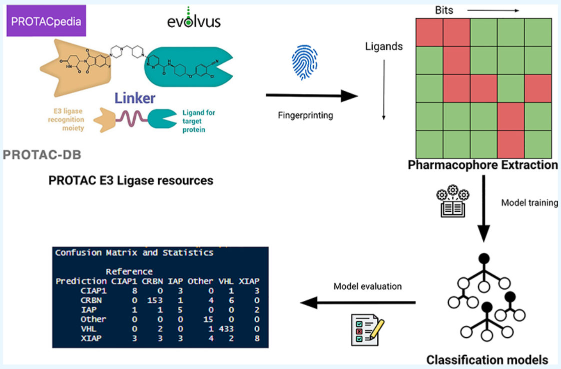
Pharmacophore-Based Machine Learning Model To Predict Ligand Selectivity for E3 Ligase Binders

Reagon Karki,* Yojana Gadiya, Philip Gribbon, and Andrea Zaliani

Cite This: *ACS Omega* 2023, 8, 30177−30185
Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: E3 ligases are enzymes that play a critical role in ubiquitin-mediated protein degradation and are involved in various cellular processes. Pharmacophore analysis is a useful approach for predicting E3 ligase binding selectivity, which involves identifying key chemical features necessary for a ligand to interact with a specific protein target cavity. While pharmacophore analysis is not always sufficient to accurately predict ligand binding affinity, it can be a valuable tool for filtering and/or designing focused libraries for screening campaigns. In this study, we present a fast and an inexpensive approach using a pharmacophore fingerprinting scheme known as ErG, which is used in a multi-class machine learning classification model. This model can assign the correct E3 ligase binder to its known E3 ligase and predict the probability of each molecule to bind to different E3 ligases. Practical applications of this approach are demonstrated on commercial libraries such as Asinex for the rational design of E3 ligase binders. The scripts and data associated with this study can be found on GitHub at https://github.com/Fraunhofer-ITMP/E3_binder_Model.



```
Confusion Matrix and Statistics

          Reference
Prediction CIAP1 CRBN IAP Other VHL XIAP
     CIAP1     8    0   3     0   1    3
     CRBN      0  153   1     4   6    0
     IAP       1    1   5     0   0    2
     Other     0    0   0    15   0    0
     VHL       0    2   0     1 433    0
     XIAP      3    3   3     4   2    8
```

INTRODUCTION

E3 ligases are a class of enzymes that are involved in ubiquitin-mediated protein degradation, and they play a critical role in many cellular processes, including cell cycle regulation, DNA repair, and apoptosis. Selective targeting of E3 ligases has emerged as a promising strategy for developing novel therapeutics for various diseases, including cancer.[1] Thus, predicting the target binding selectivity for E3 ligases using molecular fingerprint analysis can be useful in designing focused libraries for screening campaigns. With the help of this, we can not only enrich existing libraries with high probability candidates but, in the long run, also define geometric and interaction rules for each E3 ligase. Overall, this binding selectivity will facilitate rational design of future proteolysis targeting chimera (PROTAC) and novel molecular glues.

Molecular fingerprints are capable of encoding structural information and physico-chemical properties of molecules at various dimensions. 2D fingerprints, in particular, have been used widely in various scenarios and their performances are reported to outperform 3D fingerprints.[2] Their applications range from the basic task of identification of similar compounds in a library of molecules to advanced methods in drug design, such as binding pocket detection,[3] protein-ligand interaction,[4] toxicity prediction,[5] and drug repurposing.[6] Among the four main categories of 2D fingerprints (i.e., key-based, topological, circular and pharmacophore), the latter captures detailed properties, such as the number of hydrogen donors/acceptors, charges, and aromatic/lipophilic moieties required to interact with a target of interest.[7] Therefore, they are more suited to characterize interaction-selectivity modeling challenges. This can be done by analyzing the structure of known X-ray complexes and identifying common chemical features that are critical for binding.[8,9] In the case of E3 ligases, several key structural features are important for ligand binding such as the presence of a zinc-binding domain and a substrate-binding site.[10,11] However, it is important to note that pharmacophore analysis is not always sufficient to accurately predict ligand binding affinity, as there might be other factors that influence binding that are not captured by the pharmacophore model (e.g., excluded volumes).

In this work, we present a fast and inexpensive approach where ligands of known E3 ligases are described by a simple and effective pharmacophore fingerprinting scheme, known as Extended Reduced Graph (ErG).[12,13] Each ErG bit forms the basis for a multi-class classification model where singular E3 ligase target proteins are used as labels. This is the first example of such a classification approach in the E3 ligase field. The resultant statistical model showed an accuracy of 93.8% and

Received: April 26, 2023
Accepted: July 6, 2023
Published: August 9, 2023

https://doi.org/10.1021/acsomega.3c02803

thus is able to assign the correct E3 ligase binder to previously known E3 ligase. As a result of this, such an approach allows us to computationally screen and filter large compound libraries by predicting the probability of each compound to bind to different E3 ligases. We validated this model on commercial libraries for the rational design of E3 ligase binders.

■ METHODS

The first step was to gather a dataset of known E3 ligase ligands (with their respective targets), which would serve as the training set for the machine learning model. To achieve this, we merged data from three PROTAC resources, namely, PROTAC-DB 2.0,[14] PROTACpedia,[15] and a commercial subset of Proximity Degraders Database (PDD),[16] where the E3 ligase binding components of original active PROTACs are structurally identified and assigned. This yielded a total of 643 unique ligands. The merging of PDD to the classic PROTAC databases enabled the enrichment of ligases mentioned within patent documents alongside those found in the scientific literature. Additionally, we expanded the chemical space of E3 ligase binders with 19 ligands specific to DNA damage-binding protein 1 DDB1 (UniProt: Q16531) and CUL4-associated factor homolog 1 (DCAF1) (UniProt: Q9M086) which were not present in PROTAC-derived collections.[17] It is essential to note that only human protein targets were considered while building this dataset.

