Proteolysis-Targeting Chimera (PROTAC) Design using Integrated Artificial Intelligence (AI) and Quantitative System Pharmacology (QSP) Model

Jina Kim [1], Sungwoo Goo [1], Sangkeun Jung [1,2,*], Jung-woo CHAE[1,3,*], Jae-mun Choi[4,*], Hwi-yeol Yun[1,3,*]

[1]Department of Bio-AI convergence, Chungnam National University, Daejeon, South Korea
[2]College of Pharmacy, Chungnam National University, Daejeon, South Korea
[3]Department of Computer Science and Engineering, Chungnam National University, Daejeon, South Korea
[4]Calici Co., LTD, Daejeon, South Korea
*Those of authors contributed equally as correspondence

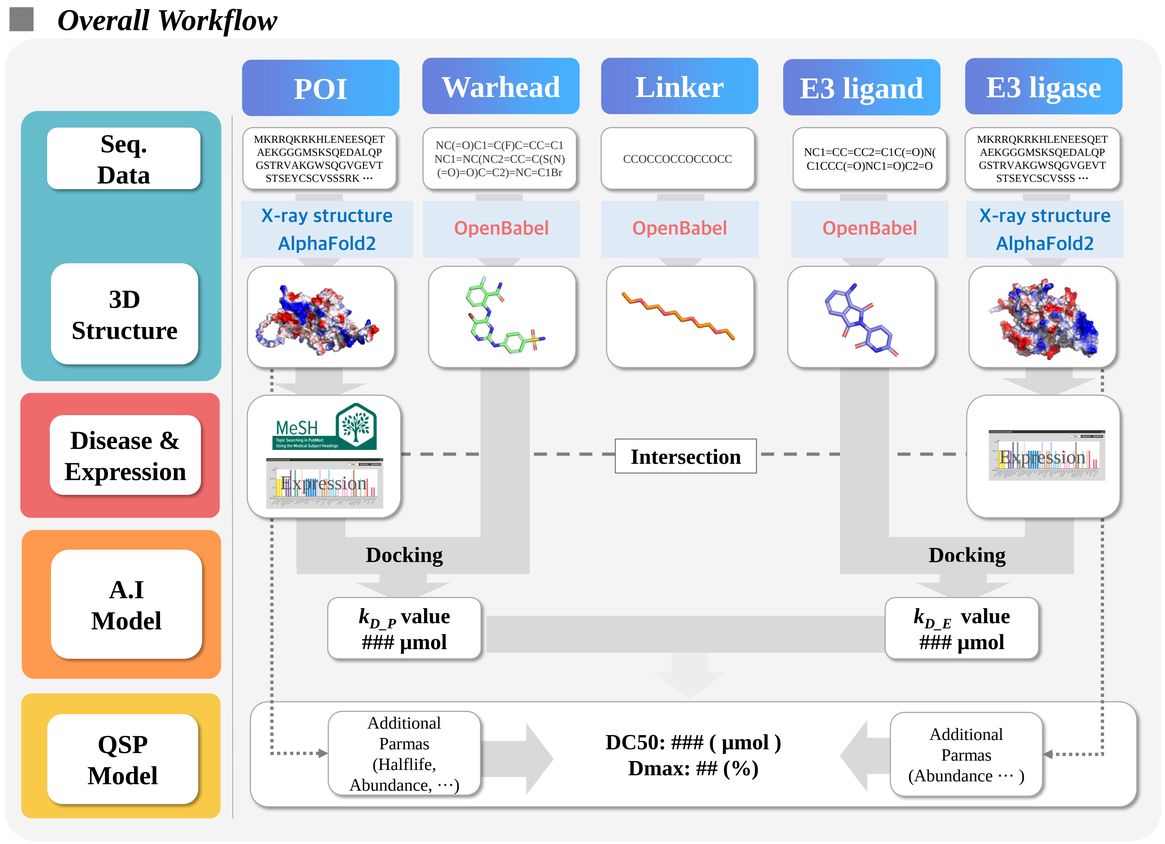
ABSTRACT

- In response to the significant **time and cost constraints associated with testing numerous combinations** of protein-targeted chimera (PROTAC) constituents, an imperative arises to develop an efficient screening method for identifying the optimal PROTAC combination.
- To address this need, **we present a novel approach integrating database, disease and expression profiles, artificial intelligence (AI), and quantitative systems pharmacology (QSP) Design Model**.
- Additionally, we developed a user-friendly **PROTAC Dashboard**, empowering researchers to autonomously adjust the DC50 and Dmax values by combining PROTAC components, thereby **facilitating the efficient identification of promising PROTAC therapeutics.**

