Pharmacophore-based ML model to predict ligand selectivity for E3 ligase binders

Reagon Karki[1,2], Yojana Gadiya[1,2,3], Phil Gribbon[1,2], Andrea Zaliani[1,2*]

Affiliations

[1] Fraunhofer Institute for Translational Medicine and Pharmacology (ITMP), Schnackenburgallee 114, 22525 Hamburg, Germany

[2] Fraunhofer Cluster of Excellence for Immune-Mediated Diseases (CIMD), Theodor Stern Kai 7, 60590 Frankfurt, Germany

[3] Bonn-Aachen International Center for Information Technology (B-IT), University of Bonn, 53113 Bonn, Germany

* **Corresponding author**: Andrea.Zaliani@itmp.fraunhofer.de

Keywords

E3 Ligase ligand, Machine Learning, Virtual Screening, Extended Reduced Graph (ErG) pharmacophore

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 inexpensive approach using a pharmacophore fingerprinting scheme known as ErG, which is used in a multiclass 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 for rational design of E3 ligase binders.

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 (Rui et al., 2023). Thus, predicting the target binding selectivity for E3 ligases using pharmacophore 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.

Pharmacophore analysis involves identification of key chemical features, or pharmacophores, that are necessary for a ligand to interact with a particular protein target cavity (Abinaya & Viswanathan, 2021). This can be done by analyzing the structure of known X-ray complexes and identifying common chemical features that are critical for binding (Lu et al., 2018; Luo et al., 2021). In the case of E3 ligases, there are several key structural features that are important for ligand binding, including the presence of a zinc-binding domain and a substrate-binding site (Chana et al., 2022; Lee et al., 2022). 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 bindings that are not captured by the pharmacophore model.

In this manuscript, we present a very 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) (Stiefl et al., 2006; Stiefl & Zaliani, 2006). Each ErG bit forms the basis for a multiclass classification model where singular E3 ligase target names are used as labels. This is the first example of such a classification approach in the E3 ligase field. The resultant statistical model has an accuracy of 93.8% and 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 predict the probability of each molecule 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 ligand complexes, which would serve as the training set for the machine learning model. This dataset was created by merging three PROTAC resources namely PROTAC-DB 2.0 (Weng et al., 2023), PROTACpedia (http://protacdb.weizmann.ac.il/ptcb/main) and a commercial subset of Evolvus's Liceptor database (http://www.evolvus.com/Data.html), where the E3 ligase binding components of original active PROTACs are structurally identified and assigned. This yielded a total of 643 unique ligands. Additionally, we expanded the chemical space of E3 binders with 19 ligands

specific to DDB1 and CUL4 associated factor 1 (DCAF1) which were not present in PROTACs-derived collections (Li et al., 2023).

Source	E3 Ligase ligands
PROTAC-DB + PROTACpedia + Evolvus	**2291 (643 unique)**
DCAF1 binders (Shi Ming Li et al.)	**19**
Total unique E3 Ligase binder	**662**

Table 1: E3 ligase ligands list from literature and patent applications collected from two public databases (PROTAC-DB and PROTACpedia) and one commercial subset of Evolvus Liceptor database

The summary of the dataset of unique 662 compounds alongside the seventeen E3 ligases targets is shown in **Figure 1**. Since certain target classes (i.e., DCAF1, DCAF15, DCAF11, and DCAF16, MDM2 proto-oncogene (MDM2), aryl hydrocarbon receptor (AHR), baculoviral IAP repeat containing 3 (cIAP2/BIRC3), ring finger proteins 4 (RNF4) and 114 (RNF114), fem-1 homolog B (FEM1B), ubiquitin-protein ligase E3 component n-recognin 1 (UBR1), and Cullin 4A (CUL4A)) 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.