The summary of the dataset of unique 662 compounds alongside the 17 E3 ligases targets is shown in Figure 1. Since certain target classes (i.e., DDB1 and CUL4 associated factors: 15 (DCAF15) (UniProt: Q66K64), 11 (DCAF11) (UniProt: Q8TEB1), and 16 (DCAF16) (UniProt: Q9NXF7), MDM2 proto-oncogene (MDM2) (UniProt: Q00987), Aryl hydrocarbon receptor (AHR) (UniProt: P35869), Baculoviral IAP repeat-containing 3 (cIAP2/BIRC3) (UniProt: Q13489), Ring finger proteins 4 (RNF4) (UniProt: Q13489) and 114 (RNF114) (UniProt: Q9Y508), Fem-1 homolog B (FEM1B) (UniProt: Q9UK73), ubiquitin-protein ligase E3 component n-recognin 1 (UBR1) (UniProt: Q8IWV7), and Cullin 4A (CUL4A) (UniProt: Q13619)) had less than 20 compounds each, we clustered them together in a common class called “Other.” This grouping was an approach to reduce the effect of imbalance on the E3 ligase set. As a result, we identified 6 target classes for the resultant 662 E3 ligase ligands i.e., Cereblon (CRBN) (UniProt: Q96SW2), Von Hippel−Lindau tumor suppressor (VHL) (UniProt: P40337), X-linked inhibitor of apoptosis (XIAP) (UniProt: P98170), Baculoviral IAP repeat-containing 2 (CIAP1/BIRC2) (UniProt: Q13490), Islet amyloid polypeptide (IAP) (UniProt: P10997), and “Other.”

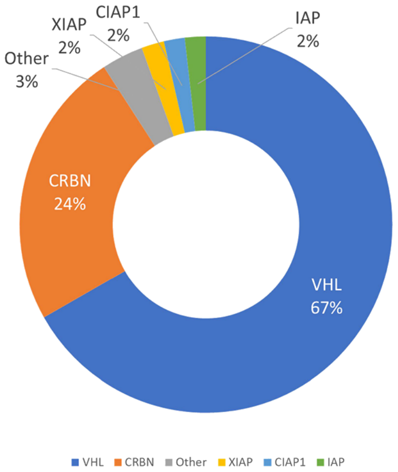


Figure 1. Representation of the E3 ligases and relative percentage of compound ligands collected. Von Hippel−Lindau tumor suppressor (VHL) and cereblon (CRBN) are the most studied E3 ligase targets with 442 and 159 ligands, respectively, in the collected dataset. Moreover, X-linked inhibitor of apoptosis (XIAP), baculoviral IAP repeat containing 2 (cIAP1/BIRC2), and islet amyloid polypeptide (IAP/IAPP) showed a consistent distribution with around 12 ligands each.

Next, we extracted the candidate pharmacophores for each ligand with the help of the ErG pharmacophoric fingerprint as implemented within Molecular Operating Environment (MOE) (version 2022.02).[12,18] ErG follows a reduced graph-based schema for extraction of pharmacophores from the compounds, unlike the subgraph schema (Figure S1). Other fingerprint schemes like the MACCS keys (MACCS),[19] RDKit fingerprint (version of DayLight fingerprint),[20] Avalon,[21] and extended connectivity fingerprint, up to four bonds (ECFP4)[22] have been utilized for comparisons. Figures S2 and S3 depict the pharmacophores extracted using fingerprints ECFP4 and RDKit, respectively.

The pharmacophoric information contained in the ErG bit names and the connections which can be easily created with relevant 3D structural biology information are exemplarily exploited in Figure 2. In this example, CRBN is found to be complexed with SALL4 and (*S*)-5-hydroxy thalidomide (CHEMBL468), generating a map of ligand interactions (PDB:7BQV). Four interacting atoms of (*S*)-5-hydroxy thalidomide can be focused and their relative fingerprint-related distances are identified. Imposing non-null values for those interatomic distances could be a straightforward filtering option.

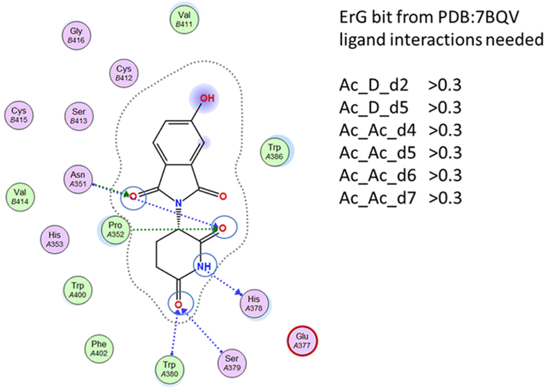


Figure 2. Using the X-ray structure of cereblon-bound ligand interactions from PDB (PDB:7BQV), non-null ErG bits can be extracted from four pharmacophoric atoms (circled are: 3 acceptors [Ac] with arrows directed toward the atoms and one donor [D] with arrows directed away from the atom). The figure has been generated using MOE version 2022.02.

Following the selection of the dataset and fingerprint, we visualized the distinct target space for each ligand. To do so, we used the t-distributed stochastic neighbor embedding (t-SNE) algorithm, a nonlinear dimensionality reduction technique.[23] This reduction in the vector space will enable finding patterns between the pharmacophore features of ligands across the E3 ligase in the nonlinear space.

