

Original article:

PARP1PRED: A WEB SERVER FOR SCREENING THE BIOACTIVITY OF INHIBITORS AGAINST DNA REPAIR ENZYME PARP-1

Tassanee Lerksuthirat^{a*}, Sermsiri Chitphuk^a, Wasana Stitchantrakul^a,
Donniphat Dejsuphong^b, Aijaz Ahmad Malik^{c*}, Chanin Nantasenamat^{d*}

^a Research Center, Faculty of Medicine Ramathibodi Hospital, Mahidol University, Bangkok 10400, Thailand

^b Program in Translational Medicine, Chakri Naruebodindra Medical Institute, Faculty of Medicine Ramathibodi Hospital, Mahidol University, Samut Prakan 10540, Thailand

^c Center of Excellence in Computational Molecular Biology, Faculty of Medicine, Chulalongkorn University, Bangkok 10330, Thailand

^d Streamlit Open Source, Snowflake Inc., USA

* **Corresponding authors:** Tassanee Lerksuthirat, Research Center, Faculty of Medicine Ramathibodi Hospital, Mahidol University, Bangkok 10400, Thailand,

E-mail: tassanee.ler@mahidol.ac.th

Aijaz Ahmad Malik, Center of Excellence in Computational Molecular Biology, Faculty of Medicine, Chulalongkorn University, Bangkok 10330, Thailand,

E-mail: ajaz_me@hotmail.com

Chanin Nantasenamat, Streamlit Open Source, Snowflake Inc., USA,

E-mail: hellodataprofessor@gmail.com

<http://dx.doi.org/10.17179/excli2022-5602>

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/4.0/>).

ABSTRACT

Cancer is the leading cause of death worldwide, resulting in the mortality of more than 10 million people in 2020, according to Global Cancer Statistics 2020. A potential cancer therapy involves targeting the DNA repair process by inhibiting PARP-1. In this study, classification models were constructed using a non-redundant set of 2018 PARP-1 inhibitors. Briefly, compounds were described by 12 fingerprint types and built using the random forest algorithm concomitant with various sampling approaches. Results indicated that PubChem with an oversampling approach yielded the best performance, with a Matthews correlation coefficient > 0.7 while also affording interpretable molecular features. Moreover, feature importance, as determined from the Gini index, revealed that the aromatic/cyclic/heterocyclic moiety, nitrogen-containing fingerprints, and the ether/aldehyde/alcohol moiety were important for PARP-1 inhibition. Finally, our predictive model was deployed as a web application called PARP1pred and is publicly available at <https://parp1pred.streamlitapp.com>, allowing users to predict the biological activity of query compounds using their SMILES notation as the input. It is anticipated that the model described herein will aid in the discovery of effective PARP-1 inhibitors.

Keywords: PARP-1, DNA repair, machine learning, QSAR, webserver, cheminformatics

INTRODUCTION

Precision medicine is becoming increasingly important in treating many cancers because it can reduce side effects compared with

conventional therapies (Baudino, 2015). Several clinical trials have shown evidence of success, especially targeting DNA repair (Brown et al., 2017). For example, an ovarian

phase 2 clinical trial, in which platinum-sensitive patients were given the PARP-1 inhibitor olaparib as a maintenance treatment, showed an improvement in progression-free survival (Ledermann et al., 2012). In a phase 3 OlympiA clinical trial, in which olaparib was administered as an adjuvant to BRCA1/2-mutated breast cancer patients following completion of local treatment and neoadjuvant or adjuvant chemotherapy, the treatment group exhibited significantly longer survival, free of invasive or distant disease, than the placebo group (Tutt et al., 2021). Moreover, the phase 2 TOPARP-A trial showed that patients who had metastatic prostate cancer, who were no longer responding to standard treatments, and who had defects in DNA-repair genes, had a high response rate toward olaparib (Mateo et al., 2015).