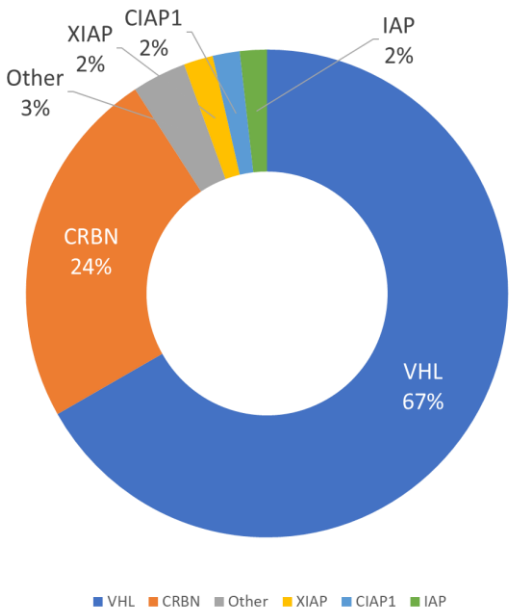

Figure 1: Representation of the E3 ligases and relative percentage of compound ligand collected. Von Hippel-Lindau tumor suppressor (VHL) and cereblon (CRBN) are the most studied E3 ligase targets with 442 and 159 ligands respectively in the collected dataset. Moreover, X-linked inhibitor of apoptosis (XIAP), baculoviral IAP repeat

containing 2 (cIAP1/BRIC2), 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 help of the ErG pharmacophoric fingerprint (Stiefl et al., 2006) as implemented within MOE 2022.02 (https://www.chemcomp.com) through the relative publicly available script (https://svl.chemcomp.com/data/Extended_Reduced_Graph__ErG__fingerprint.svlx).

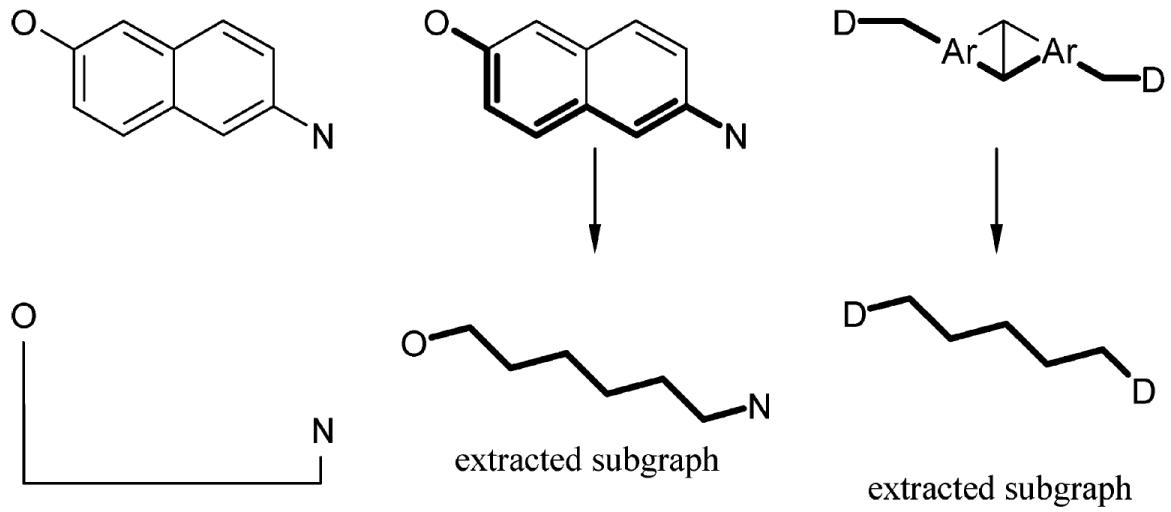

Figure 2: Pharmacophore information extraction example**.** Distances differ from 2D representation (left-hand side) to classical subgraphs representation (middle), to ErG description scheme (right-hand side) through reduced graph representation (reprinted with permission from (Stiefl et al., 2006)

The extracted pharmacophores were clustered based on similarity using methods such as hierarchical clustering or k-means clustering (see **Figure 1S** in supplementary material). Once the pharmacophores were, they were used to generate a machine learning model using techniques such as random forest, XGBoost, or simple partition tree. The model would take the chemical structure of a new ligand as input and predict the most probable E3 ligase target for the corresponding ligand based on the similarity of its pharmacophore to those in the training set.

It is important to note that the accuracy of the model depends on many factors coming from different data sources. Firstly, on the quality and size of the training set; secondly on the choice of descriptors and thirdly 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 it to predictions.

We developed several R and Python scripts to perform exploratory analyses, to compare the generated ML classifications and to analyze each model in depth. The algorithms used in the classifications were simple Partition Tree (PT), Random Forest (RF), and XGBoost (XGB) algorithms. The ErG columns (descriptors/bits) showing variance lower than 0.2 were removed to

generate a matrix of 662 rows X 123 columns, where rows are ligands collected and the columns are ErG bits remaining after variance filtration. We assumed that lower or constant variance columns should not contribute to final models (https://www.kaggle.com/code/fchmiel/low-variance-features-useless).
We generated several models with different classification algorithms and automatically selected the best algorithm on the basis of highest Cohen's kappa value (https://en.wikipedia.org/wiki/Cohen%27s_kappa) defined in here below:

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

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.

