Genomics Proteomics Bioinformatics 20 (2022) 882–898

Genomics Proteomics Bioinformatics

www.elsevier.com/locate/gpb
www.sciencedirect.com

ORIGINAL RESEARCH

Machine Learning Modeling of Protein-intrinsic Features Predicts Tractability of Targeted Protein Degradation

Wubing Zhang [1,2,#], Shourya S. Roy Burman [3,4,#], Jiaye Chen [5], Katherine A. Donovan [3,4], Yang Cao [6], Chelsea Shu [3,7], Boning Zhang [1,2], Zexian Zeng [1,2], Shengqing Gu [1,2], Yi Zhang [1,2], Dian Li [1,2], Eric S. Fischer [3,4,*], Collin Tokheim [1,2,*], X. Shirley Liu [1,2,*]

[1] *Department of Data Science, Dana-Farber Cancer Institute, Boston, MA 02215, USA*
[2] *Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, MA 02115, USA*
[3] *Department of Cancer Biology, Dana-Farber Cancer Institute, Boston, MA 02215, USA*
[4] *Department of Biological Chemistry and Molecular Pharmacology, Harvard Medical School, Boston, MA 02115, USA*
[5] *Department of Biomedical Informatics, Harvard Medical School, Boston, MA 02115, USA*
[6] *Center of Growth, Metabolism, and Aging, Key Laboratory of Bio-resource and Eco-environment, Ministry of Education, College of Life Sciences, Sichuan University, Chengdu 610064, China*
[7] *Research Scholar Initiative, Graduate School of Arts and Sciences, Harvard University, Cambridge, MA 02138, USA*

Received 15 July 2022; revised 25 October 2022; accepted 4 November 2022
Available online 6 December 2022

Handled by Feng Gao

KEYWORDS

Targeted protein degradation;
Degradability;
Protein-intrinsic feature;
Ubiquitination;
Machine learning

Abstract **Targeted protein degradation** (TPD) has rapidly emerged as a therapeutic modality to eliminate previously undruggable proteins by repurposing the cell's endogenous protein degradation machinery. However, the susceptibility of proteins for targeting by TPD approaches, termed "**degradability**", is largely unknown. Here, we developed a **machine learning** model, model-free analysis of protein degradability (MAPD), to predict degradability from features intrinsic to protein targets. MAPD shows accurate performance in predicting kinases that are degradable by TPD compounds [with an area under the precision–recall curve (AUPRC) of 0.759 and an area under the receiver operating characteristic curve (AUROC) of 0.775] and is likely generalizable to independent non-kinase proteins. We found five features with statistical significance to achieve optimal prediction, with **ubiquitination** potential being the most predictive. By structural modeling, we found

* Corresponding authors.
E-mail: xsliu.res@gmail.com (Liu XS), collintokheim@gmail.com (Tokheim C), Eric_Fischer@dfci.harvard.edu (Fischer ES).
# Equal contribution.

Peer review under responsibility of Beijing Institute of Genomics, Chinese Academy of Sciences / China National Center for Bioinformation and Genetics Society of China.

https://doi.org/10.1016/j.gpb.2022.11.008

that E2-accessible ubiquitination sites, but not lysine residues in general, are particularly associated with kinase degradability. Finally, we extended MAPD predictions to the entire proteome to find 964 disease-causing proteins (including proteins encoded by 278 cancer genes) that may be tractable to TPD drug development.

Introduction

The dominant pathway for selective protein degradation in eukaryotic cells is the ubiquitin–proteasome system (UPS), which degrades proteins that are covalently modified with ubiquitin [1]. Ubiquitination is carried out in a three-step enzymatic cascade. First, ubiquitin is activated by covalent attachment to the active site of an E1 ubiquitin-activating enzyme. Second, the activated ubiquitin is transferred from the E1 enzyme to an E2 ubiquitin-conjugating enzyme. Finally, E3 ubiquitin ligases facilitate the conjugation of ubiquitin to a substrate. For the majority of ubiquitin ligases, this process does not involve a covalent intermediate comprising a ubiquitin–E3 complex, but rather the proximity induced by an E3 ubiquitin ligase selectively binding to a substrate allows for the transfer of ubiquitin from the E2 enzyme to a lysine residue on the substrate. After repeated rounds of this process, a poly-ubiquitin chain can be formed, which often directs the substrate for degradation by the 26S proteasome [2].

Targeted protein degradation (TPD) is a novel pharmacologic modality that induces the degradation of a protein of interest (POI) by chemically repurposing the UPS [3]. The TPD molecules (degraders), epitomized by the molecular glues [4] and proteolysis targeting chimeras (PROTACs) [5], typically induce the *de novo* ternary complex formation between an E3 ligase and a POI, leading to the ubiquitin transfer to available lysines and subsequent degradation of the POI [6]. Unlike traditional inhibitors that target the catalytic binding site on a POI, degraders can induce protein degradation by binding to non-catalytic sites [7]. Therefore, previously undruggable proteins, such as transcription factors (TF), can be targeted by degraders [8]. For example, the United States Food and Drug Administration (FDA)-approved immunomodulatory drugs (IMiDs) thalidomide, pomalidomide, and lenalidomide induce degradation of TFs IKZF1 and IKZF3 by recruiting them to CRBN [9,10], the substrate recognition subunit of the E3 ubiquitin ligase complex CUL4-RBX1-DDB-CRBN. Over the last two decades, the TPD field has grown dramatically, with thousands of publicly available degraders developed for over 100 human protein targets [PROTAC-DB [11] and PROTACpedia (https://protacdb.weizmann.ac.il/ptcb/main)]. Notably, degraders targeting the androgen receptors (ARs) [12], estrogen receptor (ERs) [13], BCL-XL [14], Ikaros/Aiolos (IKZF1/3) [15], Helios (IKZF2) [16], GSPT1 [17], BTK [18], and IRAK4 have entered into clinical trials [19], and degraders targeting STAT3, BRD9, or TRK will also be tested in patients soon [19]. Despite these advances, it remains challenging to predict which proteins are susceptible and which may be resistant to the TPD approaches.

Chemoproteomic profiling approaches have emerged as a systematic approach to survey protein degradability. Rather than profiling the expression of a single protein in response to a selective degrader, these approaches use mass spectrometry to assess the proteome-wide response to treatment with pan-targeting degraders [20–23]. For example, our recent study profiled 91 multi-kinase degraders to assess the degradability of more than 400 protein kinases, identifying more than 200 kinases as degradable [20]. Using a library of degraders targeting histone deacetylases (HDACs), Xiong et al. investigated the degradability of zinc-dependent HDACs [23]. Together, these broad-targeted degrader profiling experiments have greatly expanded the known degradable proteome. Unfortunately, chemoproteomic approaches to map degradability are inapplicable for most proteins due to the absence of ligands required for target recruitment to the ligase machinery. Thus, computational prediction of protein degradability offers a potentially practical alternative.

It is widely believed that defined ternary complexes are associated with effective and selective target degradation [24,25]. A series of computational methods have been introduced to model PROTAC-mediated ternary complex formation [26–28], which have facilitated the rational and efficient optimization of PROTACs [25,29]. However, several studies have reported that although some level of binary target engagement and ternary complex formation is necessary for target recruitment and ubiquitin transfer, they are not always sufficient for TPD [20,22]. We propose that rather than drug–target interactions alone driving degradability, features intrinsic to the protein targets also heavily influence the degradability of specific targets. For instance, although ubiquitination is the initiation signal for proteasomal degradation [30], the association between protein degradability and known or potential ubiquitination sites (Ub sites) in the target protein is poorly understood.

In this study, we developed a machine learning model, model-free analysis of protein degradability (MAPD), to predict degradability from protein-intrinsic features (**Figure 1**). MAPD shows promising performance in predicting degradable kinases by multi-kinase degraders and previously reported targets of PROTAC compounds. We found that a protein's endogenous ubiquitination potential contributes the most to the degradability predictions. Structural analysis via protein–protein docking revealed the particular importance of E2-accessible Ub sites in determining degradability. Using MAPD, we have expanded our predictions to the human proteome to map protein tractability to TPD approaches. Our results are available at https://mapd.cistrome.org/, which could be a valuable resource for guiding target prioritization toward tractable TPD targets.

Results

Kinase degradability is associated with features intrinsic to the target

Substantial efforts have been invested in the optimization of degraders for any particular target with no guarantee that a

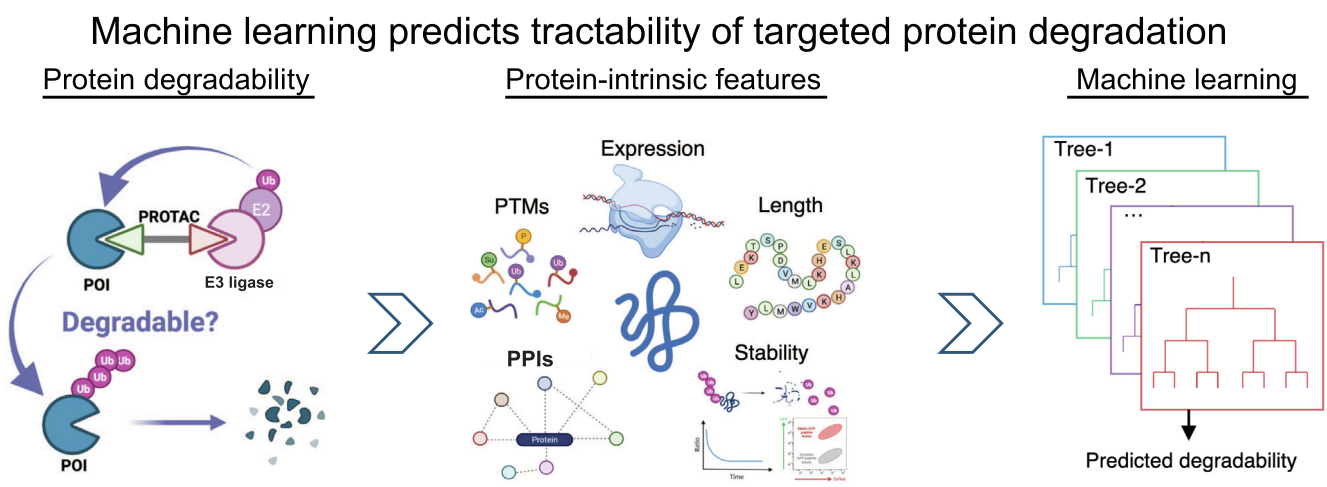


Figure 1 Study overview

The ubiquitin–proteasome system can be repurposed by a PROTAC or other small molecules to degrade a POI. However, it remains to be answered which proteins are amenable to this approach (left). Here, we associated kinase degradability with protein-intrinsic features spanning protein expression, PTMs, protein length, PPIs, protein stability, and protein half-life to identify predictive factors (middle). Based on the predictive features, we developed a machine learning model to predict protein degradability (right). PROTAC, proteolysis targeting chimera; POI, protein of interest; PTM, post-translational modification; PPI, protein–protein interaction; Ub, ubiquitination; Ac, acetylation; P, phosphorylation; Su, sumoylation; Me, methylation.

successful compound will be found [31,32]. Numerous reports indicate that stable ternary complex formation is one of the main factors that influence the effectiveness and selectivity of induced protein degradation [22,24,25,33]. However, our previous chemoproteomic study of the protein kinome indicates that drug–target engagement is insufficient by itself to predict which kinases can be degraded [20], suggesting that unexplained factors influence protein degradability. In this study, we explored factors intrinsic to POIs that may influence their degradability by comparing kinases that all have drug–target engagement, but differ in multi-kinase degrader-induced degradation. We first selected highly and lowly degradable kinases based on the number of multi-kinase degraders found to degrade each POI (**Figure 2**A), with an additional requirement of a high frequency of detection in the underlying global proteomic experiments (Figure S1A; Table S1). We next collected protein features that may be predictive of kinase degradability, including post-translational modifications (PTMs), protein stability, and protein–protein interactions (PPIs) (Table S2). Since the cellular context might influence selective protein degradation and the underlying degradability experiments were conducted mostly in the MOLT4 cell line, we collected relative protein abundance, mRNA expression, and H3K27ac regulatory potential from the MOLT4 cell line [20]. Due to the lack of protein stability in the MOLT4 cell line, we collected protein half-life and global protein stability (GPS) data from immune cells and other cell types [34–39], which resemble each other (Figure 2B). Among all these features, features within a category are highly correlated with each other, while features between categories tend to provide independent information (Figure 2B).