Machine learning models are generally referred to as "black boxes" since the reasoning behind the predictions of the model remains unclear to humans. As a result, a new field of science called eXplainable AI (XAI) grew attention, allowing humans to interpret the reasoning behind the prediction made by the model.[24] Thus, considering the importance of the XAI field, we used "transparent" machine learning models to predict the specificity of E3 ligases. We particularly used the gradient boosting model (XGBoost) which, similar to random forests, follows a bagging-based approach but, unlike the two aggregates, results in a sequential manner.[25] The model has been optimized through a random search involving different parameters such as the learning speed (eta range between 0.2 and 1) and the number of epochs [range between 1 and 10], and the final model has been cross-validated 10 times.

The fingerprint bit vectors have been expanded in single columns (descriptors/bits), and those showing variance lower than 0.2 were removed to generate a matrix of 662 rows (i.e., the total number of E3 ligands) × N columns based on the number of bits remaining after variance filtration (Table 1).

Table 1. Overview of the Total Number of Bits and the Number of Bits Preserved after Variance Filtering for each of the Fingerprints Used

fingerprint schema	number of bits	bits used (postvariance filtering)
MACCS	166	26
ECFP4	1024	78
RDKit	1024	338
Avalon	1024	224
ErG	315	73

We assumed that lower or constant variance columns should not contribute to the final models.[26] Moreover, the dataset was split in an 80:20 ratio, with 80% used to train the model and 20% for testing. Due to the high variability of ligands across the different E3 ligases, a stratification strategy was applied such that a representative set of each class of E3 ligases is present in the train and test datasets.

It is important to note that the accuracy of the model depends on many factors coming from different data sources. First, on the quality and size of the training set; second, on the choice of descriptors; and third, on the choice of machine learning algorithm and optimization parameters. Therefore, it is crucial to carefully evaluate the performances of the model using multiple and appropriate metrics and cross-validation techniques before applying them to predictions.

As mentioned previously, model evaluation enables us to understand the performance of the model. For our purpose, we use accuracy to compare the performance of each model and Cohen's kappa coefficient to compare the performance between different models.[27] Accuracy is an estimate of how good the model is in predicting the ground truth. It can be calculated as follows:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{1}$$

where TP and TN refer to true positive and true negative predictions, respectively, while FP and FN refer to false positive and false negative predictions, respectively. Cohen Kappa coefficient score, also known as the inter-rater reliability score, represents the agreement score between two compared entities. We used this score to nominate the best-performing algorithm. It is defined as follows:

$$\kappa = \frac{2 \times (\text{TP} \times \text{TN} - \text{FN} \times \text{FP})}{(\text{TP} + \text{FP}) \times (\text{FP} + \text{TN}) + (\text{TP} + \text{FN}) \times (\text{FN} + \text{TN})} \tag{2}$$

where TP and TN refer to true positive and true negative predictions, respectively, while FP and FN refer to false positive and false negative predictions, respectively.

In compliance with the XAI concept, we back-projected the most influential bits in our fingerprints to those in the molecule using RDKit.[28] For ErG, this was not needed due to its transparent schema enabling identifying the exact atoms responsible for the feature.

Data Availability. We have used KNIME[29] and R programming[30] to train and test the models described in this manuscript. The source codes and the data generated in this work are available on GitHub https://github.com/Fraunhofer-ITMP/E3_binder_Model. The structure information and representations from PDD have been omitted due to license restrictions.

■ RESULTS

A multi-class XGBoost classification model was prepared using the different fingerprints (Table 2). All the models performed well, demonstrating an accuracy of over 90%. The best model was trained with ErG fingerprints (showing an accuracy of 94% and Cohen kappa score of 0.881). Moreover, for each of the fingerprint-based models, we assessed the most dominant contributing fingerprint bits in the prediction. ErGs are the only 2D fingerprinting scheme capable of reducing 3D pharmacophore content into 2D, where the closest performing 2D fingerprint-based XGBoost model was with MACCS keys (94% accuracy and 0.877 kappa score). Interestingly, most of these fingerprint bits are hash ids that are not necessarily translated into meaningful features unlike the ErG one where the exact atom type of the contributing bits is identified. Hence, due to the high accuracy and transparent interpretability of the ErG-based model, we decided to conduct an in-depth analysis of the model in order to gain insights into the possible molecular features that may play a role in the E3-ligase ligand.

Table 2. Summary of the Performances for all the Fingerprint-Based Model Performances[a]

fingerprint	accuracy	Cohen kappa	1st most influential bit ID	2nd most influential bit ID	3rd most influential bit ID
MACCS	0.940	0.877	88	138	81
ECFP4	0.933	0.861	362	577	313
RDKit	0.910	0.810	136	844	33
Avalon	0.932	0.863	177	533	920
ErG	**0.940**	**0.881**	Ac_Ac_d4	D_D_d3	Hf_Ar_d9

[a] Highlighted in bold are the scores with the highest value. Additionally, columns mentioning the most contributing fingerprint bit for model prediction are highlighted as the 1st bit information, 2nd bit information, and 3rd bit information with the most dominant, the second most dominant, and the third most dominant bit, respectively.