DNA repair is a critical cellular process that ensures the integrity of the genome, allowing the parental cell to pass genetic information on to the progeny cell. Defective DNA repair causes accumulation of genetic mutations, thus leading to carcinogenesis. However, retaining some DNA repair activities is also important for cancer survival, especially when cells are under genotoxic stress (such as radio- and chemotherapy) (Helleday et al., 2008). DNA double-strand break (DSB) lesions are the most toxic form of DNA damage, which, if left unrepaired, result in cell death (Shibata and Jeggo, 2014). Therefore, drugs are of interest if their mode of action leads to the accumulation of DSBs (Srivastava and Raghavan, 2015).

Poly (ADP-ribose) polymerase (PARP) is an enzyme that catalyzes the ADP-ribosylation of a specific protein, resulting in the covalent binding of a single ADP-ribose unit or polymers of ADP-ribose units (Gupte et al., 2017). In humans, there are 17 members of the family, although only three (PARP-1, PARP-2, and PARP-3) are involved in DNA repair (Beck et al., 2014). Among the three, PARP-1 (EC 2.4.2.30) was identified in 1963 and is the most extensively investigated DNA repair enzyme (Gupte et al., 2017). By inhib-

iting PARP-1, DSB accumulation was induced in cancer cells deficient in *BRCA1/2*, indicating that PARP-1 is a druggable target (Mateo et al., 2019). Olaparib was the first well-known PARP-1 inhibitor, and it has been used as a targeted therapy to treat ovarian, breast, prostate, and pancreatic cancer patients with *BRCA1/2* mutations (de Bono et al., 2020; Fong et al., 2009; Golan et al., 2019; Kim et al., 2015). Recently, five more PARP-1 inhibitors, rucaparib (Balasubramaniam et al., 2017), niraparib (Mirza et al., 2016; Scott, 2017), talazoparib (Hoy, 2018), fluzoparib (Li et al., 2021), and pamiparib (Xu et al., 2021) have been approved by the Food and Drug Administration (FDA). However, access to targeted therapy has been restricted in certain countries, particularly middle- and low-income countries, because of a lack of affordability or the capability to develop domestic pharmaceutical technology, which poses a threat to health security (Fundytus et al., 2021; Ocran Mattila et al., 2021). As a result, accelerating drug discovery in such countries is an important factor to minimize such risk.

The computational-aided drug design (CADD) approach significantly reduces the time and cost associated with drug discovery (Nantasenamat and Prachayasittikul, 2015). With the availability of public bioactivity databases such as BindingDB (Gilson et al., 2016), PubChem (Kim et al., 2016), GtoPdb (Armstrong et al., 2020), and ChEMBL (Mendez et al., 2019), we can retrieve the bioactivity data and analyze the relationship between the chemical structures of compounds and their biological activities, termed the quantitative structure–activity relationship (QSAR) (Carracedo-Reboredo et al., 2021; Nantasenamat and Prachayasittikul, 2015). Developing a QSAR model involves two main steps: 1) molecular structure description; and 2) multivariate analysis to correlate molecular descriptors with observed biological activities (Nantasenamat et al., 2009). The first step is to define chemical structures as numerical representations of their physicochemical properties. The second step employs

statistical methods to establish the relationship between the independent variables (e.g., molecular descriptors) and the dependent variables (e.g., biological activities). As a result, the QSAR model is used to predict the effects of molecular descriptor changes on biological activities, as shown by the design of inhibitors against a variety of targets, such as antiviral (Malik et al., 2020; Worachartcheewan et al., 2014), anti-inflammatory (Kanan et al., 2021), and anticancer (Nantasenamat et al., 2014; Schaduengrat et al., 2021). We constructed predictive models for drug discovery using a biological dataset of PARP-1 inhibitors.