To identify features associated with protein degradability, we compared highly and lowly degradable kinases using a Wilcoxon rank-sum test. Compared to lowly degradable kinases, the highly degradable kinases have a significantly higher proportion of lysine residues that have reported ubiquitination events in PhosphoSitePlus [40] and PLMD database [41] (hereafter referred to as ubiquitination potential) (Figure 2C, Figure S1B and C). The ubiquitination potential likely reflects a protein's endogenous capacity to be ubiquitinated since the ubiquitination events are from cell lines in the absence of degrader treatment [42]. Notably, the percentage of lysine residues on POIs does not vary significantly (Figure S1D). We also observed an enrichment of proteins with a lower half-life in the highly degradable group (Figure 2C, Figure S1E). Given that protein half-life was not correlated with ubiquitination potential (Figure S1F), this indicates an independent signal for predicting protein degradability, although the underlying mechanism by which protein half-life influences protein degradability requires further study. Furthermore, protein degradability is positively associated with mRNA expression of a POI in the assayed cell lines (Figure 2C, Figure S1G), which may imply that the absolute POI expression is associated with the detection of protein degradation. Profiling the protein degradability in more cell contexts might be advantageous for further understanding of the observed association. Despite there being an incomplete understanding of the exact mechanisms that give rise to the observed associations, these results support the valid use of some of these individual features in selecting degradable targets for TPD drug development.

Development of MAPD

We next sought to build a machine learning model, named MAPD, to combine multiple features associated with protein degradability into a single score. Toward this end, we tested six commonly used machine learning methods, including naive Bayes (NB), k-nearest neighbor (KNN), logistic regression (LR), linear-kernel support vector machine (svmLinear),

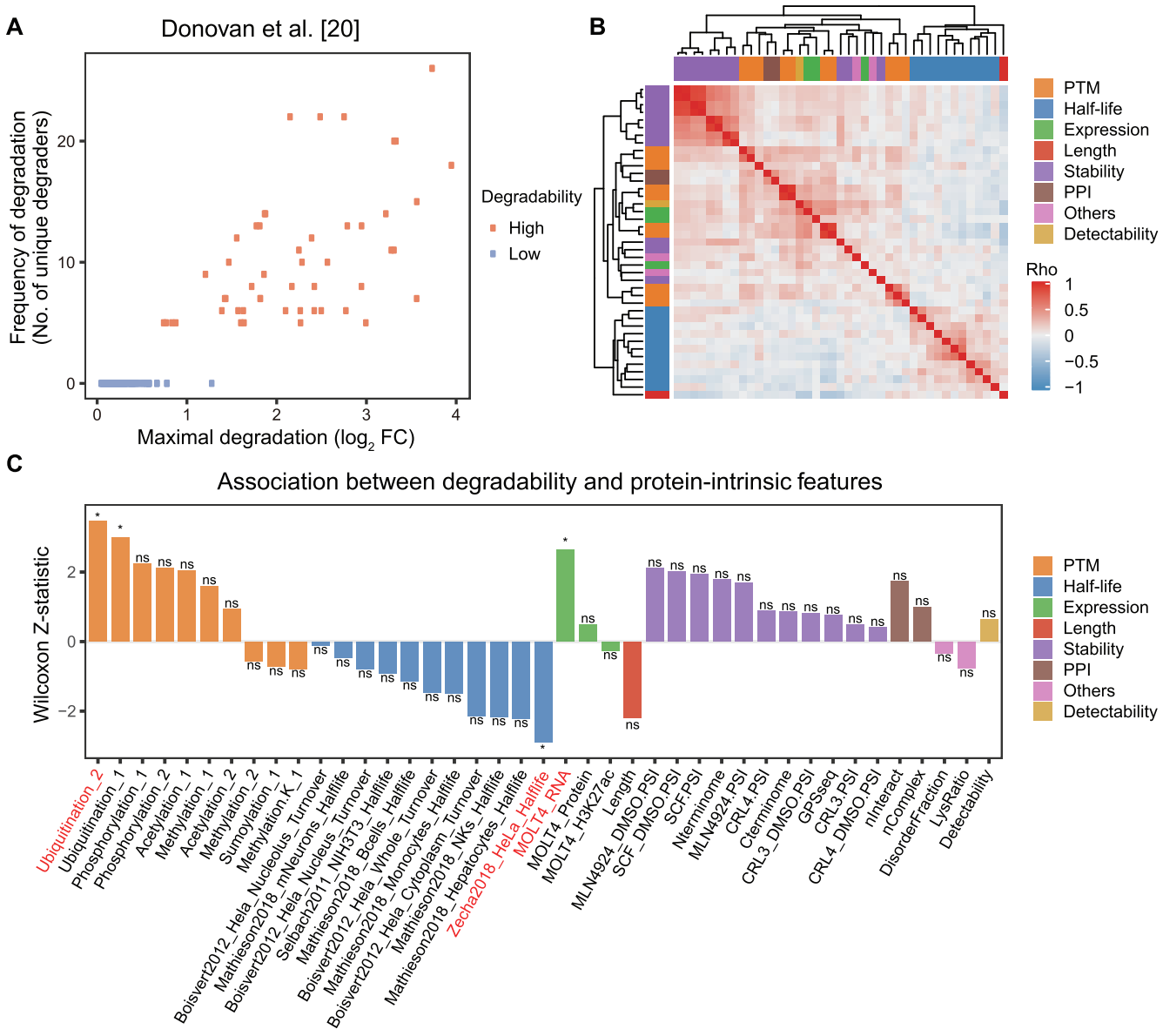


Figure 2 Kinase degradability is associated with features intrinsic to the target

A. Dot plot showing the frequency of degradation and maximal degradation of protein kinases induced by multi-kinase degraders from the study by Donovan and colleagues [20]. Orange dots represent the kinases with high degradability, and light blue dots represent the kinases with low degradability. **B.** Pairwise Spearman's correlation of 42 protein-intrinsic features spanning protein stability, PTM, PPI, protein length, protein half-life, protein expression, protein detectability, and others. **C.** Bar diagram showing the association between the degradability of kinases and their features. The x-axis shows the abbreviated name of protein-intrinsic features (see Table S2 for full details), in which the Ubiquitination_1 and Ubiquitination_2 indicate the proportion of lysine residues with reported ubiquitination events in at least one (Ubiquitination_1) or two references (Ubiquitination_2) from the PhosphoSitePlus database [40], the Zecha2018_HeLa_Halflife indicates the protein half-lives profiled in the HeLa cell line from the study by Zecha and colleagues [36], and the MOLT4_RNA indicates the mRNA expression in the MOLT4 cell line. The y-axis shows the Wilcoxon Z-statistics indicating the association between protein degradability and each protein-intrinsic feature. *, false discovery rate < 0.05; ns, not significant. FC, fold change.

radial-kernel support vector machine (svmRadial), and random forest (RF). Because of the redundancy of protein-intrinsic features, we performed forward feature selection for each method, which iteratively selects the best performing features (Table S3) until the model performance plateaus. We evaluated the performance of these methods using cross-validation, and found that the RF model outperformed other models with an area under the precision–recall curve (AUPRC) of 0.759 (**Figure 3**A) and an area under the receiver operating characteristic curves (AUROC) of 0.775 (Figure S2A). Therefore, all further analyses are based on the RF model implementation.