※ *DC50: PROTAC concentration that produces half-maximal degradation, ※ Dmax: maximal extent of degradation*

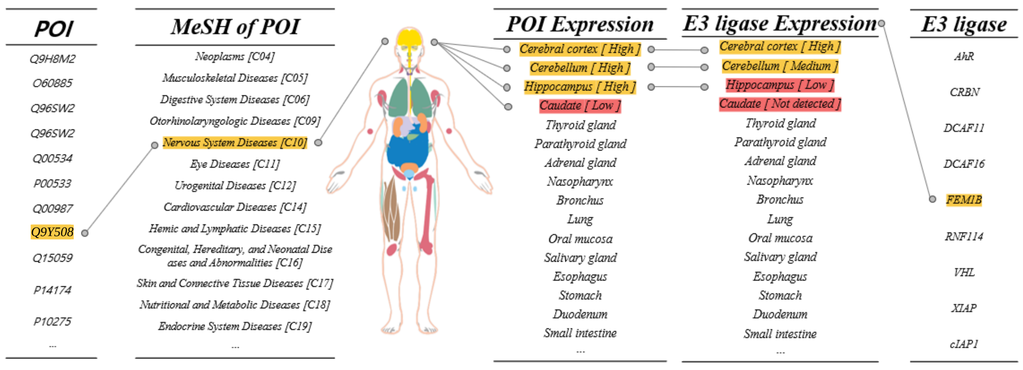
METHODS

Overall Workflow

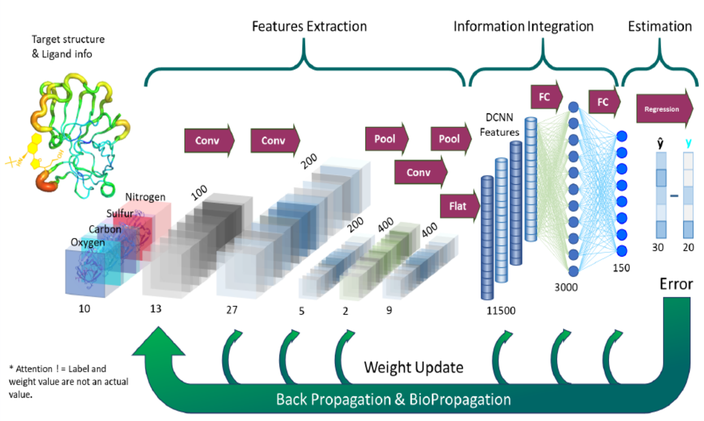


Disease & Expression Filter

- This research focuses on elucidating the disease associated with the target protein, POI, and assessing the expression levels of the E3 ligase in human tissues.
- While the disease association of E3 ligases is of secondary importance, we prioritized assessing the expression levels of E3 ligases in relevant human tissues because of their primary role in ubiquitination.
- To accomplish this, we explored the MeSH terms from Uniprot and NIH databases to identify the top diseases associated with POIs. Additionally, we referred to the E3 ligase expression data obtained from the Human Protein Atlas.



AI Model: 4D Tensor CNN Architecture



- We developed a deep learning model, 4D-CNN architecture, to investigate the binding kinetics between a target protein and a ligand.
- The 3D structures of the known protein-ligand complex with measured binding kinetics were utilized for training the model, consisting of convolutional and dense parts with different types of connections between layers.
- The data input as a protein-ligand complex is represented as a 4D tensor (x,y,z coordinates and a vector of 19 features).