Chemical Vector Space Exploration Using ErG Fingerprints. We started by visualizing the difference in the chemical space across the known E3 ligands and respective ligases in the linear and nonlinear space on the ErG-generated chemical

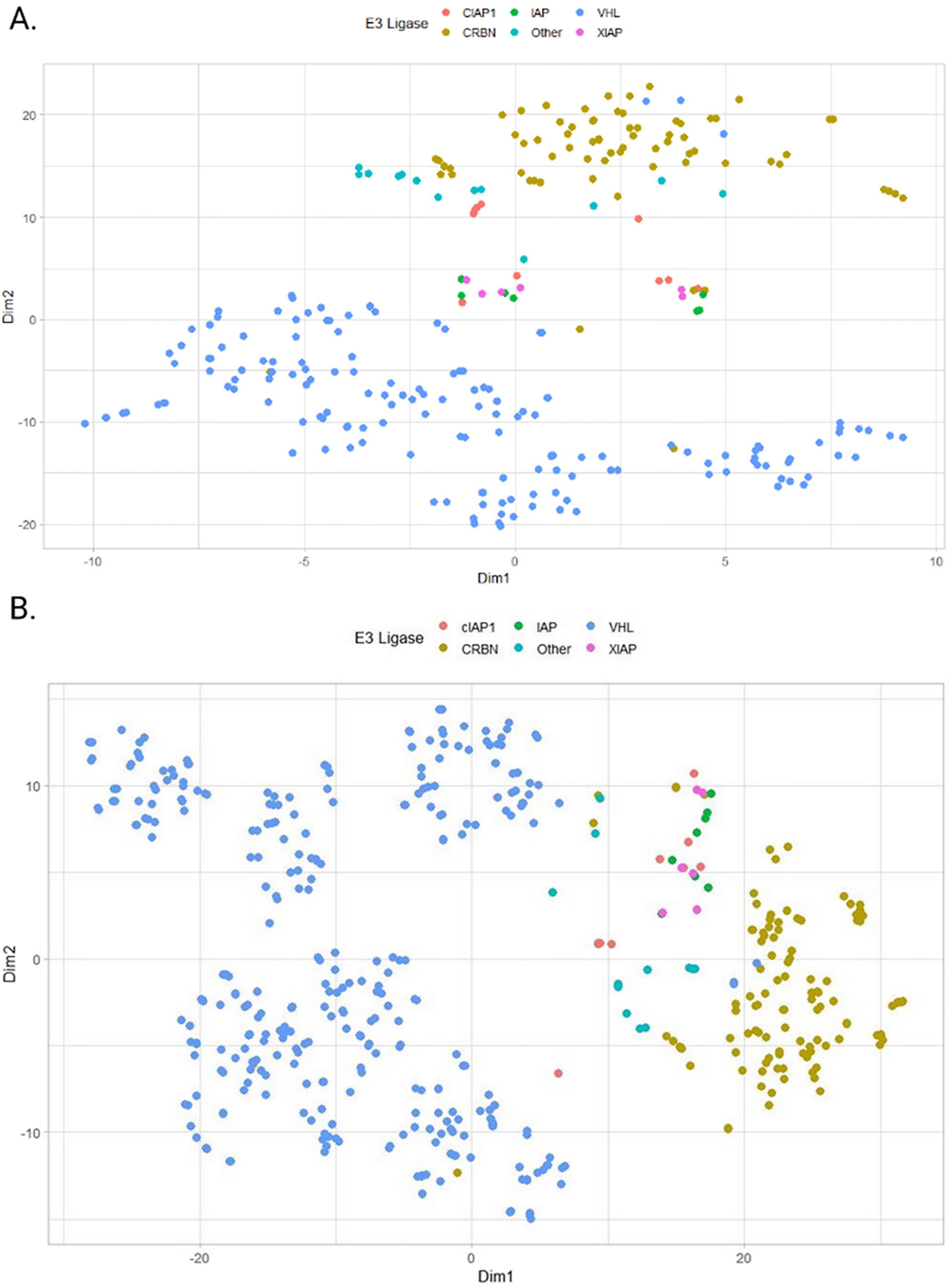


Figure 3. (A) ErG t-SNE plot for the E3 ligands colored with respect to the E3 ligases. A clear separation on the second dimension (Dim2) is noted between CRBN and some "Other" ligases and VHL, IAP, and XIAP. Interestingly, some VHL ligands are classified within the CRBN vector space indicating similarities in the respective ligand structures. (B) MACCS-derived t-SNE plot for the same dataset. While VHL and CRBN subset are pretty well separated, XIAP, CIAP1, IAP, and other groups are closer to each other.

space. Contrary to the linear space, where no distinctions were found, a clear distinction between the pharmacophoric features in the nonlinear space was seen (Figure 3). Clear separation in the vector space of ligands that bind with VHL and CRBN was found. A couple of ligands were found to be in interchanged spaces meaning that CRBN binding ligands were found in