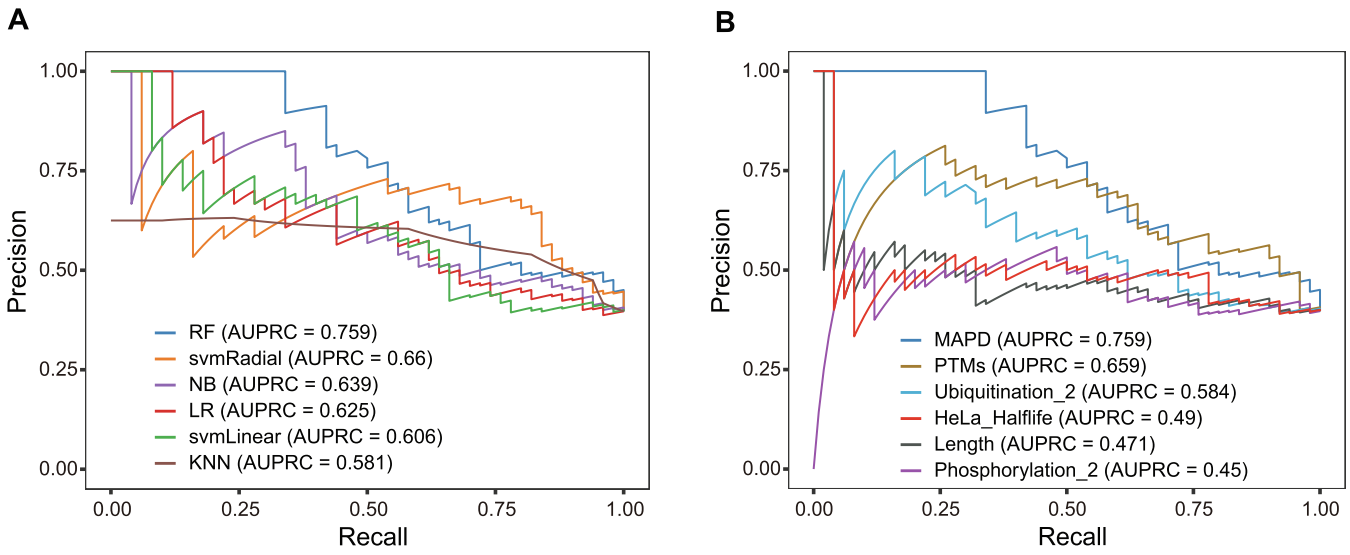


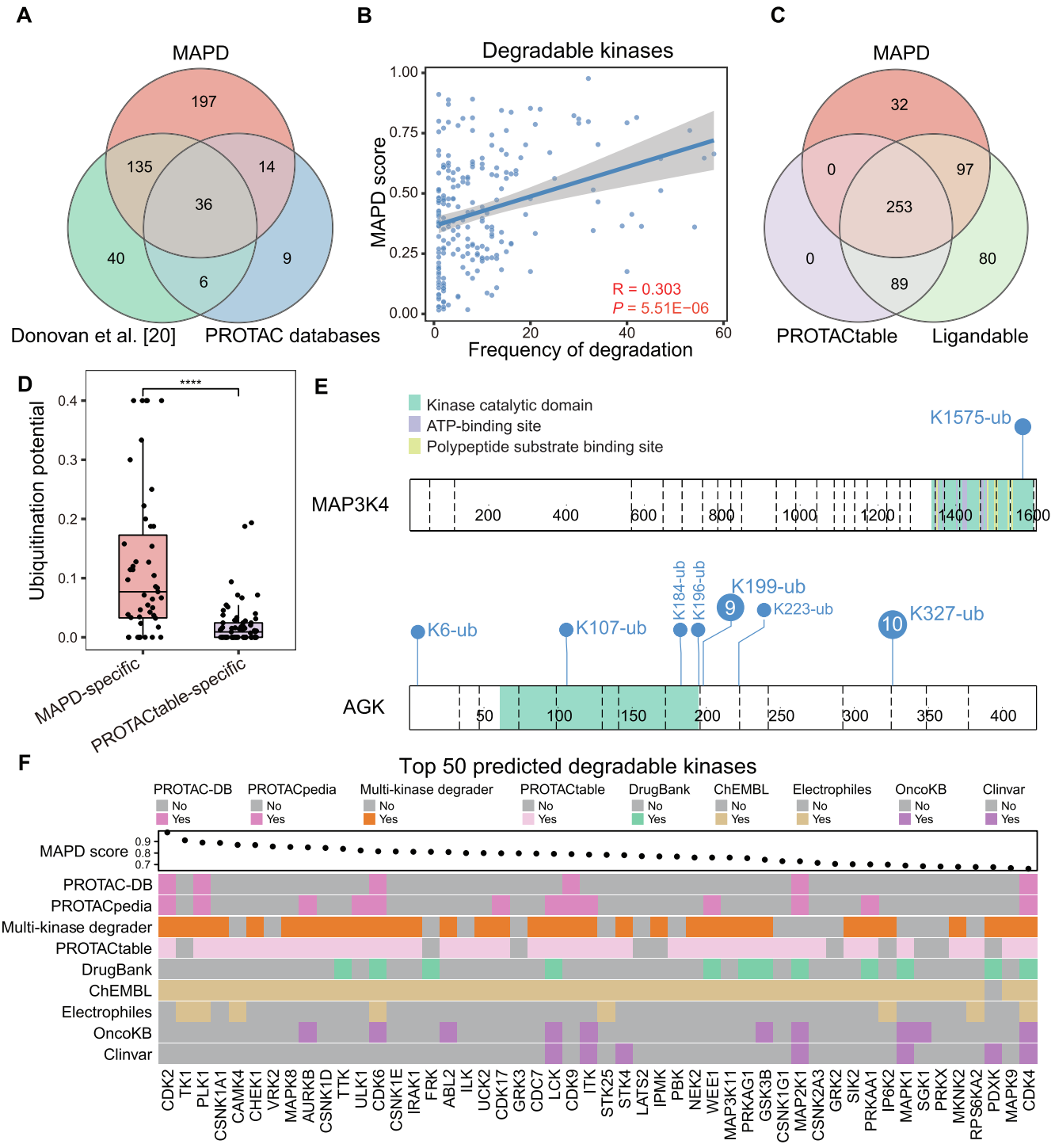
Figure 3 Development of MAPD

A. Precision–recall curves showing the performance of six machine learning models based on 20-fold cross-validation. **B.** Precision–recall curves showing the performance of MAPD and models trained on individual features or a combination of features. 'PTMs' indicates the model trained on the combination of ubiquitination potential (Ubiquitination_2), acetylation potential (Acetylation_1), and phosphorylation potential (Phosphorylation_2). 'Ubiquitination_2' indicates the model trained on ubiquitination potential. 'HeLa_Halflife' indicates the model trained on a single feature describing half-life in HeLa cells. 'Length' indicates the model trained on protein length. 'Phosphorylation_2' indicates the model trained on phosphorylation potential. MAPD, model-free analysis of protein degradability; RF, random forest; svmRadial, radial-kernel support vector machine; NB, naive Bayes; LR, logistic regression; svmLinear, linear-kernel support vector machine; KNN, k-nearest neighbor; AUPRC, area under the precision–recall curve.

Five protein-intrinsic features were identified as important in the MAPD model, including ubiquitination potential, phosphorylation potential, protein half-life, acetylation potential, and protein length (Figure S2B), in order of importance. Next, we compared the performance of MAPD to models that were trained on each individual feature using cross-validation. Consistent with the highest importance of ubiquitination potential in MAPD, the model trained on the ubiquitination potential showed the highest AUPRC (0.584) and AUROC (0.62) among all other single-featured models (Figure 3B, Figure S2C). Interestingly, the combination of the three PTM features (ubiquitination, acetylation, and phosphorylation potential) seems to achieve higher AUPRC (0.659) and AUROC (0.753) than ubiquitination potential alone ($P = 0.058$, Delong's test) (Figure 3B, Figure S2C). Like ubiquitination, acetylation occurs mainly on lysine. More than 39% of PhosphoSitePlus acetylation sites also have reported ubiquitination events (Figure S2D), suggesting that acetylation sites also serve as ubiquitination sites. Hence, the acetylation potential might compensate for the lack of Ub site data and improves the model. Unlike ubiquitination and acetylation, phosphorylation occurs mainly on serine, threonine, and tyrosine residues. It has been demonstrated that phosphorylation can impact protein degradation, and phosphorylation of evolutionarily conserved proteins often accelerates proteasomal degradation [43]. The observed association between phosphorylation potential and POI degradability suggests that phosphorylation might also regulate compound-induced protein degradation, of which the mechanism requires further investigation.

MAPD shows good performance in predicting kinase degradability

To evaluate the robustness of MAPD, we assessed the degradability of the kinome, with the predictions for training kinases collected from 20-fold cross-validation to avoid inflating the performance assessment. We first examined the degradability of kinases profiled in the study by Donovan et al. [20] and found significantly higher MAPD scores of degradable kinases than other kinases engaged by multi-kinase degraders (Figure S3A). This trend is also consistent for specific degraders, such as TL12-186 and SK-3-91 (Figure S3A), although with less significance due to the smaller number of POIs in these datasets. Using a threshold with the best cross-validation accuracy, MAPD identified 382 highly degradable kinase/kinase-related proteins, covering 78.8% (171/217) of experimentally degradable kinases (**Figure 4**A). Consistent with the low MAPD scores, the remaining 21.2% of kinases have a low frequency of degradation (Figure S3B). Furthermore, within all experimentally degraded kinases, MAPD scores show considerable Spearman's correlation with their frequencies of degradation by multi-kinase degraders ($P = 5.51E{-}06$) (Figure 4B), suggesting the ability of MAPD in prioritizing highly degradable targets. We next collected an independent set of PROTAC targets reported in databases [PROTAC-DB [11] and PROTACpedia (https://protacdb.weizmann.ac.il/ptcb/main)], and examined their overlap with degradable targets from MAPD predictions. Although some PROTAC targets were missed (Table S4), MAPD successfully identified 77% (50/65) of kinase targets (Figure 4A), supporting its ability in



distinguishing degradable kinases from other kinases. In addition, MAPD recovered 14 known PROTAC targets that were not identified by Donovan et al. [20] (Figure 4A), which highlights how computational methods can be complementary to high-throughput experimental approaches.

A binder of the target protein is required in the design of TPD molecules, so the propensity of a POI to be bound by a small molecule, also called ligandability, is relevant to the tractability of the POI by TPD molecules. Here, we leveraged knowledge of existing small molecules to refine MAPD predictions. A protein is considered ligandable if it has at least one ligand reported in PROTAC-DB [11], PROTACpedia (https://protacdb.weizmann.ac.il/ptcb/main), DrugBank [44], ChEMBL [45], or SLCABPP (ligandable cysteine database) [46] (Figure S3D). Out of the 519 ligandable kinases, MAPD identified 350 degradable kinases, including 74% (253/342) PROTACtable targets and 97 targets specifically identified by MAPD (Figure 4C). PROTACtable was introduced in a recent perspective article [47] that qualitatively assigned tractable TPD targets based on ligand records in

ChEMBL and a rule-based approach that only considers whether certain protein annotations are available. We observed a significantly lower ubiquitination potential of PROTACtable-specific targets than MAPD-specific targets (Figure 4D). For example, MAP3K4, a PROTACtable-specific target, has only one reported Ub site despite being a particularly long protein with 103 lysines [40] (Figure 4E). In contrast, the MAPD-specific target, AGK, is extensively ubiquitinated despite its short length (Figure 4E). Experimental data showed that AGK was degraded sufficiently by multi-kinase degraders [20] while MAP3K4 was not despite its strong target engagement by a multi-kinase degrader [21]. These examples highlight a potential advantage of MAPD by quantitatively assessing protein degradability.

In total, MAPD identified 132 disease-relevant kinase targets, including kinases encoded by 72 cancer genes in OncoKB and 60 kinases associated with other diseases reported in the ClinVar database [48,49] (Figure S3E). These kinases could be prospective targets for degrader development (Table S4). The most degradable kinases include PROTAC targets in PROTAC-DB [11] and PROTACpedia (https://protacdb.weizmann.ac.il/ptcb/main), such as CDK2, PLK1, CSNK1A1, CDK6, CDK9, and CDK4, and other promising targets, such as TK1, CHEK1, MAPK8, and AURKB that are degraded by multi-kinase degraders [20,21] (Figure 4F).

MAPD predicts proteome-wide degradability

We hypothesized that MAPD might also predict the degradability of non-kinase proteins. To test this, we collected 65 non-kinase targets with publicly available degraders reported in PROTAC databases [PROTAC-DB [11] and PROTACpedia (https://protacdb.weizmann.ac.il/ptcb/main)]. These PROTAC targets had significantly higher MAPD scores than other drug targets from DrugBank [44] (**Figure 5**A). To further corroborate this finding, we collected a list of TFs (such as IKZF1 and IKZF3) that are frequently degraded by thalidomide analog (IMiD)-based degraders [50]. The MAPD scores of the IMiD targets showed a notable correlation with their observed frequencies of degradation ($P = 0.0223$) (Figure 5B). Additional TFs have also been targeted by TPD molecules [8] such as degraders for ARs [12,51] and ERs [52] that have entered into clinical trials. With the exception of BCL6 which has few reported Ub sites, MAPD correctly predicted the high degradability of most TF PROTAC targets (Figure 5C). Taken together, these results indicate that MAPD could be generalizable to POIs outside of the kinome.

Given the robust performance of MAPD, we next applied MAPD proteome-wide to systematically score all proteins outside of the kinome. MAPD predicted 2648 degradable targets out of 4137 ligandable non-kinase proteins (Figure S4A and B), which was twofold more than PROTACtable [47] (Figure 5D). The MAPD-specific targets again had significantly higher levels of ubiquitination potential than the PROTACtable-specific targets (Figure 5E). We further identified 832 disease-relevant non-kinase targets that are amenable to TPD (Figure S4C; Table S4). Of these, 206 proteins are considered to be encoded by oncogenes by OncoKB, and 626 proteins are associated with other human diseases reported in the ClinVar database [48,49] (Figure S4C). The top predicted degradable targets include known PROTAC targets, such as MDM2 and BCL-XL (BCL2L1), and other potentially degradable targets. DHFR, one of the top-ranking targets, has been successfully degraded by the 2-(4-carboxyphenyl)-4,4,5,5-tetramethylimidazoline-1-oxyl-3-oxide (PTIO) in a ubiquitination-dependent manner [53]. RHOA, RHOB, and RHOC are also predicted to be degradable, which have been previously reported to be degraded by F-box-intracellular single-domain antibodies [54]. These results suggest potential opportunities for future TPD efforts (Figure 5F).