QSP Filter:

Dmax

$$\frac{a \times E_0 \times k_{cat} \times t_{1/2,P} \times 1/ln(2)}{a \times E_0 \times k_{cat} \times t_{1/2,P} \times 1/ln(2) + a \times E_0 + k_{D\_E} + k_{D\_P} + 2 \times \sqrt{k_{\mathrm{D\_E}} \times k_{D\_P}}}$$

DC50

$$\frac{1}{2} \times \left(a \times E_0 \times k_{cat} \times t_{1/2,P} \times 1/ln(2) + a \times E_0 + k_{D\_E} + k_{D\_P} + 4 \times \sqrt{k_{\mathrm{D\_E}} \times k_{D\_P}} - \sqrt{\left(a \times E_0 \times k_{cat} \times t_{1/2,P} \times 1/ln(2) + a \times E_0 + k_{D\_E} + k_{D\_P} + 4 \times \sqrt{k_{\mathrm{D\_E}} \times k_{D\_P}}\right)^2 - 4 \times k_{\mathrm{D\_E}} \times k_{D\_P}} \right)$$

As input, you need biochemical data about the **binding affinity of PROTAC (k_{D_P}, k_{D_E}, a)**, information about the **cell system involved (E_0, $t_{1/2,P}$),** the **first-order rate constant at which the target protein is degraded ($k_{\mathrm{deg,p}}$),** and the **rate constant of ubiquitination (k_{cat}).** Protein information from databases such as **ProteomicsDB** and **Protein Abundance Database** were applied to the parameters marked with " ● "

kdeg,p/kcat = 1% and half_life = ln2 / kdeg,p as additional expressions.

※ *For simplicity of screening, data from HeLa cells were used and the cooperative factor, a, was fixed at 1.*

RESULTS

Data Filter

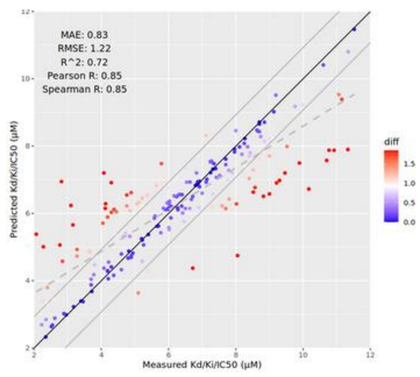
- We employed a systematic data filtering approach by importing a targeted selection of 280 proteins of interest (POIs), 13 E3 ligases, 365 warheads, and 82 E3 ligands from PROTAC-DB version 2.0.
- By carefully excluding proteins and ligands that did not meet the necessary criteria and accounting for errors in AlphaFold2 predictions, we narrowed down our dataset to 91 POIs and 9 E3 ligases for subsequent screening.
- Our final dataset consisted of 87966 combinations comprising 1086 POI-Warhead combinations and 81 E3 ligase-E3 ligand combinations, forming our primary filtered data for analysis and further investigation.

Disease & Expression Filter

- An analysis was done on 91 POIs from UniProt. Disease info was available for 42 POIs, null data for 49. Filtering for disease was applied only to POIs with disease info.
- Abundance data of 9 E3 ligases from Human Protein Atlas was examined, 7 had data. Filtering for abundance was done only for E3 ligases with data.

MeSH term	E3 Ligase to exclude
Musculoskeletal Diseases [C05]	P35869, Q8TEB1, Q9UK73, Q9Y508, P40337
Digestive System Diseases [C06]	P35869, Q8TEB1
Otorhinolaryngologic Diseases [C09]	Q9Y508, P40337
Nervous System Diseases [C10]	Q9Y508
Urogenital Diseases [C12]	Q9Y508, P98170
Cardiovascular Diseases [C14]	Q9Y508
Hemic and Lymphatic Diseases [C15]	P35869, Q9UK73, Q9Y508, P98170
Skin and Connective Tissue Diseases [C17]	Q9UK73, Q9Y508, P40337
Nutritional and Metabolic Diseases [C18]	P35869, Q9Y508
Endocrine System Diseases [C19]	P35869, Q9Y508
...	...

Artificial Inteligent Filter



 # AI Model's Kd value prediction accuracy

Using 500 protein-ligand test data, we compared the observed and predicted Kd values between them.