The models were trained on 80% of the original matrix leaving a test set of 20% for internal validation. The test set was generated by stratification on E3 ligase labels, i.e. each of the class is represented in both training and test sets. All the models have been generated after a 10X cross validation scheme. Wherever necessary, a selection of the top ten most influential ErG bits were generated to try to rationalize the model building.

Results

Model analysis
The dataset used to attempt the multi-class classification model is the largest so far reported. We found that all algorithms used for multi-class classification performed well as reported in Table 2.

Model	Accuracy	Cohen's Kappa
Simple Tree	0.939	0.879
Random Forest	0.929	0.856
XGBoost	0.938	0.872

Table 2: Summary table of model performances

The most performant models came from Simple Tree and XGBoost algorithm but we chose to focus on XGBoost only being more robust. We also extracted the most relevant ten ErG bits as reported in Figure 2. Only two out of those relate to distances more than 5, 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_D_d2 is surely present, among others, in hydroxy-proline moiety, while Hf_Ac_d2 is present, among others, in succinimide-like rings, but not in maleimide analogues. Interestingly, there is only one relevant ErG bit dealing with aromatic (Ar) groups, even if almost all the E3 ligands so far collected have at least one Aromatic group in them.

Beside 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 (see Figure 1).

Indeed, using just these two ErG bits 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 "Other" class of E3 ligases. In the confusion matrix, we report in Figure 2S in supplementary materials, it can be appreciated for the accuracy of the six classes of E3 ligases used.

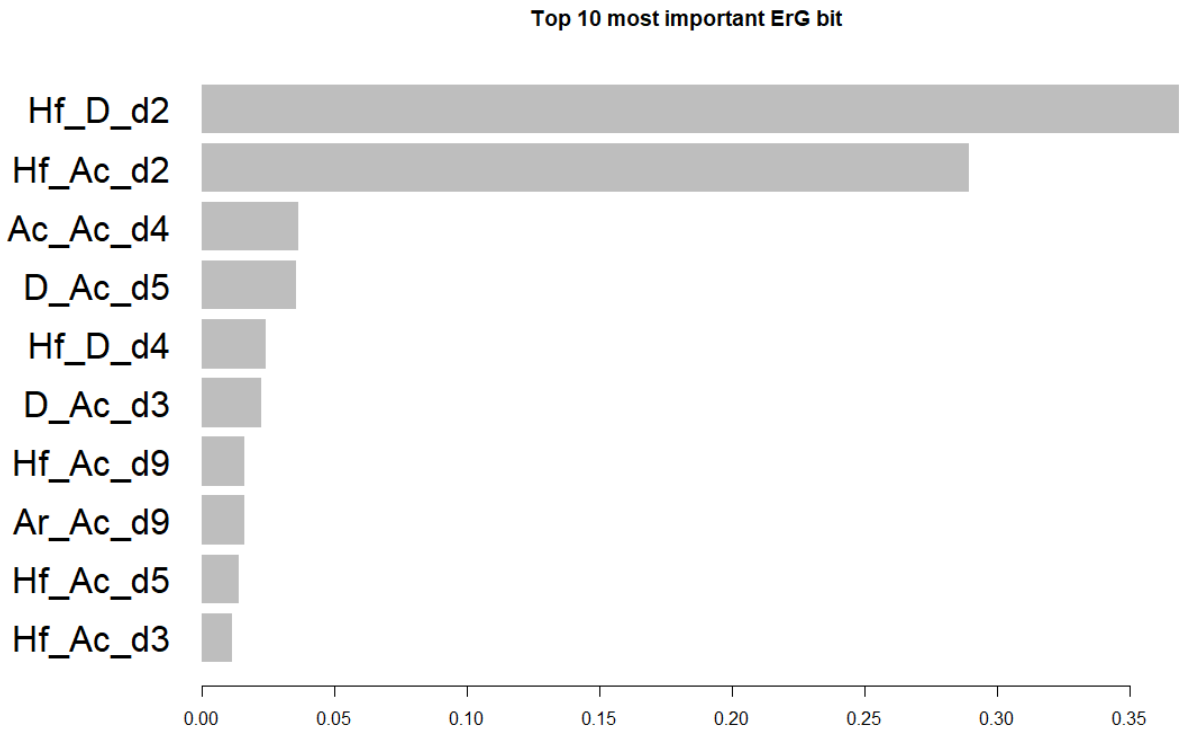

Figure 3: Top ten descriptors (ErG bits) contributing to the XGBoost model for E3 ligase selectivity predictions