The E2 accessibility of Ub sites is associated with protein degradability

Given that ubiquitination potential was the most important feature in MAPD, we hypothesized that structural properties of Ub sites could be informative of protein degradability. To test this hypothesis, we first calculated the proportion of lysine/Ub sites that had certain structural properties (Table S5), such as secondary structure, relative solvent

Figure 4 MAPD shows good performance in predicting kinase degradability

A. Venn diagram showing the overlap between kinases degraded by multi-kinase degraders from the study by Donovan and colleagues [20], PROTAC targets reported in PROTAC databases [including PROTAC-DB [11] and PROTACpedia (https://protacdb.weizmann.ac.il/ptcb/main)], and degradable kinases identified by MAPD. **B.** Scatter plot showing the Spearman's correlation between MAPD scores and frequencies of degradation of all degradable kinases from the study by Donovan and colleagues [20]. **C.** Venn diagram showing the overlap between degradable kinases identified by MAPD, PROTACtable kinases [47], and ligandable kinases. **D.** Box plot showing ubiquitination potential of MAPD-specific targets and PROTACtable-specific targets. ****, $P < 0.0001$. **E.** Lollipop diagram showing the reported Ub sites in MAP3K4 (PROTACtable-specific target) and AGK (MAPD-specific target). The number in the circles indicates the number of references for each Ub site in PhosphoSitePlus [40] and the blank circle indicates that only one reference is available. The blue text near the circle indicates the location of the Ub site. **F.** Heatmap showing annotations of the top 50 predicted degradable kinases, with MAPD scores shown at the top. 'PROTAC-DB' and 'PROTACpedia' indicate whether a kinase has a developed degrader reported in the respective databases. The 'multi-kinase degrader' indicates whether a protein is degraded by a multi-kinase degrader. 'DrugBank' indicates whether a protein has FDA-approved drugs recorded in the DrugBank database [44]. 'ChEMBL' indicates whether a protein has ligands recorded in the ChEMBL database [45]. 'Electrophiles' indicate whether a protein has ligandable cysteines from the SLCABPP [46]. The 'OncoKB' indicates whether a protein is considered to be encode by a cancer gene in the OncoKB database [48]. The 'ClinVar' indicates whether the protein is associated with a disease in the ClinVar database [49]. FDA, United States Food and Drug Administration; SLCABPP, streamlined cysteine activity-based protein profiling.

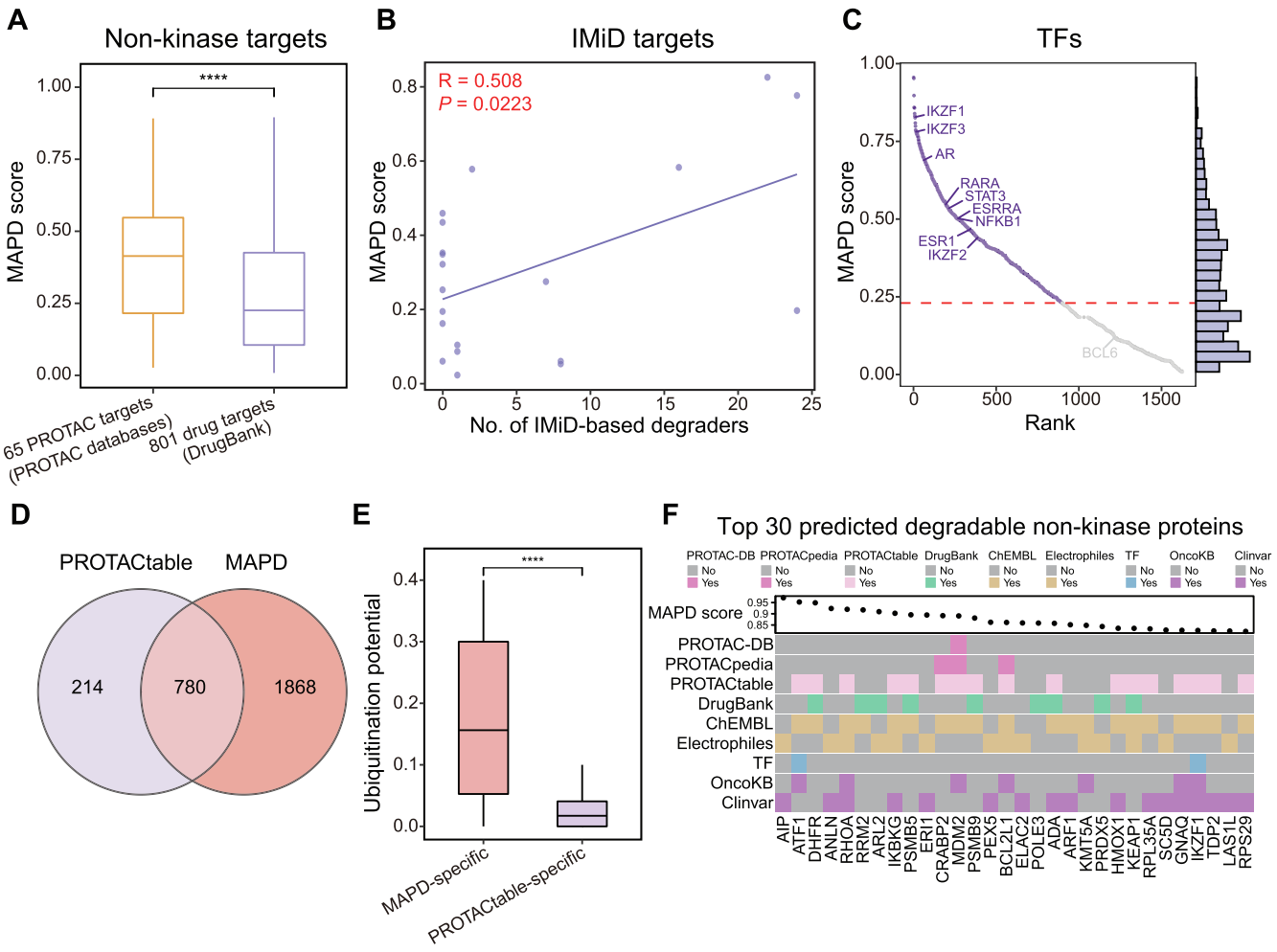
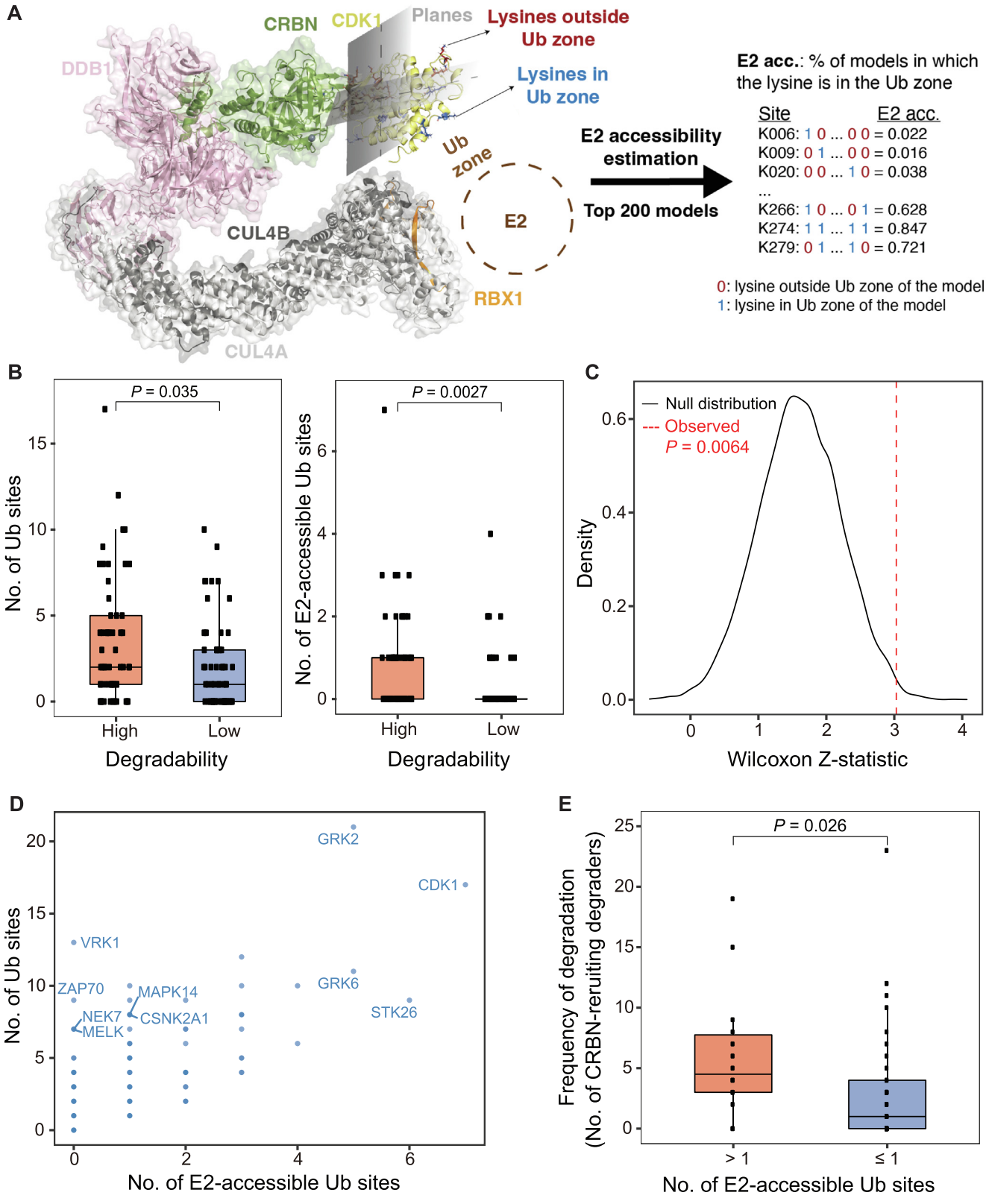


Figure 5 MAPD predicts proteome-wide degradability

A. Box plot showing the MAPD scores of non-kinase PROTAC targets from PROTAC databases [including PROTAC-DB [11] and PROTACpedia (https://protacdb.weizmann.ac.il/ptcb/main)] and other non-kinase drug targets from DrugBank [44]. ****, $P < 0.0001$. **B.** Scatter plot showing the correlation between MAPD scores and the frequencies of degradation of IMiD targets by CRBN-recruiting degraders from the study by Donovan and colleagues [20]. **C.** Ranked dot plot showing the MAPD scores of human TFs. TFs with reported degraders are labeled in the figure. The histogram at right shows the distribution of MAPD scores of all human TFs and the red dashed line shows the threshold for identifying degradable proteins by MAPD. **D.** Venn diagram showing the overlap of degradable non-kinase proteins between MAPD predictions and PROTACtable genome [47]. **E.** Box plot showing the ubiquitination potential in MAPD-specific targets and PROTACtable-specific targets. ****, $P < 0.0001$. **F.** Heatmap showing annotations of the top 30 predicted degradable non-kinase proteins, with MAPD scores shown at the top. 'PROTAC-DB' and 'PROTACpedia' annotations indicate whether a kinase has a developed degrader reported in the respective databases. 'DrugBank' indicates whether a protein has FDA-approved drugs recorded in the DrugBank database [44]. 'ChEMBL' indicates whether a protein has ligands recorded in the ChEMBL database [45]. 'Electrophiles' indicate whether a protein has ligandable cysteines from the SLCABPP [46]. 'OncoKB' indicates whether a protein is considered as a protein encoded by a cancer gene in the OncoKB database [48]. 'ClinVar' indicates whether the protein is associated with a disease in the ClinVar database [49]. TF, transcription factor; IMiD, immunomodulatory drug.

accessibility, or flexibility (as defined by B-factor). We then statistically assessed the differences of these properties between highly and lowly degradable kinases (Wilcoxon rank-sum test). Consistent with our previous results for lysine residues in general, lysine sites stratified by any of the structural properties did not show a consistent association with protein degradability (Figure S5A–C). Notably, the proportion of solvent-accessible lysine sites, defined based on either experimental structures or AlphaFold predicted structures [55], did not show a significant difference between highly and lowly degradable kinases (Figure S5D and E). Among annotated secondary structures, the Ub sites in loop regions showed a modestly higher association with protein degradability than total Ub sites (Figure S5A). However, neither relative solvent accessibility nor flexibility of Ub sites seemed to improve the association with protein degradability (Figure S5B and C). These data suggest that the local structural properties of a Ub site provide limited information for predicting protein degradability.