Many studies have investigated *in silico* screening of PARP-1 inhibitors, including QSAR, molecular modeling, molecular docking, molecular dynamics simulation (MD), and proteochemometric modeling (Abbasi-Radmoghaddam et al., 2021; Cortes-Ciriano et al., 2015; Halder et al., 2015; Li et al., 2016; Revathi et al., 2021). Halder and colleagues (2015) used comparative *in silico* studies, including 2D-QSAR, kernel-based partial least square (KPLS) analysis, pharmacophore search engine (PHASE) pharmacophore mapping, molecular docking, molecular mechanics with generalized Born and surface area solvation (MM-GBSA) analysis, and Gaussian-based 3D-QSAR analyses on docked poses to explore the structure–activity relationship of PARP-1 inhibitors (Halder et al., 2015). They used 254 compounds targeting PARP-1 from Merck Research Laboratories to conduct the analysis. They found that polar interactions play an important role to leverage the activity of PARP-1. Moreover, the positive ionizable feature of ligands also plays a key role to differentiate between highly active and inactive compounds. Revathi and colleagues (2021) used 71 compounds that were phthalazinone and 4-carboxamide benzimidazole derivatives to develop ligand-based pharmacophores (Revathi et al., 2021). They used Pharmacophore Alignment and Scoring Engine to identify the pharmacophore sites and later developed the ADHRR.1031 pharmacophore hypothesis as a 3D-QSAR model.

Furthermore, the model was validated using 1,000,000 ligands from various databases and analyzed through virtual screening. The docking analysis revealed the importance of hydrogen bonding between Gly863 and Ser904 of PARP-1 with ligands. Additionally, hydrogen bond formation with Ser864 and π - π interaction with His862, Arg878, and His909 were also observed in the docking analysis. Sahin and Durdagi (2021) aimed to identify novel piperazine-based PARP-1 inhibitors (Sahin and Durdagi, 2021). They used text mining to search for molecules containing piperazine as a main scaffold from the Specs-SC database. The sorted molecules were then analyzed by molecular docking, in which the ten highest docking scores were further subjected to molecular dynamics (MD) to calculate the free binding energy using the molecular mechanics/generalized born surface area method. They identified molecule-1388 as a potential candidate compound to selectively inhibit PARP-1. This compound had crucial hydrogen bonds with Gln759 and Met890 and π - π interaction with Tyr889. Abbasi-Radmoghaddam and colleagues (2021) conducted a QSAR and molecular modeling study that predicted the IC_{50} values (the concentration of inhibitor at which the enzymatic activity is reduced by half) of 1H-benzo[d]imidazole-4-carboxamide derivatives (Abbasi-Radmoghaddam et al., 2021). They built a QSAR model based on the genetic algorithm–multiple linear regression (GA–MLR) and least squares–support vector machine (LS–SVM) methods. Moreover, they performed molecular docking analysis to reveal the chemical interactions between the substructure in each compound and PARP-1, as well as to calculate the free energy binding. They reported nine compounds, which given the best value of IC_{50} , showed an improvement in PARP-1 inhibition of 33.394 %. Li and colleagues (2016) used a molecular docking approach to screen compounds from the ZINC database against PARP-1 (Li et al., 2016). Grid and amber scoring were used to calculate the area under the curve from the receiver operating characteristic. The selected compounds were

further analyzed through MD. Finally, they proposed ZINC67913374 as a candidate compound to inhibit PARP-1 activity. Proteochemometry was also performed by Cortés-Ciriano and colleagues (2015) to develop a model to explore the relationship between PARP inhibitors and various PARP isoforms, including PARP-1 (Cortés-Ciriano et al., 2015). They used both chemical (Morgan fingerprints) and protein (binding site amino acid (AADescs) and full protein sequence (SeqDescs) descriptors as independent variables, while thermal shift values retrieved from Differential Scanning Fluorimetry (DSF) were treated as dependent variables. The models were built based on random forests, which were then further examined for the confidence intervals to understand the reliability of the predictive performance for either new compounds or PARP isoforms. Altogether, these studies show that computational approaches are useful to identify novel inhibitors of PARP-1.