Diving deeper in the ErG bits space, we tried to evaluate statistical relevance of what has been found as top ten relevant features and how their differences are distributed along the six E3 ligase classes. Indeed, not all the distributions found around these ErG bits are statistically significant but there are some which are pretty informative (Figure 3). While Hf_D_d2 is clearly a footprint for VHL only (probably due to the already mentioned hydroxy proline residue), 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. ErG bit D_Ac_d5 seems to be another marker for VHL (highest count) especially against the "Other" group of Ligases which do not show any count on this feature distance. Hf_D_d4 and D_Ac_d3 seem to mark CIAP1, IAP and XIAP ligands as they are contained in hydrophobic aliphatic amino acids and amino acids, respectively. Both are well represented in these three groups. CIAP1 and XIAP only have the largest count of Ar_Ac_d9, while Hf_Ac_d9 has the least significant contribution according to its distribution in the six groups (Figure 3).

Application

Assuming that our XGBoost can precisely predict binding of a E3 ligase with a small molecule with highest probability, we applied it to predict some libraries which might be useful to be tested as a source of E3 ligands or as a possible source of molecular glues. There is an enormous interest to filter the most promising molecules from commercial databases (Ishida & Ciulli, 2021; Palomba et al., 2022). On the other side, E3 ligase pockets have been described in ELIOT (https://eliot.moldiscovery.com), 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 (Palomba et al., 2023). AlphaFold database (https://alphafold.ebi.ac.uk) is naturally another source, so far untapped, for E3 ligase cavities. Moreover, the opportunities to synthesize novel E3 ligase ligands have also been reported (Bricelj et al., 2021).


ErG bits distribution of most influential ErG bits for Ligase classification
Hf_D_d2
Hf_Ac_d2
Ac_Ac_d4
D_Ac_d5
Hf_D_d4
D_Ac_d3
Hf_Ac_d9
Ar_Ac_d9
value
Class
CIAP1
CRBN
IAP
Other
VHL
XIAP

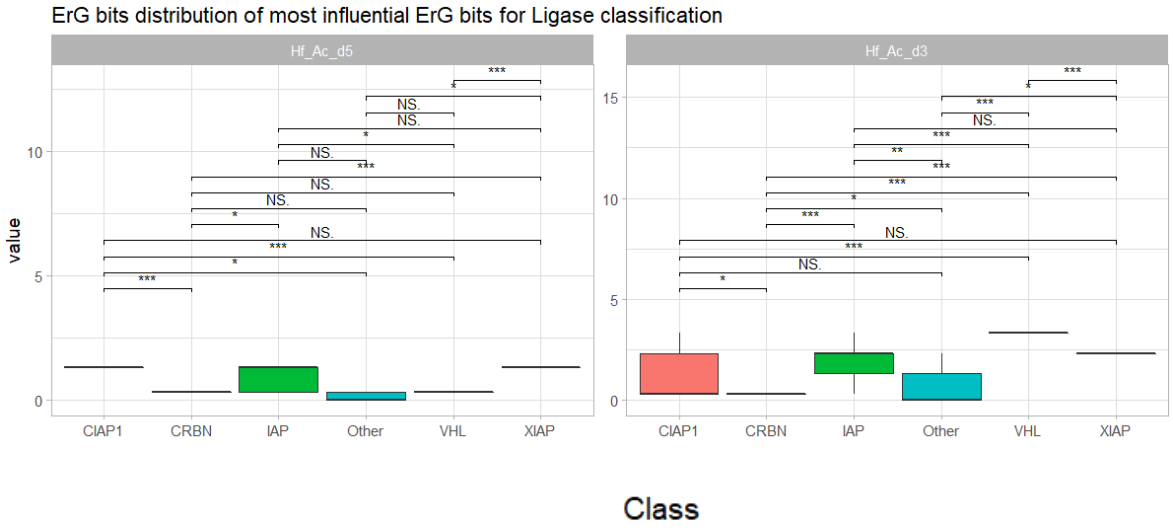

Figure 3: Box plot distribution of the top ten most influential ErG bit values according to the XGBoost model presented. While certain comparisons (t.test) are not significant (NS), some are according to calculate p-value labels (*** p-value < 0.001, ** p-value < 0.01, * p-value < 0.1)