We next investigated the property of Ub sites that facilitates the transfer of ubiquitin from the attached E2 enzyme to the


A
CRBN
CDK1
Planes
Lysines outside Ub zone
Lysines in Ub zone
DDB1
Ub zone
E2
CUL4B
RBX1
CUL4A
E2 accessibility estimation
Top 200 models
E2 acc.: % of models in which the lysine is in the Ub zone
Site E2 acc.
K006: 1 0 ... 0 0 = 0.022
K009: 0 1 ... 0 0 = 0.016
K020: 0 0 ... 1 0 = 0.038
...
K266: 1 0 ... 0 1 = 0.628
K274: 1 1 ... 1 1 = 0.847
K279: 0 1 ... 1 0 = 0.721
0: lysine outside Ub zone of the model
1: lysine in Ub zone of the model
B
P = 0.035
No. of Ub sites
High
Low
Degradability
P = 0.0027
No. of E2-accessible Ub sites
High
Low
Degradability
C
Null distribution
Observed
P = 0.0064
Density
Wilcoxon Z-statistic
D
GRK2
CDK1
VRK1
ZAP70
MAPK14
NEK7
CSNK2A1
MELK
GRK6
STK26
No. of Ub sites
No. of E2-accessible Ub sites
E
P = 0.026
Frequency of degradation (No. of CRBN-reruiting degraders)
> 1
≤ 1
No. of E2-accessible Ub sites

POI in degrader-mediated ternary complexes. We reasoned that Ub sites accessible to the E2 enzyme might be relevant to protein degradability. To quantify the accessibility of Ub sites to the E2 enzyme, we first modeled the PROTAC-induced ternary complex using a computational docking method (Figure S6A and B). We benchmarked the method using distinct experimental complex structures involving five protein targets and two E3 ligases. For each case, we docked the complex structure of a protein target and its chemical ligand to an E3 ligase without specifying a PROTAC and selected the 200 top-scoring and feasible docked models for examination (Figure S6A and B). In all cases, the docking method captured near-native binding modes (Figure S7A), including two distinct conformations of $BRD4_{BD1}$ with CRBN in particular (Table S6). As the multi-kinase degraders were based on several different parental kinase-binding moieties, we eliminated target ligands from the docking method and further examined their performance. In all but one case, near-native models were still captured (Table S6; Figure S7B). We then computationally docked 251 target kinases with experimental structures onto CRBN–IMiD and examined the accessibility of Ub sites to the E2 enzyme, of which the position is estimated by using an RBX1 fragment (Figure S6A–C). Since the CUL4 arm is highly flexible, the bound E2 can transfer ubiquitin to any site in a broad ubiquitination zone [56], hence all Ub sites in the spatial quadrant facing the E2 can be considered accessible to the E2 (**Figure 6**A, Figure S6C). We then defined E2 accessibility as the fraction of top-scoring models in which the Ub site was accessible to the E2 enzyme (Figure 6A, Figure S6C; Table S5). In comparison to the total number of Ub sites in the structure of the POI, the E2-accessible Ub sites showed a more significant positive association with protein degradability (Figure 6B, Figure S8A). In contrast, the number of E2-accessible lysine residues on the POIs did not show a significant association with POI degradability (Figure S8A and B). To further assess whether E2 accessibility is independently useful, we randomly shuffled reported Ub sites among all available lysine residues within a protein. Consistent with our initial finding, E2-accessible Ub sites were significantly more associated with protein degradability than expected based on the total number of Ub sites in each protein ($P = 0.0064$; Figure 6C).

We observed an overall positive correlation between the total number of Ub sites and E2-accessible Ub sites on kinases (Figure 6D) and noticed some POIs with outlier levels of E2-accessible and total Ub sites. For example, CDK1 had a high fraction of E2-accessible Ub sites (Figure 6D, Figure S8C), consistent with its frequent degradation by multi-kinase degraders [20]. Therefore, we hypothesize that similar proteins, such as GRK2, GRK6, and STK26, are promising targets for developing future TPD drugs if they had drug–target engagement (Figure 6D). In contrast, some kinases, such as VRK1, ZAP70, NEK7, and MAPK14, had a low number of E2-accessible Ub sites, despite having a high number of total Ub sites (Figure 6D). As expected, these kinases have a significantly lower frequency of degradation by CRBN-recruiting multi-kinase as measured by Donovan and colleagues [20] (Figure 6E).

Finally, we created an interactive web platform (https://mapd.cistrome.org) to present protein-intrinsic features, protein ligandability, disease associations, and MAPD predictions. We also incorporated the E2 accessibility of Ub sites in kinases with experimental structures. This platform could enable rational prioritization of degradable targets for developing degraders by the TPD community. Moreover, we implemented MAPD as an R package (https://github.com/liulab-dfci/MAPD), allowing researchers to extend our analysis when more chemoproteomic profiling data and/or protein features are available in the future.

Discussion

Despite the growing number of small molecule degraders, it remains challenging to predict which proteins are tractable to this approach. In this study, we investigated the degradability of kinases and their correlation with features intrinsic to protein targets. By developing a machine learning model, MAPD, we identified five features predictive of kinase degradability, including the ubiquitination potential, acetylation potential, phosphorylation potential, protein half-life, and protein length.

Figure 6 E2 accessibility of Ub sites is associated with protein degradability

A. Diagram showing how to estimate the accessibility of lysine/Ub sites to E2 enzyme in the degrader-induced ternary complex. The model of CDK1 (PDB: 4Y72) is docked to the CRBN-lenalidomide structure (PDB: 5FQD), which is shown as an example. The E3 ubiquitin ligase complex consists of CRBN, DDB1, CUL4A, and CUL4B, shown in green, pink, light gray, and gray, respectively. The CDK1 is the target protein, shown in yellow. The RBX1 fragment (shown in orange) is used to estimate the position of the E2 enzyme and the corresponding Ub zone in the target protein. Lysine/Ub sites in the Ub zone are estimated by drawing two planes with respect to the positions of CRBN and the target kinase. The sites lying in the quadrant facing the putative position of the E2 enzyme, estimated by the placement of RBX1, are considered accessible. The predicted E2-accessible and E2-inaccessible lysine residues are highlighted in blue and red, respectively. For each target protein, 200 top-scoring feasible models are selected for evaluating the accessibility of lysine residues to the E2 enzyme. For each Ub site, the fraction of feasible models with the site in the Ub zone is estimated as its E2 accessibility. **B.** Box plots showing the associations of kinase degradability with the total number of Ub sites (left) and the number of E2-accessible Ub sites (right) in the kinases, repectively. The E2-accessible Ub sites are defined as the Ub sites lying in the Ub zone of more than 50% of feasible models. **C.** Density plot showing the null distribution of Wilcoxon Z-statistics generated by shuffling Ub sites among all lysine residues 10,000 times. The red dashed line indicates the observed Wilcoxon Z-statistic representing the association between protein degradability and the number of E2-accessible Ub sites. **D.** Dot plot showing the total number of resolved Ub sites and the number of E2-accessible Ub sites. **E.** Box plot showing the number CRBN-recruiting degraders that degrade kinases with high (> 1) and low (≤ 1) levels of E2-accessible Ub sites. All kinases involved in this analysis have at least two reported Ub sites, which reduces the confounding effect derived from the difference in the total number of Ub sites.

Systematic benchmarking indicates that MAPD can predict kinase degradability and is potentially applicable to proteins outside the kinome. By integrating MAPD predictions and ligand information of POIs, we prioritized disease-associated degradable proteins as TPD drug targets.

Ternary complex formation is thought to be the most important factor in determining the degradability of protein targets. However, our analysis found that protein degradability can also be heavily influenced by protein-intrinsic features, especially the protein's endogenous ubiquitination potential. By modeling the structural relationship between target proteins and the E2 enzyme, we found that protein degradability is highly correlated with the availability of E2-accessible Ub sites. Although much remains to be understood about how the E2-accessible Ub sites influence TPD, our observations suggest that considering the protein-intrinsic features can help in assessing protein targets before a TPD drug discovery project.

Our study has several limitations. First, due to the high importance of ubiquitination potential in predicting protein degradability, inaccurate quantification of ubiquitination potential can potentially bias the prediction. For instance, BRD4, a well-known PROTAC target (https://protacdb.weizmann.ac.il/ptcb/main), has incomplete Ub site annotation, with reported Ub sites enriched in the BRD4-BD1 domain (Figure S9). Quantifying the ubiquitination potential of BRD4 at a full-protein basis leads to its low ubiquitination potential and thus a low MAPD prediction score. Second, the nature of the ubiquitin modification (mono-ubiquitination or poly-ubiquitination with varying topologies) is critical for the fate of a protein, yet this information is absent from di-Gly-based proteomics datasets [57]. Although this can be a limitation, it could also be a potential strength since what ubiquitination potential quantifies is the principal potential to be ubiquitinated, which in the case of processive ligases such as $CRL4^{CRBN}$, may lead to poly-ubiquitin chains on lysines that endogenously would be mono-ubiquitinated. Third, the activity of deubiquitinating enzymes (DUBs) that remove ubiquitin from a protein could be important for determining degradability, although existing systematic datasets were not yet sufficient to improve prediction (Figure S10A–E). Lastly, the lack of data, including matching protein-intrinsic features from the same cell lines and experimental protein structures, limits machine learning and computational docking, thereby limiting the prediction. To summarize, careful consideration of the feature data is important when interpreting the prediction results. More extensive proteomic profiling of protein-intrinsic features and induced protein degradation by multi-target degraders in disease-relevant cell lines or tissues could offer better opportunities to address these problems in the future.

Our study also reveals several future research directions to advance the field. First, computational and experimental studies investigating why certain lysines seem more susceptible to ubiquitination than others could improve the predictions for degradability by MAPD. Second, a more thorough investigation of structural features using AlphaFold predicted structures could facilitate understanding the relationship between Ub sites and induced protein degradation and further guide the selection of rational protein targets and E3 ligases to develop TPD drugs. Finally, we envision that future computational methods will not only improve the prediction of protein degradability but also predict the functional consequence of the degradation of disease-causing proteins.

Materials and methods

Kinase degradability data

We collected 151 quantitative proteomic datasets measuring the changes of protein abundance in response to the treatment of 85 unique multi-kinase degraders (degraders with allosteric linkers are excluded) [20]. We used the limma package to perform differential protein expression analysis comparing the degrader-treated samples with the samples treated with dimethyl sulfoxide (DMSO). For each protein, we calculated the frequency of degradation as the number of experiments in which the protein is significantly down-regulated [fold change (FC) > 1.25 and $P < 0.01$]. Furthermore, to aggregate the results of multiple replicates for each degrader, we aggregated $\log_2$ FC from replicate experiments using Stouffer's Z-score and corresponding P values using Fisher's method. We used a threshold of Stouffer's Z-score $< \log_2 1.5$ and Fisher's $P < 0.01$ to identify significantly down-regulated proteins as degradable proteins and then counted the number of unique degraders that can degrade each protein. We collected 5 KiNativ profiling data and 2 KinomeScan data from published studies [20,21], which profiled the target engagement of five multi-kinase degraders, including TL12-186, SK-3-91, SB1-G-187, DB0646, and WH-10417-099 [20,21]. A KinomeScan score < 15 or a KiNativ score > 35 indicates strong drug–target engagement.