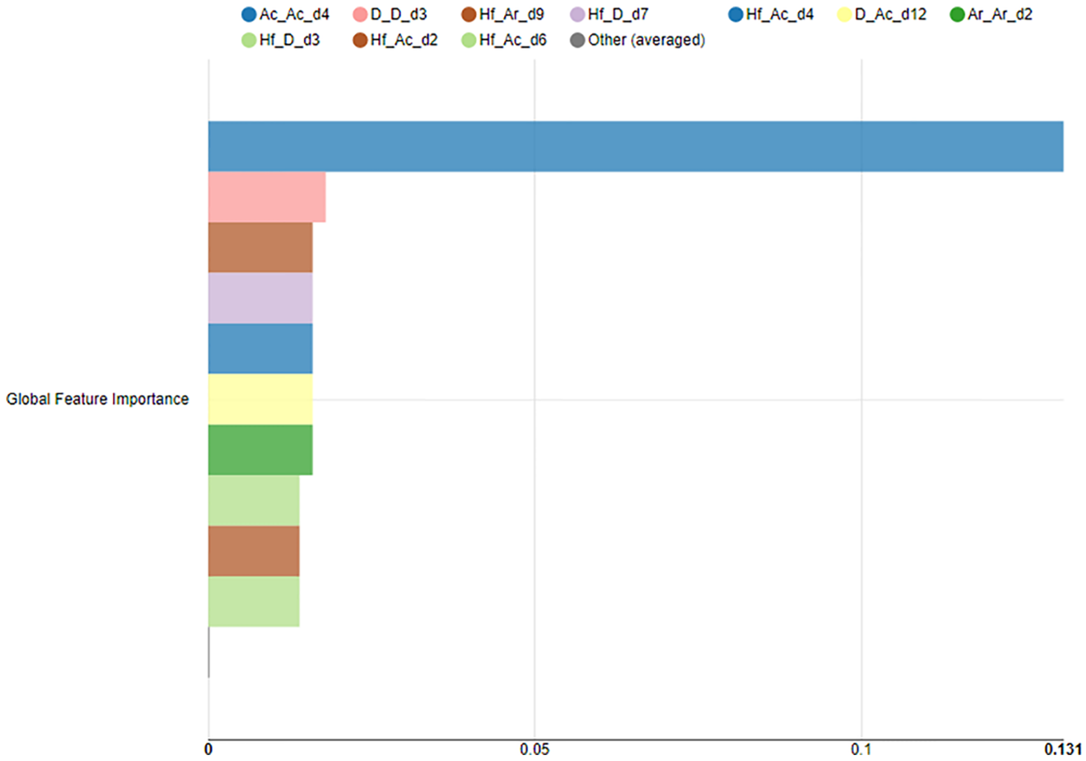


Figure 4. Top ten descriptors (ErG bits) contributing to the XGBoost model for E3 ligase selectivity predictions.

VHL-ligand space and vice versa. This raised the question of whether these ligands were mislabeled in publications (i.e., false positives) or was the dataset under study too small to understand this pattern. To answer both these questions, a larger set of known E3 ligands and their systematic mapping toward respective E3 ligases is needed. Other E3 ligases such as CIAP1, XIAP, and IAP surprisingly inhabit the vector space between the two larger clusters of VHL and CRBN. Given the low number of ligands identified for these E3 ligases, a clear difference between the space occupied by these ligands cannot be found. Using nonpharmacophoric fingerprint schemes like MACCS, the chemical space described is indeed more granular concerning the most populated labels (CRBN and VHL) but not for the less populated ones. This was another reason supporting our decision to stick with the ErG pharmacophoric description for the final model.

Insights into the Influential ErG Fingerprints. To understand the important properties of the molecule structure (i.e., the 3D pharmacophore properties), we extracted the most relevant ten ErG bits, as reported in Figure 4. Only two out of those relate to distances more than 6, suggesting that close localized pharmacophores are more important than wider ones. Six out of ten relate to distances between hydrophobic groups (Hf) and acceptor (Ac) or donor atoms (D). In the ErG scheme, every group of three or more contiguous carbon atoms are generating Hf groups, even when located in aliphatic rings. Hf_Ac_d2 is present, among others, in succinimide-like rings, but not in maleimide analogues. Interestingly, there are only two relevant ErG bits dealing with aromatic (Ar) groups, even if almost all the E3 ligands so far collected have at least one aromatic group in them. Besides this, Ar is involved in one of the only two bits dealing with higher distances (d9). This might suggest that aromatic rings can be located away from the core group of hydrogen-bond-mediated interactions. Moreover, the first two ErG bits are almost 10 times more important than the others, meaning that these two first features dictate the vast majority of selectivity recognition: indeed, between CRBN and VHL, we cover almost 91% of any training or test dataset (as seen in Figure 1).

Indeed, using just these two ErG bits (i.e., Hf-Ac_d4 and Hf-D_d3) as filters and selecting non-null values for them in the ErG description of ligands, we ended up with 92% of the entire dataset. The remaining ligands with null values with any of the two selected ErG bits are not involved with CRBN or VHL but only with the "Other" class of E3 ligases. In the confusion matrix, we report the XGBoost accuracy of the six classes of E3 ligases used (Table S1). Additionally, from our previous analysis (Figure 3), we know that some E3 ligases appear within areas where an E3 ligase label is dominant (VHL and CRBN). These ligands were found to be of the mixed nature with respect to their structures and were usually sampled from the patent description, where molecules with different scaffolds are mixed (Figure S4).

Influential ErG Fingerprints from the Ligase Perspective. Diving deeper into the ErG bits space, we tried to evaluate the statistical relevance of what has been found as the top ten relevant features and how their differences are distributed across the six ligand classes. Indeed, not all the distributions found around these ErG bits are statistically significant, but some are pretty informative (Figure 5). While Hf_Ar_d9 is clearly a footprint for VHL only and Hf_Ac_d2 is a marker for CRBN, another selective CRBN pharmacophoric point is Ac_Ac_d4 which is related to the distance between the two carbonyl oxygens in the succinimide ring and the mono or di carbonyl oxygens positioned in the attached phthalimide ring. Hf_Ar_d9 and Hf_D_d7 seem to mark CIAP1, IAP, and XIAP ligands as they are contained in hydrophobic aliphatic amino acids and amino acids, respectively. Both are well

