accessible Surface Area (SASA) was below 25%, SASA was computed using naccess[76]. From each docking experiment, the top 200 poses were extracted according to the LightDock predicted ranking.
To improve protocol accuracy, a rescoring step was pursued using the energy function from VoroMQA[49], which was sufficient in the case of the redocking experiment, and then filtered according to either the COG distance between the recruiter and warhead ligands or the existence and length of the SASD between small-molecule anchor atoms[50] for the realistic,

or unbound monomer, docking experiment. The former was computed using MDAnalysis[74,75] while the latter was computed using Jwalk[50], calibrated using the predicted PPIs and the reference structures in the benchmark dataset with the aim of recovering as many acceptable solutions as possible while also encompassing at least 90% of the SASDs computed on the crystal structures. The filtering step shifts all solutions which do not pass the criteria to the bottom of the ranking, maintaining the order obtained from the energy rescoring. For the SASD calculations, the spatial position of the anchor atoms used to build the path between the ligase and target proteins were determined from the ligands position in the reference crystal structures.

Docking accuracy

After each experiment, the accuracy of the docking protocol was evaluated using the CAPRI standards[77,78] and the DockQ score[79]. These criteria employ three metrics between docking poses and crystallographic complexes, classifying docking solutions as incorrect, acceptable, medium or high quality. The three similarity measures are the fraction of conserved native contacts (Fnat), the interface root mean squared deviation (RMSD) and the ligand RMSD, where ligand refers to the target protein, which were computed in this study using C-alphas. As PROTAC-mediated ternary complexes are more and more suspected to be flexible[43], we consider that any solution with acceptable or higher quality is a reasonable solution given that the binding interface between the protein partners is conserved. The reference structures used were the PROTAC ternary complex crystals for the redocking and realistic docking protocols respectively. Thus, accuracy was measured by evaluating the presence of solutions with at least acceptable quality in the top 1, 5, 10, 20, 50 or 100 protein-protein docking poses, using the CAPRI and DockQ criteria. Acceptable solutions thus adhere to either (i) a DockQ score ≥ 0.23, or (ii) a fraction of native contact (Fnat) >=10% and either ligand RMSD ≤ 10 Å or interface RMSD ≤ 4 Å[79]. A pose was deemed

acceptable if it passed either the CAPRI or the DockQ criteria.

AlphaFold-multimer structure predictions

Calculations using AlphaFold2-multimer[44] were carried out using a local installation of version 2.1 at the IN2P3 supercomputing center in Lyon. For the multi-sequence alignment, all templates available on the PDB up to the 30th of April 2018 were included. The AF2 confidence score was extracted to evaluate the confidence of the tool on the predicted protein-protein complexes[44]. The accuracy of AF2 predictions was assessed using the same metrics as the predictions supplied by LightDock.

Results

The docking workflow recovers PROTAC-mediated ternary complex-compatible PPIs

To evaluate the accuracy of our protocol, we first carried out redocking experiments on the benchmark dataset of 16 PROTAC-mediated complexes (**Table S1**). Here, we consider that the docking was successful if it generated at least one acceptable or higher quality solution within a given cohort (Top 1, 5, 10, 20, 50 or 100 ranked poses). From the information gathered in the literature, in particular the reports about ternary complex flexibility from Dixon and co-workers[43], we hypothesized that accurate prediction of these interfaces would be a challenge for an unbiased docking algorithm. A way to help drive the docking towards the correct interface is through the introduction of contact residue restraints, which we pursued as detailed in the experimental section. In brief, we selected all the solvent-accessible residues within a given radius around the ligands bound either to the ligase or the target protein. In the case of the redocking experiment, we removed the atoms which composed the linker up to the anchor atom on each ligand. This restraint selection approach ensures that the docking algorithm is directed towards the PPI while maintaining the overall protocol general and applicable to any ligase-target pair of interest. The redocking experiment is not realistic, as the side-chains are perfectly packed in the crystal, but represents the first test to assess protocol accuracy. **Figure 1** shows, for the 6BOY complex (**A**), a visual representation of the residues used in the restraints for the CRBN ligase at each selection threshold (**B**). **Table S3** contains an exhaustive description of the number of restraints per protein and system. As expected, unbiased docking simulations were not able to generate acceptable poses. For instance, we obtained an accuracy of 25% for the cohort of best 20 poses (**Figure 1C,** Top 20 light-blue bar). The reason behind the poor accuracy of the unbiased docking experiment is that the algorithm attempts to maximize the interaction surface as these are the most energetically favorable in the absence of restraints guiding the algorithm (**Figure S2**). On the other hand, the restrained docking calculations achieved satisfactory accuracy for the cohort of Top 20 poses (88%,

75% and 63% for calculations employing 4, 6 and 8 Å contact residue restraints, **Figure 1C**). Furthermore, we observed significant improvements in performance once the VoroMQA energy rescoring was carried out (**Figure 1D**). In particular, the protocol employing 6 Å contact residue restraints achieved 94% accuracy for the Top 20 and 100% accuracy was achieved for the Top 50 using any distance cutoff for the selection of the residues used as restraints.