Definition of highly and lowly degradable kinases

We defined highly degradable kinases as those degraded by at least five different multi-kinase degraders (50 kinases), and lowly degradable kinases that were engaged by at least one multi-kinase degrader, quantified in more than 10% underlying global proteomic experiments, but not degraded (76 kinases). The highly degradable kinases and lowly degradable kinases are used throughout the study to investigate the association between protein degradability and protein-intrinsic features.

Protein-intrinsic features

We built more than 42 protein-intrinsic features spanning PTMs [40], protein stability generated from GPS profiling [37–39], protein half-life [34–36], PPIs [58], protein expression, protein detectability, protein length, and others.

PTM features

We collected all available PTM sites from the PhosphoSitePlus database (accessed on February 17, 2021) [40]. We also collected ubiquitination sites from PLMD database [41]. For each type of PTM, such as ubiquitination, we estimated the potential of proteins being modified by calculating the fraction of

relevant amino acid residues in a protein (*e.g.*, lysine residues) that have a corresponding reported PTM site (*e.g.*, Ub site). We also included the fraction of each likely modified amino acid as additional features, such as LysRatio indicating the fraction of lysine residue in a protein.

Protein half-life and protein stability features

We downloaded protein half-lives in seven different cell types (B cells, NK cells, monocytes, hepatocytes, neurons, HeLa, and NIH3T3) from published studies [34–36]. We additionally collected seven GPS profiling data from three studies [37–39], which include the stability of full-length proteins in HEK293T cell lines treated with DMSO, MLN4924, dominant-negative CRL4, or dominant-negative CRL3 and stability of N-terminome and C-terminome peptides of the human proteome. All protein half-life data and GPS data were cross-referred for imputing the missing data. The imputation was done by using the impute.knn function (KNN) with default parameters in the impute R package.

PPI and protein complex

We downloaded PPIs from the STRING database [58] and retrieved the high-confidence PPIs using an arbitrary cutoff of experimental score > 100 and combined score > 200. The degree of each protein in the PPI network was calculated as an estimation of the likelihood of the protein interacting with others. Additionally, curated protein complex annotations were downloaded from the CORUM database [59] and the number of distinct protein complexes associated with each protein was taken as the estimation of the likelihood of a protein being complexed *in vivo*.

Gene and protein expression data

We downloaded RNA-seq data of MOLT4 from the Gene Expression Omnibus (GEO: GSE79253) [60]. RNA expression values were normalized as logarithm transcripts per million (TPM). We retrieved quantitative proteomic data of MOLT4 cell lines from a previous study by Donovan and colleagues [20]. Relative protein abundances were logarithm normalized and centered with a median value of zero per sample. The missing values in the proteomic data were imputed using the impute::impute.knn function (KNN) with Cancer Cell Line Encyclopedia (CCLE) proteomic data as reference [61].

Protein detectability

We took the frequency of detection of proteins in the proteomic datasets from the study by Donovan et al. [20] as the estimation of protein detectability by mass spectrometry.

Other features

We retrieved 20,381 reviewed human protein sequences and their length from the UniProtKB database (accessed in January, 2021). We downloaded intrinsically disordered regions (IDRs) from the MobiDB database [62], which includes manually curated annotations and predicted disorder regions. We ranked the IDR annotations based on the four types of evidence, including curated-disorder-priority, derived-missing_residues-th_90, derived-mobile_residues-th_90, and prediction-disorder-mobidb_lite. For each protein, duplicate IDRs were removed for downstream analysis.

Pairwise correlation of protein-intrinsic features

We computed pairwise Spearman's correlation of protein-intrinsic features and clustered the features based on the correlation matrix using hierarchical clustering with Euclidean distance measure and complete linkage. The data are visualized using the ComplexHeatmap R package [63].

Association between protein degradability and features intrinsic to protein targets

We tested each feature's difference in 50 highly degradable kinases and 76 lowly degradable kinases using the wilcox.test function in R and computed the Z-statistics using the wilcoxonZ function in the rcompanion R package. We used the same method to test the association between protein degradability and protein-intrinsic features in each kinase family.

Development of MAPD model

We sought to build a classification model to predict protein degradability from intrinsic protein features. We tried six different machine learning models, including NB (naivebayes), KNN, LR (LiblineaR), svmLinear (kernlab), svmRadial (kernlab), and RF (randomForest). For each model, we performed feature selection and then selected the best model trained on a set of best performing features.

Forward feature selection

We performed recursive forward feature selection for six machine learning methods separately. In each iteration, we add a feature that improves the model performance most. The performance is computed as AUPRC based on 20-fold cross-validation. This process is stopped when the addition of a new feature does not further improve the performance.

Feature importance

We evaluated the importance of features in MAPD using the varImp function in the caret R package, which computes the feature importance on permuted out-of-bag samples based on the mean decrease in the accuracy.

Performance evaluation

To evaluate the performance of each model involved in the study, we collected prediction scores of all proteins from cross-validation and computed the AUROC using the roc function from the pROC package and AUPRC using the pr.curve from the PRROC R package.

Single feature evaluation

For each individual feature, we trained a LR model. For the combination of features, we trained RF models. Finally, we

compared the model performance based on 20-fold cross-validation.

Final model training for predictions outside of the kinome

We used the caret package for parameter tuning and final model training. We evaluated the model tuning parameters based on leave-one-out cross-validation (method = "LOOCV" in the trainControl function), with the F1 score as performance metric (metric = "F" in the train function, summaryFunction = prSummary in the trainControl function). With the optimal parameters (mtry = 2), we trained a final RF model including 20,000 trees (ntree = 20,000) with 5 minimum node sizes (nodesize = 5).

Prediction

We predicted the degradability of all human proteins using the final RF model. For kinases included in the training, we took the average prediction scores collected from three repeated 20-fold cross-validation. Based on the cross-validation, we chose a cutoff (0.2327) that leads to the highest F1 score. A protein is predicted to be degradable if it has a MAPD score greater than the cutoff. To account for potential biases from missing feature data, we scored the feature completeness for each protein using a weighted sum score with the formula: $C = \sum_{x \in F} varImp(x) * I_A(x)$. The F variable represents the feature set, and x represents each feature in the feature set. The function $varImp(x)$ denotes the scaled feature importance of x and the indicator function $I_A(x)$ denotes whether x is from actual data (1 = actual, 0 = imputed). The C represents the feature completeness, with a 0–1 range. A score of 1 indicates all features are from actual data, and a score of 0 indicates all features are imputed.

Degradable proteins

We collected PROTAC targets with reported degraders in the PROTAC-DB (accessed on May 27, 2021) [11] and/or the PROTACpedia (accessed on July 8, 2021; https://protacdb.weizmann.ac.il/ptcb/main). For evaluation purposes, the targets from the study by Donovan and colleagues [20] were removed from the PROTAC databases (including PROTAC-DB and PROTACpedia). This resulted in 65 kinases and 65 proteins outside of the kinome. From the study by Donovan and colleagues, we collected 217 kinases degraded by at least one multi-kinase degrader as 'degraded' and all the others detected in the same datasets as 'not degraded' [20]. We collected 1336 PROTACtable targets, including the clinical precedence targets, discovery opportunity targets, and literature precedence targets from the PROTACtable genome [47]. We collected 24 IMiD targets from published studies [50] and assessed their frequencies of degradation by 68 CRBN-recruiting multi-kinase degraders from the study by Donovan and colleagues [20].

Protein family

We downloaded the human kinase/kinase-related proteins from four different resources, including KinMap [64], KinBase, a study by Donovan et al. [20], and a review article [65]. We collected 1626 human TFs from a review article [66].

Protein ligandability

We downloaded the cysteine reactivity data from the SLCABPP [46] and assessed protein ligandability using the number of compounds with a competition ratio greater than 4. Besides, we collected protein ligands from the ChEMBL (accessed on July 23, 2021) and DrugBank databases [44,45]. For any proteins degraded by a multi-kinase degrader or with a ligand recorded in ChEMBL (accessed on July 23, 2021), DrugBank, or SLCABPP, we considered it as a ligandable target.

Protein–disease associations

We considered a protein as a cancer driver if it is encoded by an oncogene reported in the OncoKB or it is predicted to be encoded by an oncogene by 20/20+ algorithm. 20/20+ analysis was performed on the aggregated pan-cancer dataset with default parameters. Genes with an oncogene score greater than 0.5 are considered oncogenes. To annotate potential protein targets associated with other human diseases, we also downloaded the variant–disease association from the ClinVar database [49] (accessed on April 20, 2021). For quality control, we removed annotations of likely loss-of-function variants, including indel, deletion, insertion, and microsatellite, as well as some uncertain annotations with keywords like 'conflicting', 'protective', 'uncertain', 'benign', and 'not'. This resulted in 3415 proteins associated with human diseases reported in the ClinVar database.

Structural properties of lysine residues and Ub sites

We downloaded protein structures of human models or homology models from the Protein Data Bank (PDB), SWISS-MODEL [67], and ModPipe [68]. Detailed data cleaning and processing have been described by Tokheim and colleagues [69]. Protein structures were analyzed using the Define Secondary Structure of Proteins (DSSP) program [70] in the bio3d R package, which returns the solvent accessibility and secondary structure of each residue.

Benchmarking protein–protein docking method

Structures of unbound E3 ligases and targets with the parental compounds were extracted from an existing benchmark [71]. As we were trying to determine all feasible PROTAC-mediated ternary complexes between the E3 ligase and the target (not one induced by a particular compound), redundant pairs were eliminated from the benchmark. Using Rosetta v.3.12 [72] and RosettaDock v.4.0 [73], we performed 5000 independent local docking of the target and the E3 ligase with the respective parental compounds in place with different starting points and random initial perturbations (trans = 3 Å and rot = 8) along with random spins along an axis joining the center of masses of the two proteins. Models were evaluated by the interface score metric (I_sc) and the 200 lowest-scoring models were selected for further evaluation. In each model, if a cylinder of radius 1 Å and length < 14 Å could be constructed between the atoms of the parental moieties where the linker is anchored with less than 2 protein backbone

or compound atoms (except neighboring atoms) inside the cylinder, we estimated that there exists a free path to build a linker, and hence fit the PROTAC. To evaluate the models, we used the CAPRI criteria [74] with medium- and high-quality models deemed near-native and sufficient for identifying Ub sites facing the E2 enzyme. To test the importance of docking with the target ligand, we removed the ligand and re-docked the proteins.

CRBN–kinase docking

We downloaded the protein structures of 323 kinases from the PDB. In cases where multiple structures were available, the largest structure was chosen. They were aligned to CDK2 (PDB: 1AQ1), a reference kinase, to ensure that the kinase domain was present. 251 kinase structures were alignable with root-mean-square deviation less than 3.5 Å near the ATP-binding pocket. Next, the aligned kinases were positioned in an arbitrary (but similar) orientation around the ligand-binding pocket of the CRBN–lenalidomide structure (PDB: 5FQD) [6]. They were then docked using the protein docking protocol described in benchmarking. Structural biology applications were compiled and configured by SBGrid [75].

E2 accessibility of lysine residues

We assessed the accessibility of solvent-exposed lysine residues to the E2 enzyme by calculating the fraction of protein–protein docking models among the 200 lowest-scoring models that could fit a PROTAC and in which the lysine residues are in the ubiquitination zone of the E2 enzyme. All lysines with an atom having > 2.5 Å^2 exposed surface area were considered solvent-exposed. The ability of the ternary complex to fit a PROTAC was assessed by aligning CDK2 with CDK4 inhibitor (PDB: 1GIJ) [76] to the kinase and calculating if there was a free path available between the N3 atom lenalidomide and C26 atom of the CDK4 inhibitor to build a linker as described in benchmarking. To assess which lysine residue lies within the ubiquitination zone of the E2 enzyme, we constructed two planes to split up space into quadrants. The 'vertical' plane passes through half the distance between the CRBN edge facing the kinase and the center-of-mass of the kinase. The 'horizontal' plane is approximately perpendicular to the vertical plane and passes through the center-of-mass of the kinase. The lysine residues lying in the quadrant facing the putative position of the E2 enzyme are considered accessible. Finally, if the lysine residue is more than 60 Å away from the lenalidomide or the C_α–C_β vector points in the direction opposite of the putative E2 site, the residue was considered inaccessible.