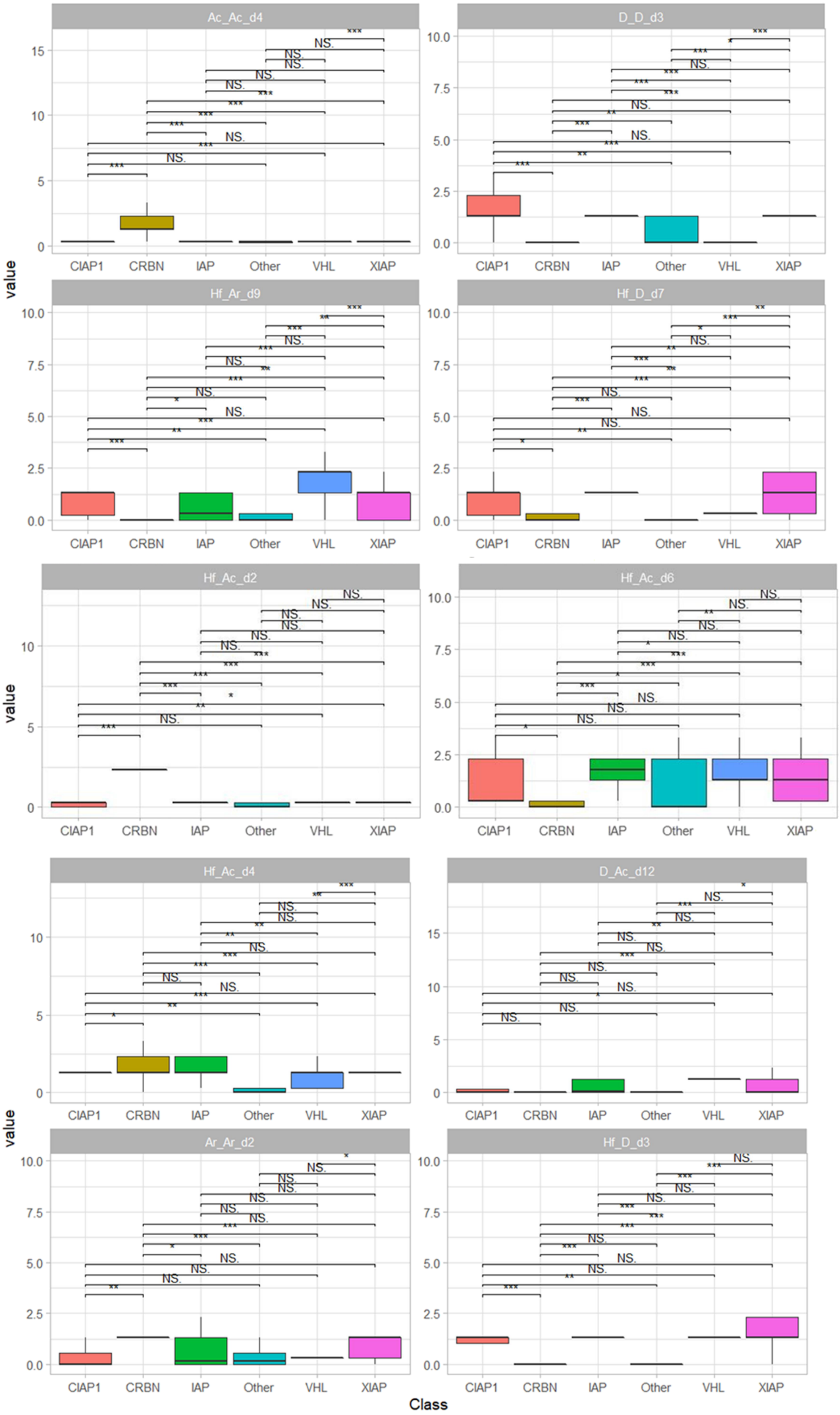


Figure 5. Box plot distribution of the top ten most influential ErG bit values according to the XGBoost model presented. While certain comparisons (t.test) are not significant (NS), some are according to calculate *p*-value labels (****p*-value <0.001, ***p*-value <0.01, **p*-value <0.1).

represented in these three groups. CIAP1, IAP, and XIAP only have the largest mean count of Hf_D_d7, while Hf_Ac_d6 has the least significant contribution according to its distribution in the six groups.

Influence of the Data Source in the ErG-Based XGBoost Model. PDD's subset with compounds manually selected from patents surely enriched the general dataset of E3 Ligase binders we used. Some molecule examples from the same E3 ligase selectivity but different chemical space is given in Figure 6.

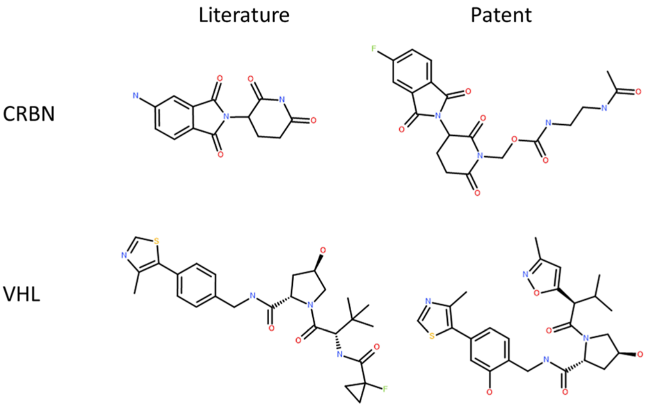


Figure 6. Four examples of E3 ligase binders with the relative source and E3 ligase specificity.

Applicability of the ErG-Based XGBoost Model. Assuming that our XGBoost can precisely predict the binding of an E3 ligase with a small molecule with the highest probability, we applied it to predict potential novel E3 ligands from commercial libraries. There is an enormous interest in filtering the most promising molecules from commercial databases.[31,32] On one side, E3 ligase pockets have been described in ELIOT,[33] a platform containing the E3 ligase pocketome to enable navigation and selection of new E3 ligases and new ligands for the design of new PROTACs, while on the other are large AI-based models like AlphaFold database,[34] so far untapped, for E3 ligase cavity detection.

To demonstrate the applicability of the model to known degrader libraries, we used the Asinex molecular degrader collection which contained about 1257 compounds.[35] Upon running our ErG-based XGBoost model on these compounds, it was revealed that this commercial collection was heavily skewed toward probable CRBN binders (66%) and with only 2% possibly selecting VHL. Surprisingly, a good 32% is addressing the mixed "other" E3 ligase class. To assess the applicability of the prediction, we had to compare the training set chemical space with the predicted library space. In case the library molecules would have not been similar (Tanimoto similarity accepted >0.5) to those in training, the relative predictions should be taken with very less confidence. We found eight exactly identical compounds in the commercial dataset while other 74 compounds showed a Tanimoto similarity higher than 0.5. A complete experimental validation would require specific biophysical binding assays with the E3 ligase predicted. Important for us was to show that commercial libraries have the potential to deliver novel candidates but that, even with lower level of applicability, these potentialities are confined if the selection is oriented toward specific E3 ligases.