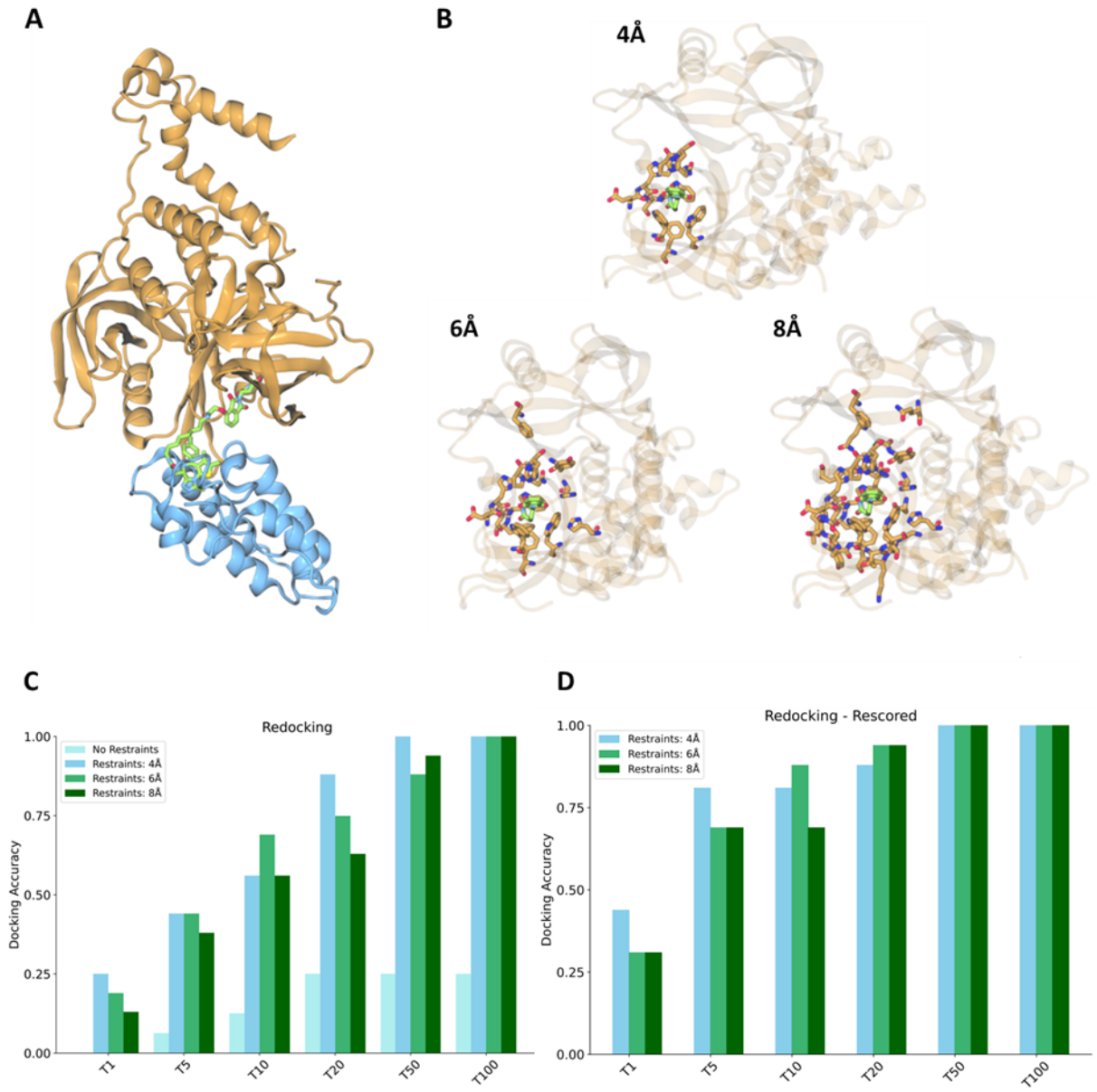


Figure 1 - A) Crystal structure of the PROTAC-mediated complex 6BOY. The ligase component cereblon (CRBN, orange) is bound to the bromodomain target BRD4 (blue), with the complex being stabilized by the PROTAC ligand dBET6 (green). B) The center-of-mass of the thalidomide-like CRBN

ligand from 6BOY (green) is used to select surface-accessible residues to act as docking restraints (orange) at different cutoff radii. C) Redocking calculations performed either with or without contact residue restraints. D) Redocking calculations after rescoring with the VoroMQA energy function.

Minimal solvent-accessible path is essential to recover acceptable Ligase-Target binding poses in realistic docking scenarios

The same protein pairs that were used in the benchmark described above, were adopted for a realistic docking experiment. In this case, the docking uses the crystallographic structures of the unbound ligase and target proteins, each respectively in complex with the recruiter and warhead ligands. Importantly, we did not enforce any prior knowledge on which linker composition and length might fit the ternary complex. As expected, there is a significant drop in accuracy when comparing the redocking with the realistic experiments (88% to 38% for the best performing setups at the Top 20 level, respectively), as illustrated by **Figures 2C** and **2D**. One possible reason may be that the sidechain packing in the redocking experiment is optimal and thus finding appropriate solutions is possible whereas in the realistic case optimal side-chain orientations are not accessible. Nonetheless, we are able to prioritize more acceptable poses after VoroMQA energy rescoring[49], leading to an increase in accuracy, from 38% to 44% at the Top 20 level for the best performing protocol in realistic docking conditions (restraints defined with a 6 Å cutoff). To further improve the accuracy of the realistic docking calculations, we decided to impose a filtering scheme to remove highly ranked but physically unreasonable poses, where the COG euclidean distance between ligands (**Figure 2A**) is too large and/or the ligand binding sites are occluded. To account for binding site occlusion, we filtered the predicted docking poses through the existence and length of the solvent-accessible surface distance between the two ligands (SASD, **Figure 2B**).

The selected SASD filter was an interval from 3 to 14 Å (**Table S1**). As shown in **Figure 3A**, the filters improved remarkably the performance of the workflow, with the accuracy rising

from 44% to 70% (Top 20, 6 Å contact residue restraints). However, the enhancement in performance observed here is still a product of a dual effect: the docking algorithms' ability to produce meaningful solutions and the accuracy of the post-processing procedure, i.e. the energy-rescoring and SASD filtering. To ascertain the performance of the post-processing approach, we removed those systems for which no meaningful docking solutions were found for the best performing setup, which employs 6 Å contact residue restraints (**Figure 3B**). For this protocol, the systems for which the docking produced no solution were two CRBN cases (PDB IDs 5HXB and 6XK9) and two VHL cases (PDB IDs 7JTP, 7KHH). Since the structural differences of the ligase-target systems between their ternary complex and monomeric states are small (see **Table S2**), and their interfaces in the ternary complex are very small (which may be as small as two contact residues), we hypothesized that even small perturbations to sidechain orientation may be sufficient to induce the algorithm into error. This further highlights the difficulty in PROTAC-mediated docking problems, which are exacerbated by the relatively small interface size when compared with standard PPI. We found that after removal of those systems, the protocol exhibits an accuracy of 92% for the Top 20 poses. Interestingly, if we only consider the most well studied ligases (CRBN and VHL), discarding the two systems employing inhibitor of apoptosis E3 ligases (cIAPs), the accuracy of the procedure rises to 90% and 100% at the Top 5 and Top 20 levels respectively. Thus, these results indicate that our post-processing is accurate and the major limitation of the overall protocol appears to be the generation of suitable protein-protein docking poses. In conclusion, we achieved good accuracy if one considers that unbound monomer docking in the case of PROTACs is challenging and no structural refinement was performed. As examples, the highest ranked structures found after PROTACability filtering for three complexes are given in **Figure 3C-E** whereas **Figure S3** illustrates the highest ranking acceptable solutions for all systems here explored.