Association between protein degradability and characteristics of Ub sites

We first counted each protein's lysine residues/Ub sites in different secondary structures (coil, strand, and loop), and then tested whether there is a difference between highly degradable and lowly degradable kinases using the Wilcoxon Z-statistics. Similarly, we assessed the associations between kinase degradability and the number of lysine residues/Ub sites with a specific range of solvent accessibility or B-factor. A positive Wilcoxon Z-statistic indicates the positive correlation between kinase degradability and the number of Ub sites/lysine residues in the proteins.

We also tested the association between kinase degradability and the number of E2-accessible Ub sites/lysine residues (E2 accessibility greater than a specific threshold) in each protein. To further demonstrate the specific importance of E2-accessible Ub sites, we randomly shuffled the Ub sites among all lysine residues and re-evaluated the association between kinase degradability and the number of E2-accessible Ub sites in each kinase. We generated a null distribution by repeating the shuffling process 10,000 times and calculated the P value by counting the percentage of shuffling that led to a higher Wilcoxon Z-statistic than the observed Wilcoxon Z-statistics.

Association between deubiquitination and kinase degradability

For each protein, we estimated the deubiquitination activity using four different metrics: 1) the highest protein-level correlation between the protein and all DUBs based on proteomic data on CCLE cancer cell lines [61]; 2) the number of DUBs interacted with the protein based on curated DUB–substrate interaction data from the UbiBrowser database [77]; 3) the differential expression of the protein upon treatment of a pan-DUB inhibitor UbVS [78]; and 4) the maximal down-regulation of the protein after DUB inhibitions with 11 different compounds [78,79]. We statistically tested the difference of each DUB feature between highly and lowly degradable kinases (Wilcoxon rank-sum test). We also trained a RF model based on the four features and evaluated its performance based on 20-fold cross-validation.

Code availability

The source codes for reproducing the model are available in the MAPD repository at https://github.com/liulab-dfci/MAPD. The source codes for reproducing the figures are available in the Degradability2021 repository at https://github.com/liulab-dfci/Degradability2021.

Data availability

The datasets supporting the conclusions of this study are accessible at http://mapd.cistrome.org.

CRediT author statement

Wubing Zhang: Conceptualization, Methodology, Software, Formal analysis, Writing - original draft. **Shourya S. Roy Burman:** Conceptualization, Methodology, Formal analysis, Writing - original draft, Funding acquisition. **Jiaye Chen:** Software, Formal analysis. **Katherine A. Donovan:** Resources, Writing - review & editing. **Yang Cao:** Methodology. **Chelsea Shu:** Formal analysis. **Boning Zhang:** Writing - review & editing. **Zexian Zeng:** Methodology. **Shengqing Gu:** Methodology. **Yi Zhang:** Methodology. **Dian Li:** Software. **Eric S. Fischer:** Conceptualization, Writing - review & editing, Supervision, Funding acquisition. **Collin Tokheim:** Conceptualization, Methodology, Writing - original draft, Supervision,

Funding acquisition. **X. Shirley Liu:** Conceptualization, Writing - review & editing, Supervision, Funding acquisition. All authors have read and approved the final manuscript.

Competing interests

X. Shirley Liu is a cofounder, board member, and CEO of GV20 Therapeutics. Eric S. Fischer is a founder, member of the scientific advisory board (SAB), and equity holder of Civetta Therapeutics, Lighthorse Therapeutics, Proximity Therapeutics, and Neomorph Inc (board of directors), SAB member and equity holder in Avilar Therapeutics and Photys Therapeutics, and a consultant to Astellas, Sanofi, Novartis, Deerfield, and EcoR1 capital. The Fischer laboratory receives or has received research funding from Novartis, Deerfield, Ajax, Interline, and Astellas. Katherine A. Donovan is a consultant to Kronos Bio and Neomorph Inc. All the other authors declared no competing interests.

Acknowledgments

This study was supported by grants from the Breast Cancer Research Foundation (Grant No. BCRF-19-100 to X. Shirley Liu), the Mark Foundation for Cancer Research (Mark Foundation Emerging Leader Award; Grant No. 19-001-ELA to Eric S. Fischer), the National Institutes of Health (NIH; Grant Nos. R01CA218278 and R01CA214608 to Eric S. Fischer), and Cancer Research Institute (Irvington Postdoctoral Fellowship; Grant No. CRI 3442 to Shourya S. Roy Burman), USA. Collin Tokheim is a Damon Runyon Fellow supported by the Damon Runyon Cancer Research Foundation, USA (Grant No. DRQ-04-20). We acknowledge the Research Computing Group at Harvard Medical School and Dana-Farber Cancer Institute for cluster time. We also would like to thank Dr. Chris Sander for his valuable suggestions on this study.

Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.gpb.2022.11.008.

ORCID

ORCID 0000-0003-2233-299X (Wubing Zhang)
ORCID 0000-0001-9274-9104 (Shourya S. Roy Burman)
ORCID 0000-0001-9654-1968 (Jiaye Chen)
ORCID 0000-0002-8539-5106 (Katherine A. Donovan)
ORCID 0000-0002-1925-2123 (Yang Cao)
ORCID 0000-0002-0463-1809 (Chelsea Shu)
ORCID 0000-0001-8516-137X (Boning Zhang)
ORCID 0000-0002-3905-3244 (Zexian Zeng)
ORCID 0000-0002-4200-0864 (Shengqing Gu)
ORCID 0000-0002-7453-6188 (Yi Zhang)
ORCID 0000-0002-5374-7314 (Dian Li)
ORCID 0000-0001-7337-6306 (Eric S. Fischer)
ORCID 0000-0003-1395-5378 (Collin Tokheim)
ORCID 0000-0003-4736-7339 (X. Shirley Liu)

References

[1] Glickman MH, Ciechanover A. The ubiquitin-proteasome proteolytic pathway: destruction for the sake of construction. Physiol Rev 2002;82:373–428.
[2] Baumeister W, Walz J, Zühl F, Seemüller E. The proteasome: paradigm of a self-compartmentalizing protease. Cell 1998;92:367–80.
[3] Burslem GM, Crews CM. Small-molecule modulation of protein homeostasis. Chem Rev 2017;117:11269–301.
[4] Liu J, Farmer Jr JD, Lane WS, Friedman J, Weissman I, Schreiber SL. Calcineurin is a common target of cyclophilin-cyclosporin A and FKBP-FK506 complexes. Cell 1991;66:807–15.
[5] Sakamoto KM, Kim KB, Kumagai A, Mercurio F, Crews CM, Deshaies RJ. Protacs: chimeric molecules that target proteins to the Skp1-Cullin-F box complex for ubiquitination and degradation. Proc Natl Acad Sci U S A 2001;98:8554–9.
[6] Petzold G, Fischer ES, Thomä NH. Structural basis of lenalidomide-induced CK1α degradation by the $CRL4^{CRBN}$ ubiquitin ligase. Nature 2016;532:127–30.
[7] Burslem GM, Smith BE, Lai AC, Jaime-Figueroa S, McQuaid DC, Bondeson DP, et al. The advantages of targeted protein degradation over inhibition: an RTK case study. Cell Chem Biol 2018;25:67–77.e3.
[8] Henley MJ, Koehler AN. Advances in targeting 'undruggable' transcription factors with small molecules. Nat Rev Drug Discov 2021;20:669–88.
[9] Ito T, Ando H, Suzuki T, Ogura T, Hotta K, Imamura Y, et al. Identification of a primary target of thalidomide teratogenicity. Science 2010;327:1345–50.
[10] Krönke J, Udeshi ND, Narla A, Grauman P, Hurst SN, McConkey M, et al. Lenalidomide causes selective degradation of IKZF1 and IKZF3 in multiple myeloma cells. Science 2014;343:301–5.
[11] Weng G, Shen C, Cao D, Gao J, Dong X, He Q, et al. PROTAC-DB: an online database of PROTACs. Nucleic Acids Res 2021;49:D1381–7.
[12] Petrylak DP, Gao X, Vogelzang NJ, Garfield MH, Taylor I, Moore MD, et al. First-in-human phase I study of ARV-110, an androgen receptor (AR) PROTAC degrader in patients (pts) with metastatic castrate-resistant prostate cancer (mCRPC) following enzalutamide (ENZ) and/or abiraterone (ABI). J Clin Oncol 2020;38:3500.
[13] Flanagan JJ, Qian Y, Gough SM, Andreoli M, Bookbinder M, Cadelina G, et al. ARV-471, an oral estrogen receptor PROTAC degrader for breast cancer [abstract]. Cancer Res 2019;79:P5-04-18.
[14] He Y, Koch R, Budamagunta V, Zhang P, Zhang X, Khan S, et al. DT2216—a Bcl-xL-specific degrader is highly active against Bcl-xL-dependent T cell lymphomas. J Hematol Oncol 2020;13:95.
[15] Hansen JD, Correa M, Nagy MA, Alexander M, Plantevin V, Grant V, et al. Discovery of CRBN E3 ligase modulator CC-92480 for the treatment of relapsed and refractory multiple myeloma. J Med Chem 2020;63:6648–76.
[16] Wang ES, Verano AL, Nowak RP, Yuan JC, Donovan KA, Eleuteri NA, et al. Acute pharmacological degradation of Helios destabilizes regulatory T cells. Nat Chem Biol 2021;17:711–7.
[17] Powell CE, Du G, Che J, He Z, Donovan KA, Yue H, et al. Selective degradation of GSPT1 by cereblon modulators identified via a focused combinatorial library. ACS Chem Biol 2020;15:2722–30.
[18] Dobrovolsky D, Wang ES, Morrow S, Leahy C, Faust T, Nowak RP, et al. Bruton tyrosine kinase degradation as a therapeutic strategy for cancer. Blood 2019;133:952–61.
[19] Mullard A. Targeted protein degraders crowd into the clinic. Nat Rev Drug Discov 2021;20:247–50.