In this study, we used Python-based programming to retrieve the biological activities of human PARP-1 from ChEMBL (Mendez et al., 2019). We extracted a total of 2018 non-redundant compounds with known IC_{50} values. All the inhibitors were converted to 12 different molecular descriptors and further built with 12 different machine learning models. Of the 144 models, the PubChem random forest model was chosen, because it was interpretable and it robustly classified substances as active or inactive, as indicated by MCC values > 0.7 of the training and CV sets in all three sampling approaches. Additionally, the important chemical fingerprints that contributed to the constructed model were examined. In-depth analysis of the top 20 descriptors demonstrated that aromatic/heterocyclic and nitrogen-containing characteristics are important for PARP-1 inhibition. Lastly, a web server was built to make this prediction accessible in the public domain. This will accelerate the discovery of new and diverse inhibitors against PARP-1.

MATERIALS AND METHODS

Data compilation and curation

The dataset of PARP-1 (ChEMBL ID: ChEMBL3105) inhibitors was compiled using data from the ChEMBL database, release 29 (Mendez et al., 2019), which includes an initial set of 5094 bioactivity data points and 3738 compounds. The data were retrieved through a Python-based library (<https://pypi.org/project/chembl-webresource-client/>) which enables users to cache all results in the local file system for faster retrieval (Davies et al., 2015). The IC_{50} values, containing 2815 data points and 2429 compounds, were chosen for further curation. Because the purpose of this study was to create a classification model for PARP-1 inhibition, we defined active as $\leq 1 \mu M$ ($n = 1720$) and inactive as $\geq 10 \mu M$ ($n = 298$). The intermediates with concentrations ranging between 1 and $10 \mu M$ were discarded ($n = 334$). Finally, we obtained 2018 non-redundant and curated active and inactive compounds for further analysis.

Molecular descriptor analysis

The PaDEL-Descriptor software was used to calculate molecular fingerprints for each compound in the dataset (Yap, 2011). As previously described by Malik and colleagues (2020), molecular fingerprints are numerical values that represent both qualitative and quantitative chemical structures (Malik et al., 2020). Thus, they are crucial for QSAR studies. The software computes 12 types of fingerprints which belong to nine classes, namely, Atom Pairs 2D, CDK, CDK extended, CDK graph only, E-state, Klekota–Roth, MACCS, PubChem, and Substructure. Moreover, Atom Pairs 2D, Klekota–Roth, and Substructure are available in two versions. The first version indicates the presence or absence of the descriptors using the values 1 and 0, while the second version indicates the descriptor's frequency value. The structures in SMILES format were pre-processed by removing salt, detecting aromaticity, standardizing nitro groups, and standardizing tautomers, before

being subjected to molecular fingerprint calculation.

Data filtering

During the feature selection process, low variance variables were not useful for the model's predictive capability. Therefore, constant and near constant variables were omitted from the selection of fingerprint descriptor sets to reduce model complexity and bias. The constants of the fingerprint descriptors were calculated using a standard deviation (SD) of 0.1 as a cut-off value. Thus, variables with SD values of less than 0.1 were selected for further analysis.

Data splitting for model construction

The Kennard–Stone algorithm was used to divide the data into an 80/20 ratio (Kennard and Stone, 1969), of which 80 % was assigned as an internal set (1614 compounds, active = 1380, inactive = 234) and the remaining 20 % was used as an external set (404 compounds, active = 340, inactive = 64) to validate the model. The internal dataset was further divided into balanced and imbalanced datasets and used as the training dataset, which was subjected to five-fold cross-validation.

Statistical analysis

We present chemical descriptors of each molecule according to the previous study by Schaduangrat and colleagues (2021). Briefly, this uses six common descriptive statistical parameters: minimum (Min), first quartile (Q1), median, mean, third quartile (Q3), and maximum (Max). All the parameters were visualized as a box plot using the seaborn and matplotlib data visualization packages in Python. Lipinski's rule-of-five parameters were compared between active and inactive groups using the Mann–Whitney *U* test, with $p < 0.05$ indicating a significant difference.