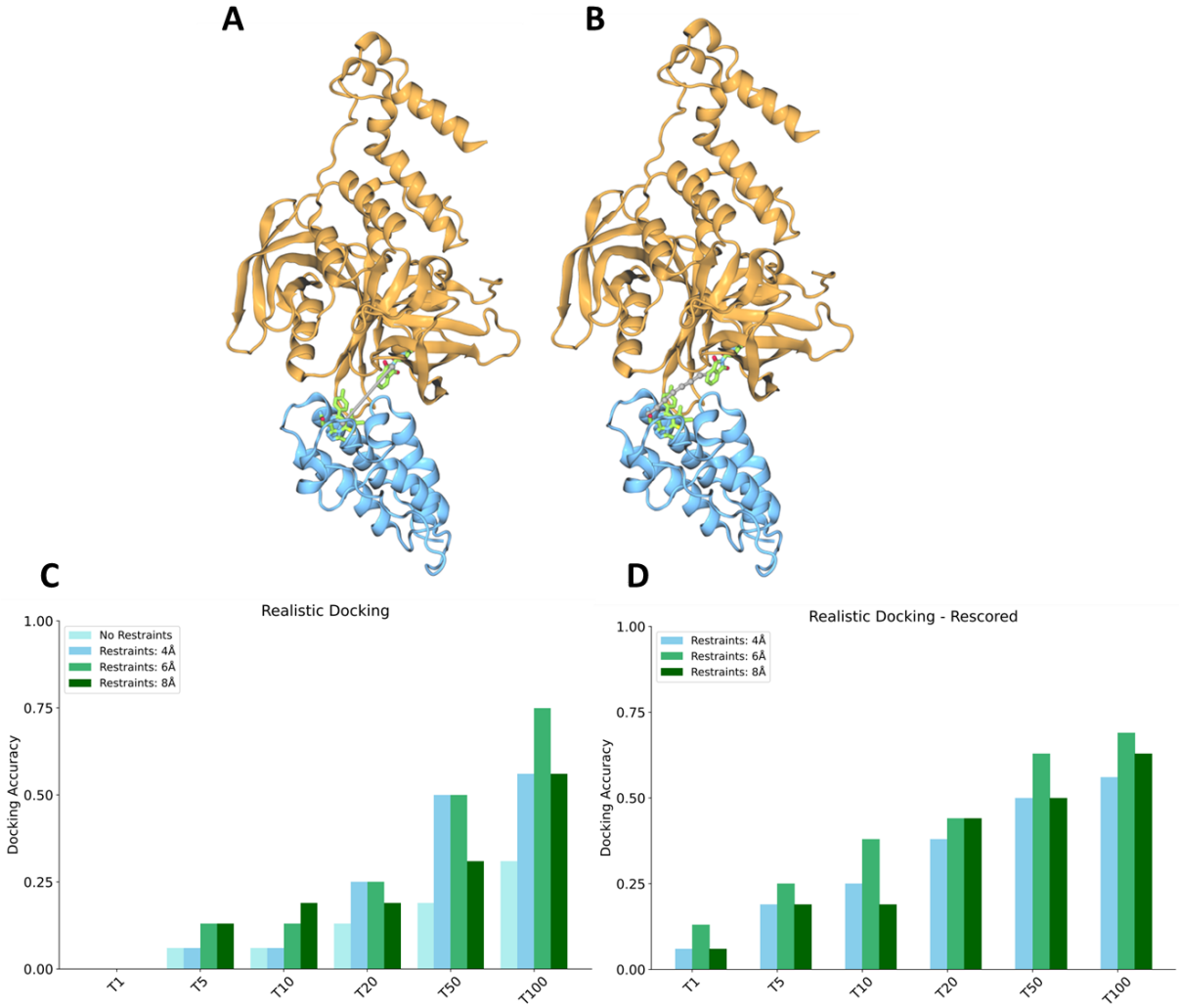


Figure 2 - Realistic docking experiments carried out for the PROTAC benchmark dataset. A) Illustrative figure depicting the COG-COG euclidean distance (grey) between the two ligands composing the dBet6 PROTAC (green), bound to the E3 ligase (orange) and a bromodomain target (blue). B) Illustrative figure highlighting the minimal SASD path (grey) between the two ligands. C) Accuracy of realistic docking in the presence or absence of contact residue restraints. D) Accuracy of realistic docking followed by VoroMQA rescoring.