For what concerns the possibility of enriching publicly known datasets with E3 ligase binders, i.e., the possibility to find specific E3 ligase binders within chemical biology collections or even within repurposing libraries, we repeated the prediction experiment using one of the major sources of repurposing compounds, the compounds from the Broad Institute's Drug Repurposing Hub.[36] Assuming again, of course wrongly, that all compounds could be E3 ligase binders, we wanted to check which ligases could be eventually predicted as more probable for those compounds and found that 24% of the molecules collected there could be indeed a CRBN binder. Here as well, we checked how far these molecules lied from the model training set and found a comfortable set of about 650 compounds with a Tanimoto similarity higher than 0.5. For this promising reason, we explored experimentally this collection to find possible degraders (J. Reinshagen et al., manuscript in preparation).

■ DISCUSSION

As hydroxy proline is a key residue for interactions with VHL protein, and as succinimide ring plays a key interaction role with CRBN protein cavity, we have demonstrated that the ErG bits are well designed to drive selectivity of E3 ligase binders by showing that the most relevant bits for the model are indeed essential in known ligase-ligand interactions. Not surprisingly, the ErG pharmacophore scheme resulted sufficiently general to be applied across different classes of ligands.

While it is true that the dataset used for training our machine learning models could be biased and not structurally homogeneous enough, we took several steps to address this potential issue. First, we carefully curated the dataset to include only high-quality experimental data with well-established and accepted literature sources. Second, we performed rigorous cross-validation to evaluate the generalizability of the model to unseen data. Third, we used feature selection techniques to identify the most informative features that contribute to binders' probability and selectivity toward specific E3 ligases. Finally, we validated the model on an independent test set and observed convincing performances, indicating that our model was not simply memorizing the training data. We are aware of the dynamic nature of the field: each novel ligase ligand should be added to the training set to improve generality of the model, so we are constantly keeping track of changes to enrich the training set and to provide the community with novel tools.

We are well aware that the current modeling approach is limited only to known E3 ligase binders. However, the definition and the inclusion of the nonbinder dataset together with experimental validation of the multi-class predictions needs to be considered as a large extension of this study and will be published in due course. Moreover, as a future prospect, a classification of E3 ligases through their druggable cavities extracted, for instance, either from the cited ELIOT database or from the Alpha Fold collection of ligase 3D models that will also be considered as a natural playground to apply our predictions.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acsomega.3c02803.

Example of pharmacophore information extraction using Erg; atoms and bonds and their bits used in pharmacophore extraction in ECFP4 and RDKit fingerprint schemas, respectively; two examples of compounds with the composite structure and their E3 ligase specificity assignments; and Confusion Matrix from the XGBoost model for the ligases used in this study (PDF)

■ AUTHOR INFORMATION

Corresponding Author

Reagon Karki − *Fraunhofer Institute for Translational Medicine and Pharmacology (ITMP), 22525 Hamburg, Germany; Fraunhofer Cluster of Excellence for Immune-Mediated Diseases (CIMD), 60590 Frankfurt, Germany;* orcid.org/0000-0002-1815-0037; Email: Reagon.Karki@itmp.fraunhofer.de

Authors

Yojana Gadiya − *Fraunhofer Institute for Translational Medicine and Pharmacology (ITMP), 22525 Hamburg, Germany; Fraunhofer Cluster of Excellence for Immune-Mediated Diseases (CIMD), 60590 Frankfurt, Germany; Bonn-Aachen International Center for Information Technology (B-IT), University of Bonn, 53113 Bonn, Germany;* orcid.org/0000-0002-7683-0452

Philip Gribbon − *Fraunhofer Institute for Translational Medicine and Pharmacology (ITMP), 22525 Hamburg, Germany; Fraunhofer Cluster of Excellence for Immune-Mediated Diseases (CIMD), 60590 Frankfurt, Germany*

Andrea Zaliani − *Fraunhofer Institute for Translational Medicine and Pharmacology (ITMP), 22525 Hamburg, Germany; Fraunhofer Cluster of Excellence for Immune-Mediated Diseases (CIMD), 60590 Frankfurt, Germany*

Complete contact information is available at: https://pubs.acs.org/10.1021/acsomega.3c02803

Author Contributions

A.Z. and R.K. conceived the work. R.K. programmed mapping of the E3 binders in accordance with the three database sources cited. Y.G., A.Z., and P.G. performed the analysis and contributed to ideation. A.Z., Y.G., and R.K. have written the manuscript. All the authors have reviewed, read, and approved the final manuscript.

Funding

This work was supported by the German Federal Ministry of Education and Research (BMBF) Project: **03ZU1109KB** 'PROXIDRUGS: Datenmanagement, Transfer und Innovation (INNODATA)'.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

We acknowledge PROTAC-DB and PROTACpedia curators for the access to PROTACs data and Dr. Aniket Ausekar from Evolvus inc. for providing license to access PDD data.

■ REFERENCES

(1) Rui, H.; et al. Protein-protein interfaces in molecular glue-induced ternary complexes: classification, characterization, and prediction. *RSC Chem. Biol.* **2023**, *4*, 192−215.

(2) Hu, G.; et al. Performance evaluation of 2D fingerprint and 3D shape similarity methods in virtual screening. *J. Chem. Inf. Model.* **2012**, *52*, 1103−1113.

(3) Zhao, R.; et al. Protein pocket detection via convex hull surface evolution and associated Reeb graph. *Bioinformatics* **2018**, *34*, i830−i837.