Multivariate analysis

Twelve machine learning classification models were constructed from the internal da-

taset: decision trees, extra trees, Gaussian Naive Bayes, Gaussian process, gradient boosting, K-neighbors, light gradient boosted machine, multi-layer perceptron, quadratic discriminant analysis, random forest, C-support vector, and extreme gradient boosting. The model construction was developed using the scikit-learn library (Pedregosa et al., 2011) in Python. Each type of model had different characteristics to determine the relationship between the dependent variables and the independent variables. Gradient boosting, random forest, extra trees, light gradient boosted machine, and extreme gradient boosting were grouped as ensemble methods, which generate many models and combine them to get the best model. Multi-layer perceptron was part of the neural network, which was considered a black box model and could not be interpreted. Decision tree was used to learn simple decision rules retrieved from the data features. K-neighbors is a type of instance-based learning in which the classification of certain data is based on most of its nearest neighbors. Support vector machine draws a hyperplane to separate two or more classes in the best possible manner. The Gaussian process uses a Gaussian distribution to fit random points of data, whereas quadratic discriminant analysis estimates the means and covariances from the data and assigns a new observed data point to the class with the greatest likelihood. Lastly, Gaussian Naive Bayes assumes each feature follows Gaussian distribution, calculates the probability from each feature at a given class, and multiplies all the probabilities of each feature.

Model validation

We used a variety of statistical parameters to evaluate the performance of the models, including true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN). The model's fitness was determined using the following statistical parameters: overall prediction accuracy (Ac), sensitivity (Sn), specificity (Sp), and Matthews correlation coefficient (MCC).

$$Ac = \frac{TP + TN}{(TP + TN + FP + FN)}$$

$$Sn = \frac{TP}{(TP + FN)}$$

$$Sp = \frac{TN}{(TN + FP)}$$

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

Applicability domain analysis

To estimate the chemical space in which the model can make reliable and accurate predictions for compounds based on similarity with the compounds on which the model was constructed, we used the PCA bounding box to determine the applicability domain (AD) of compounds from the training (internal) and test (external) sets. Compounds that fall inside the AD of the model are typically predicted reliably.

Reproducibility research

The data and code used in the study are deposited on GitHub at <https://github.com/tlerk-suthirat/data-driven-PARP1>.

Development of the PARP-1 web server

The best predictive model was exported as model.pkl and is used in the deployed web server developed in Python using Streamlit version 1.12.0. Particularly, the Streamlit web app accepts the input SMILES notation of query molecule and converts this into an image file of the 2D chemical structure via rdkit-pypi version 2022.3.5. Subsequently, the SMILES notation is used to compute the PubChem molecular fingerprint using padelpy version 0.1.10. The best machine learning model, which was built using the random forest algorithm with scikit-learn version 1.0.2, is applied on the computed fingerprint of the query molecule where the bioactivity is predicted. The PARP1pred web app is publicly available at <https://parp1pred.streamlit.app/> while the data and code used for building this app is deposited on GitHub at <https://github.com/dataprofessor/parp1>.

RESULTS AND DISCUSSION

The entire workflow for constructing the model is summarized in Figure 1.