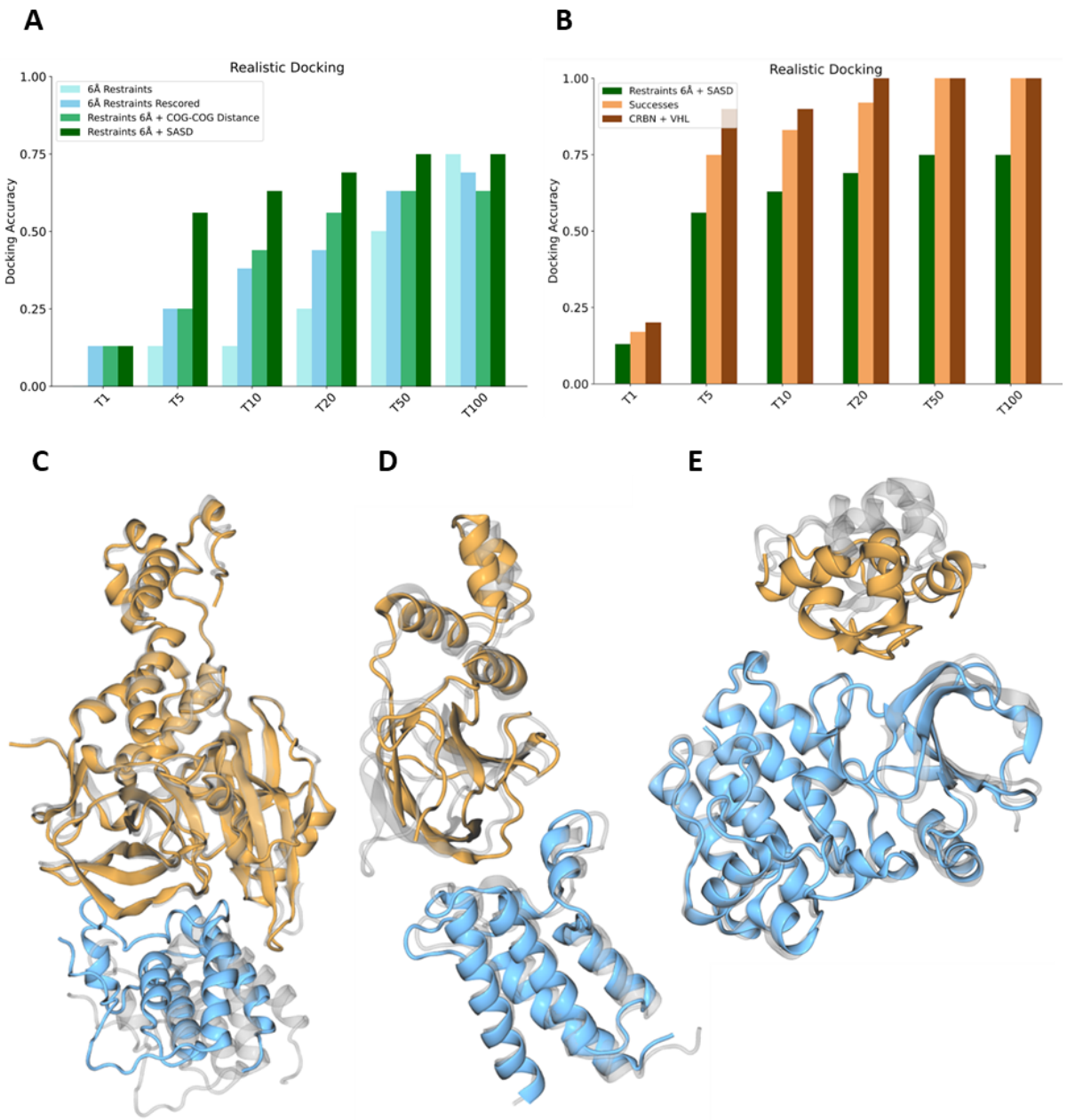


Figure 3 - Accuracy of the realistic docking followed by rescoring and SASD filtering. A) Docking accuracy before and after filtering the predicted poses using either the COG-COG euclidean distance between ligands or the SASD computed using Jwalk. B) Summary of the performance of the best protocol. The first bar corresponds to the best performing setup, employing both VoroMQA energy rescoring and SASD filtering (dark green). The success cases are those (12 out of 16) for which at least acceptable-scored solutions were found by the docking engine in the full top 200 predicted structures (light orange). The last bar (dark brown) highlights the performance of the protocol when considering the successful cases that contain either CRBN or VHL ligases (10 out of 12 in total). C) Showcase comparing the 6BOY (CRBN-BRD4) crystal structure (grey) to the best prediction (ranked

first) obtained using our workflow (colored). D) Showcase comparing the 6HAX (VHL-BRD4) to crystal structure (grey) to the best prediction (ranked third) obtained using our workflow (colored). E) Showcase comparing the 6W8I (cIAP-BTK) crystal structure (grey) to the best prediction (ranked sixth) obtained using our workflow (colored).

Discussion

In this study, we proposed a novel protocol for the generation of PROTAC-compatible PPI structures in the case no known PROTAC molecule exists, and the POI and ligase ligands are known. To this end, we manually curated a dataset of PROTAC-containing ternary complexes from the PDB to serve as a benchmark dataset. To the best of our knowledge, this is the most structurally diverse and curated dataset of PROTAC-mediated ternary complexes available. Using this dataset for protocol benchmarking, we carried out both redocking and realistic docking experiments (see **Results**) where the accuracy of the docking was measured by evaluating the quality of the predicted structures using CAPRI and DockQ criteria[77–79]. In particular, we considered that a docking experiment was successful at a given cohort if at this cohort (for example, the 5 top ranked solutions), at least one structure with acceptable quality was found. For this study, we consider that the docking accuracy is only meaningful until the top 20 cohort.

We found that both the inclusion of contact residue restraints and the rescoring using the energy function of VoroMQA[49], were fundamental steps to increase the accuracy of the calculations in both redocking and realistic docking experiments. In the case of redocking, the best performing protocol (restraints defined with a 6 Å cutoff, rescored) achieved a striking accuracy of 88% and 94% for the 10 or 20 best ranked predicted PPIs, respectively. The unbiased docking experiments yielded poor accuracy even when the solutions were rescored, hinting that the docking solutions were frequently those that maximized the PPI area and not those that were compatible with ternary complex formation. This was expected, as most docking scoring functions are developed to reproduce interfaces with large contact areas and PROTAC-mediated complexes exhibit typically small and transient PPIs, as illustrated in **Figure S1**. In the realistic docking experiments, the accuracy was much lower, even with restraints (below 50% for the Top 20). We noticed that part of the decrease in accuracy was due to the production of high ranking solutions that presented occlusion of the ligand binding sites in either or both of the proteins. This is less frequent in the redocking

experiments as the side-chain packing takes into account the small-molecule ligands and linker atoms. As such, an extra step was taken to allow for better filtering of the predicted solutions. These results suggest that a PPI is PROTAC-compatible when the two ligand binding sites are accessible to the solvent, and it is possible to thread a solvent-accessible path that connects the two ligands. Based on this idea, we decided to use the minimal SASD[50] as a filter and applied it to the VoroMQA-rescored predicted PPIs, named PROTACability.

The application of the PROTACability filter improves the docking accuracy, which rises from 44% to 70% for the Top 20 cohort. Nonetheless, the docking algorithm was unable to produce acceptable solutions in four (4) out of 16 (16) cases. Thus, if we consider only these 12 cases, we state that the post-processing workflow (rescoring + SASD filtering) is highly accurate for both the Top 10 and Top 20 cohorts (83% and 92% accuracy, respectively). Furthermore, if we consider only those systems containing either VHL or CRBN E3 ligases, which are the most well studied E3 ligases in the literature, our method achieves an accuracy of 90% and 100% at the level of the Top 10 or Top 20, respectively. As the PROTACability filter was calibrated on an asymmetrically distributed benchmark dataset, which is dominated by VHL and, to a lesser extent, Cereblon, ligases, we expected it to be better suited to complexes formed by one of these two ligases than to the cIAP based systems, as observable from the Top 5 accuracy data in **Figure 3B**. Nonetheless, we are still able to recover acceptable solutions for the two cIAP systems, at the cost of overall performance at the early stages (Top 5 to Top 20, **Figure 3E**).

Since our protocol only requires information already available in the ligand-bound monomeric forms of the proteins, it is generally applicable to any ligase-target protein pair. Furthermore, it is computationally cheap as only three steps are necessary: (1) restrained docking using LightDock, (2) energy rescoring using VoroMQA and (3) filtering by SASD. As such, we think our protocol may be useful to accelerate PROTAC-based drug discovery campaigns where no known PROTAC is available while being computationally tractable, with

an average run-time per system of 23 minutes using 20 cpu cores (460 minutes per system on a single core (**Table S4**)), and simple to employ.