[20] Donovan KA, Ferguson FM, Bushman JW, Eleuteri NA, Bhunia D, Ryu S, et al. Mapping the degradable kinome provides a resource for expedited degrader development. Cell 2020;183:1714–31.e10.
[21] Huang HT, Dobrovolsky D, Paulk J, Yang G, Weisberg EL, Doctor ZM, et al. A chemoproteomic approach to query the degradable kinome using a multi-kinase degrader. Cell Chem Biol 2018;25:88–99.e6.
[22] Bondeson DP, Smith BE, Burslem GM, Buhimschi AD, Hines J, Jaime-Figueroa S, et al. Lessons in PROTAC design from selective degradation with a promiscuous warhead. Cell Chem Biol 2018;25:78–87.e5.
[23] Xiong Y, Donovan KA, Eleuteri NA, Kirmani N, Yue H, Razov A, et al. Chemo-proteomics exploration of HDAC degradability by small molecule degraders. Cell Chem Biol 2021;28:1514–27.
[24] Gadd MS, Testa A, Lucas X, Chan KH, Chen W, Lamont DJ, et al. Structural basis of PROTAC cooperative recognition for selective protein degradation. Nat Chem Biol 2017;13:514–21.
[25] Nowak RP, DeAngelo SL, Buckley D, He Z, Donovan KA, An J, et al. Plasticity in binding confers selectivity in ligand-induced protein degradation. Nat Chem Biol 2018;14:706–14.
[26] Drummond ML, Williams CI. In silico modeling of PROTAC-mediated ternary complexes: validation and application. J Chem Inf Model 2019;59:1634–44.
[27] Zaidman D, Prilusky J, London N. PRosettaC: Rosetta based modeling of PROTAC mediated ternary complexes. J Chem Inf Model 2020;60:4894–903.
[28] Bai N, Miller SA, Andrianov GV, Yates M, Kirubakaran P, Karanicolas J. Rationalizing PROTAC-mediated ternary complex formation using Rosetta. J Chem Inf Model 2021;61:1368–82.
[29] Farnaby W, Koegl M, Roy MJ, Whitworth C, Diers E, Trainor N, et al. BAF complex vulnerabilities in cancer demonstrated via structure-based PROTAC design. Nat Chem Biol 2019;15:672–80.
[30] Lecker SH, Goldberg AL, Mitch WE. Protein degradation by the ubiquitin-proteasome pathway in normal and disease states. J Am Soc Nephrol 2006;17:1807–19.
[31] Cheng B, Ren Y, Cao H, Chen J. Discovery of novel resorcinol diphenyl ether-based PROTAC-like molecules as dual inhibitors and degraders of PD-L1. Eur J Med Chem 2020;199:112377.
[32] McCoull W, Cheung T, Anderson E, Barton P, Burgess J, Byth K, et al. Development of a novel B-cell lymphoma 6 (BCL6) PROTAC to provide insight into small molecule targeting of BCL6. ACS Chem Biol 2018;13:3131–41.
[33] Roy MJ, Winkler S, Hughes SJ, Whitworth C, Galant M, Farnaby W, et al. SPR-measured dissociation kinetics of PROTAC ternary complexes influence target degradation rate. ACS Chem Biol 2019;14:361–8.
[34] Mathieson T, Franken H, Kosinski J, Kurzawa N, Zinn N, Sweetman G, et al. Systematic analysis of protein turnover in primary cells. Nat Commun 2018;9:689.
[35] Schwanhäusser B, Busse D, Li N, Dittmar G, Schuchhardt J, Wolf J, et al. Global quantification of mammalian gene expression control. Nature 2011;473:337–42.
[36] Zecha J, Meng C, Zolg DP, Samaras P, Wilhelm M, Kuster B. Peptide level turnover measurements enable the study of proteoform dynamics. Mol Cell Proteomics 2018;17:974–92.
[37] Emanuele MJ, Elia AEH, Xu Q, Thoma CR, Izhar L, Leng Y, et al. Global identification of modular cullin-RING ligase substrates. Cell 2011;147:459–74.
[38] Yen HCS, Elledge SJ. Identification of SCF ubiquitin ligase substrates by global protein stability profiling. Science 2008;322:923–9.
[39] Yen HCS, Xu Q, Chou DM, Zhao Z, Elledge SJ. Global protein stability profiling in mammalian cells. Science 2008;322:918–23.
[40] Hornbeck PV, Zhang B, Murray B, Kornhauser JM, Latham V, Skrzypek E. PhosphoSitePlus, 2014: mutations, PTMs and recalibrations. Nucleic Acids Res 2015;43:D512–20.
[41] Zhang W, Tan X, Lin S, Gou Y, Han C, Zhang C, et al. CPLM 4.0: an updated database with rich annotations for protein lysine modifications. Nucleic Acids Res 2022;50:D451–9.
[42] Xu G, Jaffrey SR. Proteomic identification of protein ubiquitination events. Biotechnol Genet Eng Rev 2013;29:73–109.
[43] Wu C, Ba Q, Lu D, Li W, Salovska B, Hou P, et al. Global and site-specific effect of phosphorylation on protein turnover. Dev Cell 2021;56:111–24.e6.
[44] Wishart DS, Feunang YD, Guo AC, Lo EJ, Marcu A, Grant JR, et al. DrugBank 5.0: a major update to the DrugBank database for 2018. Nucleic Acids Res 2018;46:D1074–82.
[45] Mendez D, Gaulton A, Bento AP, Chambers J, De Veij M, Félix E, et al. ChEMBL: towards direct deposition of bioassay data. Nucleic Acids Res 2019;47:D930–40.
[46] Kuljanin M, Mitchell DC, Schweppe DK, Gikandi AS, Nusinow DP, Bulloch NJ, et al. Reimagining high-throughput profiling of reactive cysteines for cell-based screening of large electrophile libraries. Nat Biotechnol 2021;39:630–41.
[47] Schneider M, Radoux CJ, Hercules A, Ochoa D, Dunham I, Zalmas LP, et al. The PROTACtable genome. Nat Rev Drug Discov 2021;20:789–97.
[48] Chakravarty D, Gao J, Phillips SM, Kundra R, Zhang H, Wang J, et al. OncoKB: a precision oncology knowledge base. JCO Precis Oncol 2017;2017:PO.17.00011.
[49] Landrum MJ, Chitipiralla S, Brown GR, Chen C, Gu B, Hart J, et al. ClinVar: improvements to accessing data. Nucleic Acids Res 2020;48:D835–44.
[50] Sievers QL, Petzold G, Bunker RD, Renneville A, Słabicki M, Liddicoat BJ, et al. Defining the human C2H2 zinc finger degrome targeted by thalidomide analogs through CRBN. Science 2018;362:eaat0572.
[51] Han X, Zhao L, Xiang W, Qin C, Miao B, Xu T, et al. Discovery of highly potent and efficient PROTAC degraders of androgen receptor (AR) by employing weak binding affinity VHL E3 ligase ligands. J Med Chem 2019;62:11218–31.
[52] Bihani T, Patel HK, Arlt H, Tao N, Jiang H, Brown JL, et al. Elacestrant (RAD1901), a selective estrogen receptor degrader (SERD), has antitumor activity in multiple ER breast cancer patient-derived xenograft models. Clin Cancer Res 2017;23:4793–804.
[53] Cai Z, Lu Q, Ding Y, Wang Q, Xiao L, Song P, et al. Endothelial nitric oxide synthase-derived nitric oxide prevents dihydrofolate reductase degradation via promoting S-nitrosylation. Arterioscler Thromb Vasc Biol 2015;35:2366–73.
[54] Bery N, Keller L, Soulié M, Gence R, Iscache AL, Cherier J, et al. A targeted protein degradation cell-based screening for nanobodies selective toward the cellular RHOB GTP-bound conformation. Cell Chem Biol 2019;26:1544–58.e6.
[55] Varadi M, Anyango S, Deshpande M, Nair S, Natassia C, Yordanova G, et al. AlphaFold protein structure database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. Nucleic Acids Res 2022;50:D439–44.
[56] Fischer ES, Scrima A, Böhm K, Matsumoto S, Lingaraju GM, Faty M, et al. The molecular basis of CRL4DDB2/CSA ubiquitin ligase architecture, targeting, and activation. Cell 2011;147:1024–39.
[57] Fulzele A, Bennett EJ. Ubiquitin diGLY proteomics as an approach to identify and quantify the ubiquitin-modified proteome. Methods Mol Biol 2018;1844:363–84.
[58] Szklarczyk D, Gable AL, Lyon D, Junge A, Wyder S, Huerta-Cepas J, et al. STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. Nucleic Acids Res 2019;47:D607–13.
[59] Giurgiu M, Reinhard J, Brauner B, Dunger-Kaltenbach I, Fobo G, Frishman G, et al. CORUM: the comprehensive resource of

mammalian protein complexes—2019. Nucleic Acids Res 2019;47:D559–63.
[60] Winter GE, Mayer A, Buckley DL, Erb MA, Roderick JE, Vittori S, et al. BET bromodomain proteins function as master transcription elongation factors independent of CDK9 recruitment. Mol Cell 2017;67:5–18.e19.
[61] Nusinow DP, Szpyt J, Ghandi M, Rose CM, McDonald 3rd ER, Kalocsay M, et al. Quantitative proteomics of the cancer cell line encyclopedia. Cell 2020;180:387–402.e16.
[62] Potenza E, Di Domenico T, Walsh I, Tosatto SCE. MobiDB 2.0: an improved database of intrinsically disordered and mobile proteins. Nucleic Acids Res 2015;43:D315–20.
[63] Gu Z, Eils R, Schlesner M. Complex heatmaps reveal patterns and correlations in multidimensional genomic data. Bioinformatics 2016;32:2847–9.
[64] Eid S, Turk S, Volkamer A, Rippmann F, Fulle S. KinMap: a web-based tool for interactive navigation through human kinome data. BMC Bioinformatics 2017;18:16.
[65] Buljan M, Ciuffa R, van Drogen A, Vichalkovski A, Mehnert M, Rosenberger G, et al. Kinase interaction network expands functional and disease roles of human kinases. Mol Cell 2020;79:504–20.e9.
[66] Lambert SA, Jolma A, Campitelli LF, Das PK, Yin Y, Albu M, et al. The human transcription factors. Cell 2018;172:650–65.
[67] Waterhouse A, Bertoni M, Bienert S, Studer G, Tauriello G, Gumienny R, et al. SWISS-MODEL: homology modelling of protein structures and complexes. Nucleic Acids Res 2018;46:W296–303.
[68] Pieper U, Webb BM, Dong GQ, Schneidman-Duhovny D, Fan H, Kim SJ, et al. ModBase, a database of annotated comparative protein structure models and associated resources. Nucleic Acids Res 2014;42:D336–46.
[69] Tokheim C, Bhattacharya R, Niknafs N, Gygax DM, Kim R, Ryan M, et al. Exome-scale discovery of hotspot mutation regions in human cancer using 3D protein structure. Cancer Res 2016;76:3719–31.
[70] Kabsch W, Sander C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers 1983;22:2577–637.
[71] Weng G, Li D, Kang Y, Hou T. Integrative modeling of PROTAC-mediated ternary complexes. J Med Chem 2021;64:16271–81.
[72] Leman JK, Weitzner BD, Lewis SM, Adolf-Bryfogle J, Alam N, Alford RF, et al. Macromolecular modeling and design in Rosetta: recent methods and frameworks. Nat Methods 2020;17:665–80.
[73] Marze NA, Roy Burman SS, Sheffler W, Gray JJ. Efficient flexible backbone protein-protein docking for challenging targets. Bioinformatics 2018;34:3461–9.
[74] Méndez R, Leplae R, De Maria L, Wodak SJ. Assessment of blind predictions of protein-protein interactions: current status of docking methods. Proteins 2003;52:51–67.
[75] Morin A, Eisenbraun B, Key J, Sanschagrin PC, Timony MA, Ottaviano M, et al. Collaboration gets the most out of software. Elife 2013;2:e01456.
[76] Ikuta M, Kamata K, Fukasawa K, Honma T, Machida T, Hirai H, et al. Crystallographic approach to identification of cyclin-dependent kinase 4 (CDK4)-specific inhibitors by using CDK4 mimic CDK2 protein. J Biol Chem 2001;276:27548–54.
[77] Wang X, Li Y, He M, Kong X, Jiang P, Liu X, et al. UbiBrowser 2.0: a comprehensive resource for proteome-wide known and predicted ubiquitin ligase/deubiquitinase-substrate interactions in eukaryotic species. Nucleic Acids Res 2022;50:D719–28.
[78] Rossio V, Paulo JA, Chick J, Brasher B, Gygi SP, King RW. Proteomics of broad deubiquitylase inhibition unmasks redundant enzyme function to reveal substrates and assess enzyme specificity. Cell Chem Biol 2021;28:487–502.e5.
[79] Bushman JW, Donovan KA, Schauer NJ, Liu X, Hu W, Varca AC, et al. Proteomics-based identification of DUB substrates using selective inhibitors. Cell Chem Biol 2021;28:78–87.e3.