Chemical space analysis

The aim of performing chemical space analysis between active and inactive compounds is to understand the difference in chemical characteristics between two groups. We first explored the relationship between molecular weight (MW) and the Ghose–Crippen–Viswanadhan octanol-water partition coefficient (LogP), as shown in Figure 2 (Wildman and Crippen, 1999). LogP is a lipophilic descriptor that can be used to determine the permeability of molecules to the cell membrane, thereby indicating their drug-likeness molecule (van de Waterbeemd, 2008). Next, Lipinski's rule-of-five (Ro5) descriptors were employed to investigate the difference in chemical features between active and inactive compounds, as shown in Figure 3. The Ro5 are composed of four parameters, namely MW (< 500 kDa), LogP (< 5), the number of H-bond donors (NumHDonors < 5), and the number of H-bond acceptors (NumHAcceptors < 10) (Lipinski et al., 2001). If any compounds have values out of range for two parameters, they are likely to have poor absorption or permeability, and thus a higher rate of drug development failure. As illustrated in Figure 2, most compounds clustered between 300 and 500 MW with a LogP of 2–4. Moreover, the Ro5 analysis and statistical analysis revealed that most of the active and inactive compounds following the Ro5 as illustrated by the box plots were under the cut-off values (dashed line, Figure 3). The Mann–Whitney *U* test found a significant difference in MW, NumHDonors, and NumHAcceptors between active and inactive molecules, but no difference in LogP. Active molecules had a higher MW, NumHDonors, and NumHAcceptors than inactive molecules, as demonstrated by the circle in the boxplot (Figure 3). The mean \pm SD of MW in the active and inactive groups were 381.66 ± 87.93 and 349.35 ± 119.32 , respectively. NumHDonors had a mean \pm SD of 1.80 ± 0.82 for active molecules and 1.39 ± 0.79