Nonetheless, there are some inherent limitations to our work. Having taken the decision to keep it computationally cheap and fast, there are no structural refinement steps. This is clearly fundamental when looking at the showcases in **Figures 3C, 3D** and **3E**. Comparing the reference structures (grey) to the predicted poses (colored), it appears that even when guided by restraints the docking algorithm still tries to maximize the interface contacts, leading to a super-packed binding mode. One of the reasons for this behavior is that the training set of the DFIRE[72] scoring function was based on typical protein-protein complexes, which tend to have larger interface areas than those present in PROTAC-mediated ternary complexes (**Figure S2**) and contained no PROTAC-mediated protein-protein complex. Possible avenues for improvement could pass by including protein flexibility by means of Normal Mode Analysis (NMA) in internal coordinates[80], energy minimization steps to relax the produced structures or the inclusion of the ligand molecules in the structures during docking. The first two suggestions would improve side-chain packing at the level of the realistic docking experiments at the cost of docking throughput. However, it would probably increase even further the realistic dockings' accuracy. The last suggestion would be useful at preventing the ligand-binding site from becoming part of the protein-protein interaction surface, which is exactly what happens in the case of the realistic docking calculations for 5HXB and 6XK9 (**Figure S4**).

Further refinement of the docking-produced structures may pass through molecular dynamics simulations. By simulating the predicted PPI structures and enforcing soft restraints on the region where the PROTAC molecule would be expected to be located, we allow the structure the possibility to relax, while maintaining the predicted binding mode, and optimize the side-chain packing, which would in turn lead to a higher quality structure for further studies. Finally, Jwalk[50] does not consider the ligand molecules when computing the minimal SASD, meaning that this path may run across the ligand molecular structures

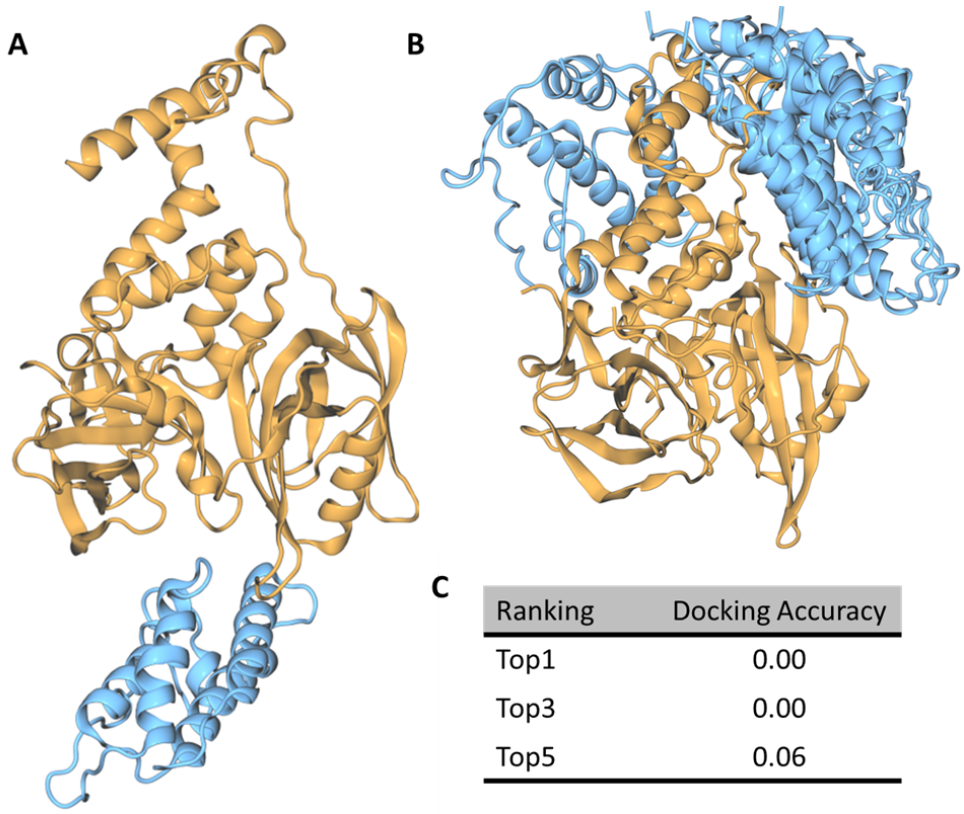
depending on the positions of the anchor atoms. A further potential improvement can come from the application of an *in-house* developed tool which does exclude from the path construction the voxels occupied by ligand atoms (**Figure S5**). Additionally, one key aspect in any docking simulation is the scoring function used. Here we have not carried out a comprehensive benchmark of docking scoring functions and it is possible that DFIRE[72] is not the best amongst the multitude of options available. This is particularly important because DFIRE was trained on a dataset of common protein-protein complexes and includes no PROTAC-mediated complexes[72]. As such, the proposed workflow is accurate but improvable, and future work in our group is aimed at increasing the quality of the docking-predicted PPIs.

Other computational studies based on molecular docking have been pursued in recent years, aiming at the establishment of easy-to-use and accurate PROTAC drug design strategies. For example, the group behind PROsettaC[28] and the group of Tingjun Hou[29] both aimed at reproducing the crystallographic PPI of PROTAC-mediated ternary complexes. On one hand, the PRosettaC procedure was only successful in redocking experiments, as noted previously[29]. On the other hand, the workflow by Weng and co-workers[29] was shown to be able to generate medium to high quality solutions for 12 out of the 14 PROTAC systems queried in a realistic docking scenario. However, both setups employed several docking steps, side-chain optimization and, most importantly, *a priori* knowledge of the PROTAC to use, which was critical for the filtering and retrieval of the most relevant protein-protein docking poses in their work. Furthermore, the emphasis of the studies is on generating binding poses which resemble the crystal structure. We hypothesize that this is not the most optimal approach, since according to Dixon and co-workers[43], the free energy landscape of these complexes is not dominated by a single minima but instead is composed of several minima separated by low-energy barriers. As such, the crystal structure is one possible solution but it is not necessarily the only acceptable solution and trying to reproduce it may

lead us to discard possible conformations which are indeed biologically relevant and may be used in drug design campaigns.

Since DeepMind AlphaFold is now able to predict protein-protein complexes in the multimer variant, we also questioned whether AF2-multimer[44] would be able to correctly predict the PROTAC-mediated PPI interfaces. We retrieved the protein sequences from the complexes in the benchmark set and submitted them to a local installation of AF2-multimer. As shown in **Figure 4C**, the methods' accuracy according to DockQ and/or CAPRI was low. As an example, **Figure 4B** shows the predictions obtained for 6BOY when run through AF2-Multimer, which are in complete disagreement with the experimental structure shown in **Figure 4A**. Nonetheless, it appears that AF2-Multimer knows that the predicted interface is possibly wrong, judging from the low confidence score it assigns (Score < 0.5). A possible avenue to improve on these predictions is the inclusion of ligands in the modeling process, as proposed by Hekkelman and co-workers[81].