(4) Cang, Z.; Wei, G.-W. Integration of element specific persistent homology and machine learning for protein-ligand binding affinity prediction. *Int. J. Numer. Methods Biomed. Eng.* **2018**, *34*, No. e2914.

(5) Chen, D.; Gao, K.; Nguyen, D. D.; et al. Algebraic graph-assisted bidirectional transformers for molecular property prediction. *Nat. Commun.* **2021**, *12*, 3521.

(6) Feng, H.; et al. Machine-learning repurposing of DrugBank compounds for opioid use disorder. *Comput. Biol. Med.* **2023**, *160*, No. 106921.

(7) Abinaya, R. V.; Viswanathan, P. Biotechnology-based therapeutics. *Translational Biotechnology*; Academic Press, 2021; pp 27−52.

(8) Lu, X.; et al. The development of pharmacophore modeling: Generation and recent applications in drug discovery. *Curr. Pharm. Des.* **2018**, *24*, 3424−3439.

(9) Luo, M.; Li, Z.; Li, S.; Lee, T.-Y. A representation and deep learning model for annotating ubiquitylation sentences stating E3 ligase-substrate interaction. *BMC Bioinf.* **2021**, *22*, 1−18.

(10) Chana, C. K.; et al. Discovery and structural characterization of small molecule binders of the human CTLH E3 ligase subunit GID4. *J. Med. Chem.* **2022**, *65*, 12725−12746.

(11) Lee, J.; et al. Discovery of E3 ligase ligands for target protein degradation. *Molecules* **2022**, *27*, 6515.

(12) Stiefl, N.; Watson, I. A.; Baumann, K.; Zaliani, A. ErG: 2D pharmacophore descriptions for scaffold hopping. *J. Chem. Inf. Model.* **2006**, *46*, 208−220.

(13) Stiefl, N.; Zaliani, A. A knowledge-based weighting approach to ligand-based virtual screening. *J. Chem. Inf. Model.* **2006**, *46*, 587−596.

(14) Weng, G.; et al. PROTAC-DB 2.0: an updated database of PROTACs. *Nucleic Acids Res.* **2023**, *51*, D1367−D1372.

(15) PROTACpedia. Last accessed: February 04, 2023. https://protacdb.weizmann.ac.il/ptcb/main.

(16) Proximity Degraders Database. Last accessed: April 04, 2023. https://www.evolvus.com/PD.html.

(17) Li, A. S.; et al. Discovery of Nanomolar DCAF1 Small Molecule Ligands. *J. Med. Chem.* **2023**, *66*, 5041−5060.

(18) Chemical Computing Group. Last accessed: April 08, 2023. https://www.chemcomp.com, 2023.

(19) Durant, J. L.; Leland, B. A.; Henry, D. R.; Nourse, J. G. Reoptimization of MDL keys for use in drug discovery. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 1273−1280.

(20) *DAYLIGHT*. Last accessed: April 08, 2023. https://www.daylight.com/dayhtml/doc/theory.

(21) Gedeck, P.; Rohde, B.; Bartels, C. QSAR- how good is it in practice? Comparison of descriptor sets on an unbiased cross section of corporate data sets. *J. Chem. Inf. Model.* **2006**, *46*, 1924−1936.

(22) Rogers, D.; Hahn, M. Extended-connectivity fingerprints. *J. Chem. Inf. Model.* **2010**, *50*, 742−754.

(23) van der Maaten, L.; Hinton, G. Visualizing data using t-SNE. *J. Mach. Learn. Res.* **2008**, *9*, 2579−2605.

(24) Jonathon, P. P.; Hahn, C. A.; Fontana, P. C.; Broniatowski, D. A.; Przybocki, M. A. *Four principles of explainable artificial intelligence*, 2020, vol *18*.

(25) Friedman, J.; Hastie, T.; Tibshirani, R. Additive logistic regression: a statistical view of boosting (with discussion and a rejoinder by the authors). *Ann. Stat.* **2000**, *28*, 337−407.

(26) kaggle: low variance features. Last accessed: April 08, 2023. https://www.kaggle.com/code/fchmiel/low-variance-features-useless.

(27) Banerjee, M.; Capozzoli, M.; McSweeney, L.; Sinha, D. Beyond kappa: A review of interrater agreement measures. *Can. J. Stat.* **1999**, *27*, 3−23.

(28) RDKit: Open-Source Cheminformatics Software. Last accessed: 25.02.2023. https://www.rdkit.org.

(29) Berthold, M. R.; et al. KNIME-the Konstanz information miner: version 2.0 and beyond. *ACM SIGKDD Explor. Newsletter* **2009**, *11*, 26−31.

(30) R Core Team. *R: A language and environment for statistical computing*, 2013; pp 275−286.

(31) Palomba, T.; et al. Exploiting ELIOT for scaffold-repurposing opportunities: TRIM33 a possible novel E3 ligase to expand the toolbox for PROTAC design. *Int. J. Mol. Sci.* **2022**, *23*, 14218.

(32) Ishida, T.; Ciulli, A. E3 ligase ligands for PROTACs: how they were found and how to discover new ones. *SLAS Discov.* **2021**, *26*, 484−502.

(33) Palomba, T.; Baroni, M.; Cross, S.; Cruciani, G.; Siragusa, L. ELIOT: A platform to navigate the E3 pocketome and aid the design of new PROTACs. *Chem. Biol. Drug Des.* **2023**, *101*, 69−86.

(34) Varadi, M.; 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−D444.

(35) Asinex. Last accessed: April 04, 2023. https://www.asinex.com/protein-degradation.

(36) Drug Repurposing Hub (version: 9/7/2018). Last accessed: April 10, 2023. https://clue.io/repurposing.