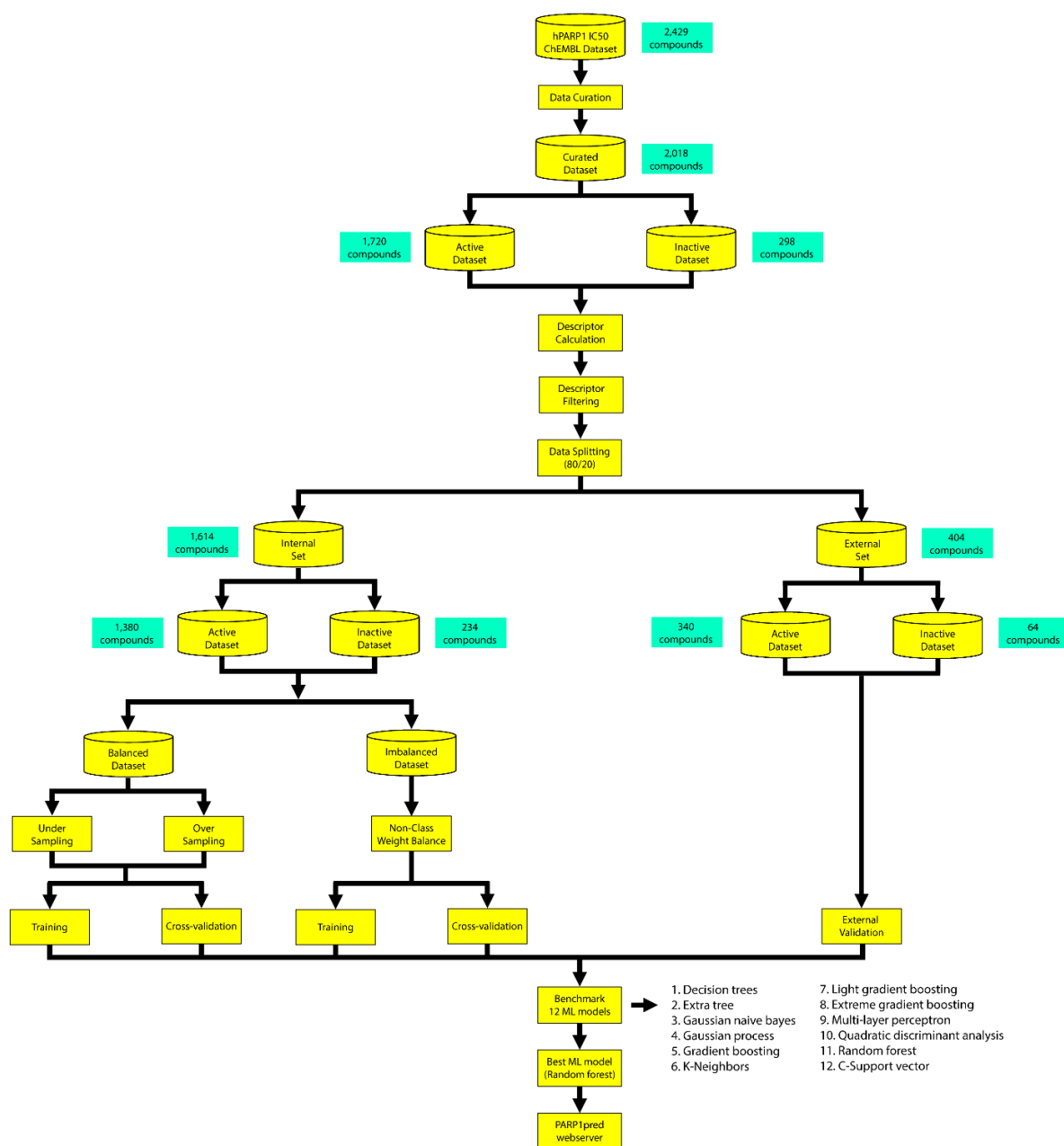


Figure 1: Overall workflow of the development of the webserver for PARP-1 inhibitors

for inactive molecules, whereas the mean SD for NumHAcceptors was 4.56 ± 1.53 for active molecules and 4.22 ± 2.06 for inactive molecules. Between the active and inactive molecules, the logP value was 2.66 ± 1.17 for active molecules and 2.61 ± 1.52 for inactive molecules.

QSAR modeling

To develop a robust QSAR model, we followed the guidelines of the Organization for

Economic Co-operation and Development (OECD, 2014). Briefly, a robust model should include, at least: 1) a defined endpoint for the dataset; 2) an unambiguous learning algorithm; 3) a defined applicability domain of the QSAR model; 4) appropriate measures of goodness-of-fit, robustness, and predictability; and 5) mechanistic interpretation of the QSAR model. Thus, to develop interpretable QSAR models, the molecular fingerprints indicated in Table 1 were calculated using the

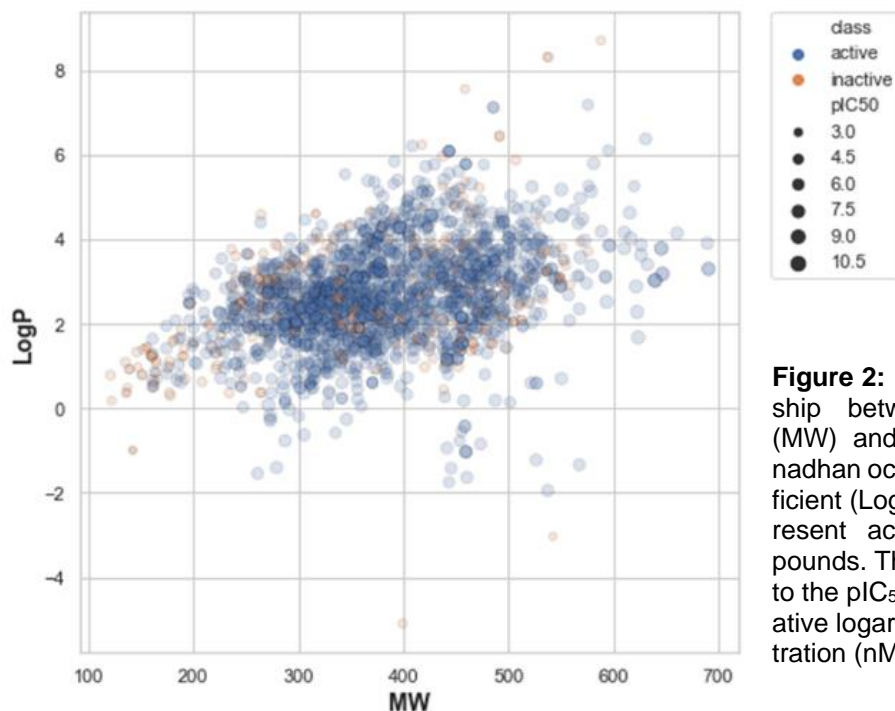


Figure 2: Illustration of the relationship between molecular weight (MW) and Ghose–Crippen–Viswanadhan octanol–water partition coefficient (LogP). Blue and orange represent active and inactive compounds. The size of the circle refers to the pIC_{50} value, which is the negative logarithmic of the IC_{50} concentration (nM).