As such, it appears that AF2 is either not able to correctly identify ligand-mediated PPIs which are natively low affinity or is predicting the native ligase-target PPI in absence of ligand. However, the latter hypothesis is unlikely since the PPI surface predicted for all CRBN-based cases is where another protein in the ubiquitination machinery binds (DNA damage-binding protein 1, DDB1). Thus, it appears that AF2-multimer was not trained to predict ligand-mediated PPIs, such as those occurring in PROTAC-mediated complexes, highlighting the critical contribution of the restraints. So, for the moment, rational and physical-based approaches combined, as described here, seem to be a better choice towards computational design of PROTACs-compatible PPIs.



C

Ranking	Docking Accuracy
Top1	0.00
Top3	0.00
Top5	0.06

Figure 4 - Calculations using AlphaFold2-Multimer. A) Reference structure from 6BOY, where the PROTAC molecule dBET6 was removed for clarity, illustrating the interaction interface between cereblon (orange) and the BRD4 bromodomain (blue). B) All five models produced by AF2-Multimer, superimposed on cereblon. C) Table showing the docking accuracy calculated for all 16 systems using AF2-Multimer.

Conclusion

In this study we have described a new LightDock-based protocol for PROTAC drug design. Our workflow combines restraint-enabled LightDock simulations with energy-based rescoring from VoroMQA and SASD calculations from Jwalk to yield at least acceptable quality PPI complex structures, starting only from the ligand-bound monomeric form of the E3 ligase and the protein target. The protocol is generalizable to any E3 ligase/target pair, fast and accurate, achieving 94% and 70% accuracy for the top 20 cohort on the redocking and realistic docking experiments, respectively, when benchmarked against a manually curated and diverse dataset of PROTAC-mediated ternary complexes. In the case that no PROTAC

molecule is known to bind the ligase/target pair of interest, our approach is able to provide initial configurations which reproduce the protein-protein binding interface and provide the minimal solvent accessible path connecting the ligand and target ligands, potentially accelerating PROTAC drug design campaigns by narrowing down the chemical space to be searched for linker selection. Further improvements to the protocol are possible, notably through the inclusion of side-chain flexibility via NMA calculations or through molecular dynamics simulations to relax and promote correct packing of residue side-chains within the complex.

Financial support

Brian Jiménez-García (BJG) is employed by Zymvol Biomodeling on a project which received funding from the European Union's Horizon 2020 research and innovation programme under Marie Skłodowska-Curie grant agreement No. 801342 (Tecniospring INDUSTRY) and the Government of Catalonia's Agency for Business Competitiveness (ACCIÓ). PCTS, JM, GPP, GL and RP are supported by the French National Center for Scientific Research (CNRS). Further funding of GPP and PCTS came from a research collaboration with PharmCADD.

Conflicts of Interest

All authors declare no existing conflict of interest.

Acknowledgements

Computer hours awarded to the GENCI project number 2022-A0120713456 on the Jean-Zay and Joliot-Curie Rome clusters of the French National Supercomputing Center and the

computing time available at the IN2P3 computer cluster in Lyon are gratefully acknowledged. The authors also acknowledge the support provided by the CNRS and PharmCADD.

References

1. Ryan, D. P. & Matthews, J. M. Protein–protein interactions in human disease. *Curr. Opin. Struct. Biol.* **15**, 441–446 (2005).
2. Keskin, O., Tuncbag, N. & Gursoy, A. Predicting Protein–Protein Interactions from the Molecular to the Proteome Level. *Chem. Rev.* **116**, 4884–4909 (2016).
3. Kuzmanov, U. & Emili, A. Protein-protein interaction networks: probing disease mechanisms using model systems. *Genome Medicine* **5**, 37 (2013).
4. Schweppe, D. K. *et al.* Mitochondrial protein interactome elucidated by chemical cross-linking mass spectrometry. *Proc. Natl. Acad. Sci. U. S. A.* **114**, 1732–1737 (2017).
5. Cerdan, A. H., Martin, N. É. & Cecchini, M. An Ion-Permeable State of the Glycine Receptor Captured by Molecular Dynamics. *Structure* **26**, 1555–1562.e4 (2018).
6. Kleijnen, M. F. *et al.* The hPLIC proteins may provide a link between the ubiquitination machinery and the proteasome. *Mol. Cell* **6**, 409–419 (2000).
7. Liu, J. & Nussinov, R. The mechanism of ubiquitination in the cullin-RING E3 ligase machinery: conformational control of substrate orientation. *PLoS Comput. Biol.* **5**, e1000527 (2009).
8. Ozbabacan, A., Ece, S., Engin, H. B., Gursoy, A. & Keskin, O. Transient protein–protein interactions. *Protein Eng. Des. Sel.* **24**, 635–648 (2011).
9. Ghadie, M. A. & Xia, Y. Are transient protein-protein interactions more dispensable? *PLoS Comput. Biol.* **18**, e1010013 (2022).
10. Garber, K. The PROTAC gold rush. *Nat. Biotechnol.* **40**, 12–16 (2022).
11. Sakamoto, K. M. *et al.* Protacs: Chimeric molecules that target proteins to the Skp1–Cullin–F box complex for ubiquitination and degradation. *Proc. Nat. Acad. Sci.* **98**, 8554–8559 (2001).
12. Crew, A. P. *et al.* Identification and Characterization of Von Hippel-Lindau-Recruiting Proteolysis Targeting Chimeras (PROTACs) of TANK-Binding Kinase 1. *J. Med. Chem.* **61**, 583–598 (2018).