Thus, using the 1257 compounds from Asinex molecular degrader collection (https://www.asinex.com/protein-degradation), and assuming that all the compounds would be binding an E3 ligase, as actual stand of molecular glue mediated degradation suggests, we ended up revealing that the commercial collection is heavily skewed towards probable CRBN binders (35%) and with only 3% possibly selecting VHL. We repeated this experiment using one of the major sources of repurposing compounds publicly available, i.e., the compounds from the Broad library collection (https://clue.io/repurposing, version: 9/7/2018). 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 18% of the molecules collected there could be indeed a CRBN binder. We are experimentally exploring this collection in order to find degrader (J. Reinshagen et al., manuscript in preparation). This is just an application of the many possibilities: so, for example, any collections or proprietary compounds could be filtered taking advantage of the simplicity of the ErG scheme.

In our dataset, by keeping track of whether the E3 ligase ligands have scientific literature or patent provenance, we have tried to analyze possible differences under this aspect (Figure 3S). As expected in patent literature, the chemical diversity and/or complexity is higher. For all non-significant pharmacophoric features, comparisons of the boxes indicating 95% percentiles are clearly close to one another, while for seven out of ten there are significant differences, suggesting that the latent patent pharmacophoric space is in some aspects different from published one. Evolvus's liceptor 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 4.

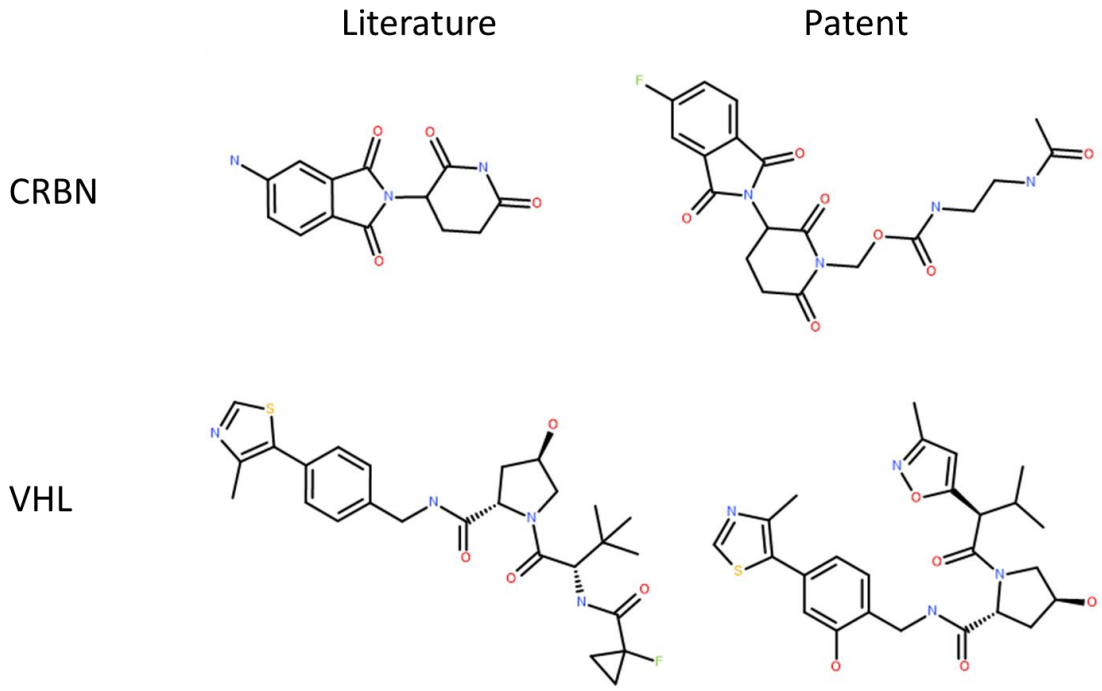

Figure 4: Four examples of E3 Ligase binders with relative source and E3 ligase specificity

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.