	R-square	Pearson cor	RMSE	MAE
Kd	0.72	0.85	1.22	0.83

※ RMSE: Root Mean Squared Error, ※ MAE: Mean Absolute Error

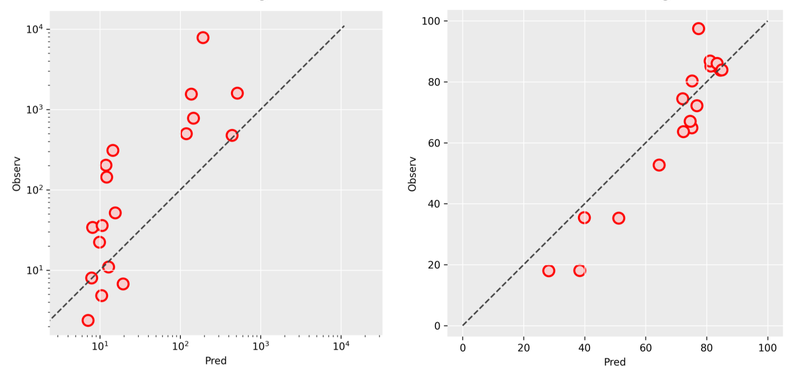
 # Kd value of PROTAC derived using the AI Model

1036 sets of POI and Warhead and 81 sets of E3 Ligase and E3 Ligand were docked, and the binding kinetics (k_{D_P} and k_{D_E}) were obtained.

QSP Filter

 # QSP Model's DC50, Dmax prediction accuracy

▼ DC50 Predict Regression ▼ Dmax Predict Regression



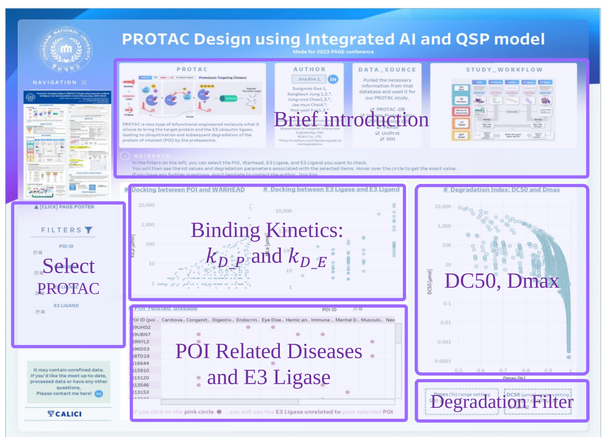
- The predicted and observed values of DC50 and Dmax of 18 PROTACs were compared.
- To evaluate the prediction ability of the QSP Model, all parameters of QSP were referenced to the experimental values in the reference paper.

※ The observed DC50 and Dmax were extracted using a digitizer.

	Pearson cor	RMSE	MAE
DC50	0.8049	1.7957	1.4791
Dmax	0.9512	9.6606	7.6331

A total of 37,321 degradation parameters were acquired for combinations of POI, Warhead, E3 Ligase, and E3 Ligand that successfully met all the top filters.

PROTAC Dashboard



 # Introduction for usage

- The user has the ability to choose the specific POI, Warhead, E3 Ligase, and E3 Ligand combinations they wish to examine.
- By doing so, they can view the associated kd values and degradation parameters. Additionally, the user has the option to set their own degradation value range for personalized filtering.
- By hovering over the circle, precise values can be obtained.

CONCLUSION & DISCUSSION

So far:

- Our PROTACs design is a leading attempt to integrate AI and QSP into the field of PROTACs. It should be helpful to make candidates for optimal PROTAC combination based on highly efficient screening results as well as suggestion the PROTAC research paradigm for combining AI and QSP.

From now on:

- Along with the refinement of the PROTAC design, further approaches will need to be taken to integrate experimental results to create sophisticated models.

Related material and program:

 # PROTAC Dashboard
Follow the QR code to access the PROTAC Dashboard. It is designed for desktop computers.

 # AI Model Platform: Pharmaco-Net
AI-assisted SB-RDD(Structure-Based Rational Drug Design/Discovery) Platform. There is no Licensing fee for subscription. Feel free to use the new token system.

Acknowledgement

- This research was funded by Chungnam National University and an Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No.RS-2022-00155857, Artificial Intelligence Convergence Innovation Human Resources Development (Chungnam National University)) and supported by the Research Foundation of Korea (NRF) grants from the Korean government (MSIT) (Nos. NRF-2022R1A2C1010929 and 2022R1F1A1071047) and Research Foundation of Korea (NRF) grant funded by the Korea government(MSIT) (No. 2022R1A5A7085156, Senior Health Convergence Research Center based on Life Cycle (Chungnam National University)).