13. Békés, M., Langley, D. R. & Crews, C. M. PROTAC targeted protein degraders: the past is prologue. *Nat. Rev. Drug Discov.* **21**, 181–200 (2022).
14. Bondeson, D. P. *et al.* Catalytic in vivo protein knockdown by small-molecule PROTACs. *Nat. Chem. Biol.* **11**, 611–617 (2015).
15. Pettersson & Crews. PROteolysis TArgeting Chimeras (PROTACs)—past, present and future. *Drug Discov. Today:Technol.* 15–27 (2019).
16. Gadd, M. S. *et al.* Structural basis of PROTAC cooperative recognition for selective protein degradation. *Nat. Chem. Biol.* 514–521 (2017).
17. Jaime-Figueroa, S., Buhimschi, A. D., Toure, M., Hines, J. & Crews, C. M. Design, synthesis and biological evaluation of Proteolysis Targeting Chimeras (PROTACs) as a BTK degraders with improved pharmacokinetic properties. *Bioorg. Med. Chem. Lett.* **30**, 126877 (2020).
18. Zeng, M. *et al.* Exploring Targeted Degradation Strategy for Oncogenic KRASG12C. *Cell Chemical Biology* **27**, 19–31.e6 (2020).
19. Lai, A. C. *et al.* Modular PROTAC Design for the Degradation of Oncogenic BCR-ABL. *Angew. Chem. Int. Ed Engl.* **55**, 807–810 (2016).
20. He, M. *et al.* PROTACs: great opportunities for academia and industry (an update from 2020 to 2021). *Signal Transduction and Targeted Therapy* **7**, (2022).
21. Cheng, M. *et al.* Discovery of Potent and Selective Epidermal Growth Factor Receptor (EGFR) Bifunctional Small-Molecule Degraders. *J. Med. Chem.* **63**, 1216–1232 (2020).
22. Gao, X. *et al.* Phase 1/2 study of ARV-110, an androgen receptor (AR) PROTAC degrader, in metastatic castration-resistant prostate cancer (mCRPC). *J. Clin. Orthod.* **40**, 17–17 (2022).
23. Hamilton, E. *et al.* First-in-human safety and activity of ARV-471, a novel PROTAC® estrogen receptor degrader, in ER+/HER2-locally advanced or metastatic breast cancer. *Cancer Res.* **82**, PD13–08 (2022).
24. Weng, G. *et al.* PROTAC-DB: an online database of PROTACs. *Nucleic Acids Res.* **49**, D1381–D1387 (2021).

25. Burley, S. K. *et al.* RCSB Protein Data Bank: powerful new tools for exploring 3D structures of biological macromolecules for basic and applied research and education in fundamental biology, biomedicine, biotechnology, bioengineering and energy sciences. *Nucleic Acids Res.* **49**, D437–D451 (2021).
26. Drummond, M. L. & Williams, C. I. In Silico Modeling of PROTAC-Mediated Ternary Complexes: Validation and Application. *J. Chem. Inf. Model.* **59**, 1634–1644 (2019).
27. Drummond, M. L., Henry, A., Li, H. & Williams, C. I. Improved Accuracy for Modeling PROTAC-Mediated Ternary Complex Formation and Targeted Protein Degradation via New In Silico Methodologies. *J. Chem. Inf. Model.* **60**, 5234–5254 (2020).
28. Zaidman, D., Prilusky, J. & London, N. PRosettaC: Rosetta Based Modeling of PROTAC Mediated Ternary Complexes. *J. Chem. Inf. Model.* **60**, 4894–4903 (2020).
29. Weng, G., Li, D., Kang, Y. & Hou, T. Integrative Modeling of PROTAC-Mediated Ternary Complexes. *J. Med. Chem.* **64**, 16271–16281 (2021).
30. Bai, N. *et al.* Modeling the CRL4A ligase complex to predict target protein ubiquitination induced by cereblon-recruiting PROTACs. *J Biol Chem* **298**, 101653 (2022).
31. Śledź, P. & Caflisch, A. Protein structure-based drug design: from docking to molecular dynamics. *Curr. Opin. Struct. Biol.* **48**, 93–102 (2018).
32. Chen, R., Li, L. & Weng, Z. ZDOCK: an initial-stage protein-docking algorithm. *Proteins* **52**, 80–87 (2003).
33. Dominguez, C., Boelens, R. & Bonvin, A. M. J. J. HADDOCK: a protein-protein docking approach based on biochemical or biophysical information. *J. Am. Chem. Soc.* **125**, 1731–1737 (2003).
34. Cheng, T. M. K., Blundell, T. L. & Fernandez-Recio, J. pyDock: Electrostatics and desolvation for effective scoring of rigid-body protein–protein docking. *Proteins* **68**, 503 (2007).
35. Karaca, E. & Bonvin, A. M. J. J. A multidomain flexible docking approach to deal with large conformational changes in the modeling of biomolecular complexes. *Structure* **19**, 555–565 (2011).

36. Jiménez-García, B. *et al.* LightDock: a new multi-scale approach to protein-protein docking. *Bioinformatics* **34**, 49–55 (2018).
37. Soler, D., Westermaier, Y. & Soliva, R. Extensive benchmark of rDock as a peptide-protein docking tool. *J. Comput. Aided Mol. Des.* **33**, 613–626 (2019).
38. Roel-Touris, J., Bonvin, A. M. J. J. & Jiménez-García, B. LightDock goes information-driven. *Bioinformatics* **36**, 950–952 (2020).
39. Ramírez-Aportela, E., López-Blanco, J. R. & Chacón, P. FRODOCK 2.0: fast protein-protein docking server. *Bioinformatics* **32**, 2386–2388 (2016).
40. Schueler-Furman, O., Wang, C. & Baker, D. Progress in protein-protein docking: atomic resolution predictions in the CAPRI experiment using RosettaDock with an improved treatment of side-chain flexibility. *Proteins* **60**, 187–194 (2005).
41. Kilambi, K. P., Reddy, K. & Gray, J. J. Protein-protein docking with dynamic residue protonation states. *PLoS Comput. Biol.* **10**, e1004018 (2014).
42. Marze, N. A. *et al.* Modeling oblong proteins and water-mediated interfaces with RosettaDock in CAPRI rounds 28-35. *Proteins* **85**, 479–486 (2017).
43. Dixon, T. *et al.* Predicting the structural basis of targeted protein degradation by integrating molecular dynamics simulations with structural mass spectrometry. *Nat Comm* **13***,* 5884 (2022).
44. Evans, R. *et al.* Protein complex prediction with AlphaFold-Multimer. *bioRxiv* 2021.10.04.463034 (2022) doi:10.1101/2021.10.04.463034.
45. Du, Z. *et al.* The trRosetta server for fast and accurate protein structure prediction. *Nat. Protoc.* **16**, 5634–5651 (2021).
46. Zheng, W. *et al.* Folding non-homologous proteins by coupling deep-learning contact maps with I-TASSER assembly simulations. *Cell Rep Methods* **1**, (2021).
47. Thornton, J. M., Laskowski, R. A. & Borkakoti, N. AlphaFold heralds a data-driven revolution in biology and medicine. *Nat. Med.* **27**, 1666–1669 (2021).
48. Roel-Touris, J., Jiménez-García, B. & Bonvin, A. M. J. J. Integrative modeling of membrane-associated protein assemblies. *Nat. Commun.* **11**, 6210 (2020).

49. Olechnovič, K. & Venclovas, Č. VoroMQA: Assessment of protein structure quality using interatomic contact areas. *Proteins* **85**, 1131–1145 (2017).