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. Secondly, we performed rigorous cross-validation to evaluate the generalizability of the model to unseen data. Thirdly, we used feature selection techniques to identify the most informative features that contribute to binders' probability and selectivity towards 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. As a future prospect, a classification of E3 ligases through their druggable cavities extracted, for instance,

either from cited ELIOT database or from an Alpha Fold collection of Ligase 3D models could represent the natural playground to apply our predictions.

Data availability

All input data and workflow have been made available either on github (https://github.com/Fraunhofer-ITMP/E3_binder_Model/tree/main/Data). The structure information from Liceptor subset from Evolvus has been omitted due to license restrictions.

Author contribution

AZ and RK conceived the work. RK programmed mapping of the E3 binders in accordance with the three database sources cited. JR, YG, AZ and PG performed the analysis and contributed to ideation. AZ and RK have written the manuscript. All the authors contributed towards reviewing the manuscript. PG reviewed the manuscript and all authors have read and approved the final manuscript.

Competing Interests

None declared

Funding

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

References

Abinaya, R. V., & Viswanathan, P. (2021). Biotechnology-based therapeutics. In *Translational Biotechnology* (pp. 27–52). Elsevier.

Bricelj, A., Steinebach, C., Kuchta, R., Gütschow, M., & Sosič, I. (2021). E3 ligase ligands in successful PROTACs: an overview of syntheses and linker attachment points. *Frontiers in Chemistry*, *9*, 707317.

Chana, C. K., Maisonneuve, P., Posternak, G., Grinberg, N. G. A., Poirson, J., Ona, S. M., Ceccarelli, D. F., Mader, P., St-Cyr, D. J., Pau, V., & others. (2022). Discovery and structural characterization of small molecule binders of the human CTLH E3 ligase subunit GID4. *Journal of Medicinal Chemistry*, *65*(19), 12725–12746.

Ishida, T., & Ciulli, A. (2021). E3 ligase ligands for PROTACs: how they were found and how to discover new ones. *SLAS DISCOVERY: Advancing the Science of Drug Discovery*, *26*(4), 484–502.

Lee, J., Lee, Y., Jung, Y. M., Park, J. H., Yoo, H. S., & Park, J. (2022). Discovery of E3 ligase ligands for target protein degradation. *Molecules*, *27*(19), 6515.

Li, A. S. M., Kimani, S., Wilson, B., Noureldin, M., González-Álvarez, H., Mamai, A., Hoffer, L., Guilinger, J. P., Zhang, Y., von Rechenberg, M., & others. (2023). Discovery of Nanomolar DCAF1 Small Molecule Ligands. *Journal of Medicinal Chemistry*.

Lu, X., Yang, H., Chen, Y., Li, Q., He, S., Jiang, X., Feng, F., Qu, W., & Sun, H. (2018). The development of pharmacophore modeling: Generation and recent applications in drug discovery. *Current Pharmaceutical Design*, *24*(29), 3424–3439.

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

Palomba, T., Baroni, M., Cross, S., Cruciani, G., & Siragusa, L. (2023). ELIOT: A platform to navigate the E3 pocketome and aid the design of new PROTACs. *Chemical Biology & Drug Design*, *101*(1), 69–86.

Palomba, T., Tassone, G., Vacca, C., Bartalucci, M., Valeri, A., Pozzi, C., Cross, S., Siragusa, L., & Desantis, J. (2022). Exploiting ELIOT for Scaffold-Repurposing Opportunities: TRIM33 a Possible Novel E3 Ligase to Expand the Toolbox for PROTAC Design. *International Journal of Molecular Sciences*, *23*(22), 14218.

Rui, H., Ashton, K. S., Min, J., Wang, C., & Potts, P. R. (2023). Protein-protein interfaces in molecular glue-induced ternary complexes: classification, characterization, and prediction. *RSC Chemical Biology*.

Stiefl, N., Watson, I. A., Baumann, K., & Zaliani, A. (2006). ErG: 2D pharmacophore descriptions for scaffold hopping. *Journal of Chemical Information and Modeling*, *46*(1), 208–220.

Stiefl, N., & Zaliani, A. (2006). A knowledge-based weighting approach to ligand-based virtual screening. *Journal of Chemical Information and Modeling*, *46*(2), 587–596.

Weng, G., Cai, X., Cao, D., Du, H., Shen, C., Deng, Y., He, Q., Yang, B., Li, D., & Hou, T. (2023). PROTAC-DB 2.0: an updated database of PROTACs. *Nucleic Acids Research*, *51*(D1), D1367–D1372.