50. Bullock, J. M. A., Sen, N., Thalassinos, K. & Topf, M. Modeling Protein Complexes Using Restraints from Crosslinking Mass Spectrometry. *Structure* **26**, 1015–1024.e2 (2018).

51. Martin, J. When Alphafold2 predictions go wrong for protein–protein complexes, is there something to be learnt? *Q. Rev. Biophys.* **55**, e6 (2022).

52. Farnaby, W. *et al.* BAF complex vulnerabilities in cancer demonstrated via structure-based PROTAC design. *Nat. Chem. Biol.* **15**, 672–680 (2019).

53. Chung, C.-W. *et al.* Structural Insights into PROTAC-Mediated Degradation of Bcl-xL. *ACS Chem. Biol.* **15**, 2316–2323 (2020).

54. Yu, X. *et al.* A selective WDR5 degrader inhibits acute myeloid leukemia in patient-derived mouse models. *Sci. Transl. Med.* **13**, eabj1578 (2021).

55. Dragovich, P. S. *et al.* Antibody-Mediated Delivery of Chimeric BRD4 Degraders. Part 2: Improvement of In Vitro Antiproliferation Activity and In Vivo Antitumor Efficacy. *J. Med. Chem.* **64**, 2576–2607 (2021).

56. Law, R. P. *et al.* Discovery and characterisation of highly cooperative FAK-degrading PROTACs. *Angew. Chem. Int. Ed Engl.* **60**, 23327–23334 (2021).

57. Kraemer, A. *et al.* Quaternary Complex of human WDR5 and pVHL:ElonginC:ElonginB bound to PROTAC Homer. Structure at https://doi.org/10.2210/pdb7q2j/pdb (2021).

58. Matyskiela, M. E. *et al.* A novel cereblon modulator recruits GSPT1 to the CRL4CRBN ubiquitin ligase. *Nature* **535**, 252–257 (2016).

59. Nowak, R. P. *et al.* Plasticity in binding confers selectivity in ligand-induced protein degradation. *Nat. Chem. Biol.* **14**, 706–714 (2018).

60. Surka, C. *et al.* CC-90009, a novel cereblon E3 ligase modulator, targets acute myeloid leukemia blasts and leukemia stem cells. *Blood* **137**, 661–677 (2021).

61. Wang, E. S. *et al.* Acute pharmacological degradation of Helios destabilizes regulatory T cells. *Nat. Chem. Biol.* **17**, 711–717 (2021).

62. Schiemer, J. *et al.* Snapshots and ensembles of BTK and cIAP1 protein degrader ternary complexes. *Nat. Chem. Biol.* **17**, 152–160 (2021).
63. Kamura, T. *et al.* Activation of HIF1α ubiquitination by a reconstituted von Hippel-Lindau (VHL) tumor suppressor complex. *Proc. Nat. Acad. Sci.* **97**, 10430–10435 (2000).
64. Lee, C. C. Crystal Structure of Focal Adhesion Kinase Domain Complexed with 7H-Pyrrolo [2,3-d] pyrimidine Derivative. Structure at https://doi.org/10.2210/pdb2etm/pdb (2006).
65. Schuetz, A. *et al.* Structural basis for molecular recognition and presentation of histone H3 by WDR5. *EMBO J.* **25**, 4245–4252 (2006).
66. Ku, B. *et al.* Structural and Biochemical Bases for the Inhibition of Autophagy and Apoptosis by Viral BCL-2 of Murine γ-Herpesvirus 68. *PLoS Pathog.* **4**, e25 (2008).
67. Matyskiela, M. E. *et al.* A Cereblon Modulator (CC-220) with Improved Degradation of Ikaros and Aiolos. *J. Med. Chem.* **61**, 535–542 (2018).
68. Wanior, M. *et al.* Pan-SMARCA/PB1 Bromodomain Inhibitors and Their Role in Regulating Adipogenesis. *J. Med. Chem.* **63**, 14680–14699 (2020).
69. Li, J. *et al.* Discovery of the natural product 3’,4',7,8-tetrahydroxyflavone as a novel and potent selective BRD4 bromodomain 2 inhibitor. *J. Enzyme Inhib. Med. Chem.* **36**, 903–913 (2021).
70. Humphrey, W., Dalke, A. & Schulten, K. VMD: visual molecular dynamics. *J. Mol. Graph.* **14**, 33–8, 27–8 (1996).
71. Krishnanand, K. N. & Ghose, D. A Glowworm Swarm Optimization Based Multi-robot System for Signal Source Localization. in *Design and Control of Intelligent Robotic Systems* (eds. Liu, D., Wang, L. & Tan, K. C.) 49–68 (Springer Berlin Heidelberg, 2009).
72. Zhang, C., Liu, S. & Zhou, Y. Docking prediction using biological information, ZDOCK sampling technique, and clustering guided by the DFIRE statistical energy function. *Proteins* **60**, 314–318 (2005).
73. Faraggi, E., Yang, Y., Zhang, S. & Zhou, Y. Predicting continuous local structure and the effect of its substitution for secondary structure in fragment-free protein structure

prediction. *Structure* **17**, 1515–1527 (2009).

74. Michaud-Agrawal, N., Denning, E. J., Woolf, T. B. & Beckstein, O. MDAnalysis: a toolkit for the analysis of molecular dynamics simulations. *J. Comput. Chem.* **32**, 2319–2327 (2011).
75. Gowers, R. J. *et al.* MDAnalysis: a Python package for the rapid analysis of molecular dynamics simulations. in *Proceedings of the 15th Python in Science Conference* (eds. Benthall, S. & Rostrup, S.) 98–105 (2016).
76. Hubbard, S. J. & Thornton, J. M. *Naccess*. (1993).
77. Méndez, R., Leplae, R., Lensink, M. F. & Wodak, S. J. Assessment of CAPRI predictions in rounds 3-5 shows progress in docking procedures. *Proteins* **60**, 150–169 (2005).
78. Lensink, M. F. & Wodak, S. J. Docking, scoring, and affinity prediction in CAPRI. *Proteins* **81**, 2082–2095 (2013).
79. Basu, S. & Wallner, B. DockQ: A Quality Measure for Protein-Protein Docking Models. *PLoS One* **11**, e0161879 (2016).
80. Cirauqui Diaz, N., Frezza, E. & Martin, J. Using normal mode analysis on protein structural models. How far can we go on our predictions? *Proteins* **89**, 531–543 (2021).
81. Hekkelman, M. L., de Vries, I., Joosten, R. P. & Perrakis, A. AlphaFill: enriching AlphaFold models with ligands and cofactors. *Nat. Methods* **20**, 205–213 (2023).