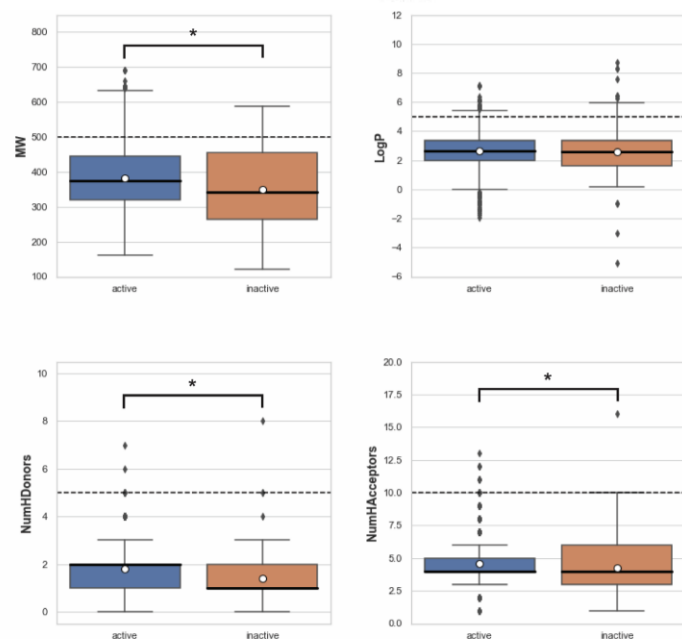


Figure 3: Box plots of Lipinski's rule-of-five descriptors comparing between active and inactive groups. The dashed line represents cut-off values indicating drug-like molecules: molecular weight (MW) < 500, Ghose–Crippen–Viswanadhan octanol–water partition coefficient (LogP) < 5, number of hydrogen bond donors (NumHDonors) < 5, number of hydrogen bond acceptors (NumHAcceptors) < 10. A circle represents the mean, and an asterisk indicates a significant difference between two groups ($p < 0.05$).

Table 1: Twelve different sets of fingerprint descriptors derived from the PaDEL-Descriptor software

Fingerprint	Number	Description
Atom Pairs 2D	780	Presence of atom pairs at various topological distances
Atom Pairs 2D Count	780	Count of atom pairs at various topological distances
CDK	1024	Fingerprint of length 1024 and search depth of 8
CDK extended	1024	Extends the fingerprinter with additional bits describing ring features
CDK graph only	1024	Specialized version of the fingerprinter which does not take bond orders into account
E-state	79	E-State fragments
Klekota-Roth	4860	Presence of chemical substructures
Klekota-Roth Count	4860	Count of chemical substructures
MACCS	166	MACCS keys
PubChem	881	PubChem fingerprint
Substructure	307	Presence of SMARTS patterns for functional group classification
Substructure Count	307	Count of SMARTS patterns for functional group classification

PaDEL-Descriptor software, from which three fingerprints (PubChem, Substructure, and Klekota–Roth) are readily interpretable.

We constructed 12 machine learning models from 12 molecular fingerprints to determine which model gave the best performance and was the most robust and interpretable. Because our imbalanced data contained more active compounds ($n = 1720$) than inactive compounds ($n = 298$), we compared the models generated from both balanced and imbalanced approaches. Prior to data splitting, we reduced the dimensionality of the data by selecting the fingerprint that rendered $SD < 0.1$. The data were split into external and internal sets in an 80:20 ratio. The internal dataset ($n = 1614$), which contained 1380 active and 234 inactive compounds, was further divided into balanced and imbalanced datasets. For the balanced dataset, the models were created based on two methods: 1) undersampling, which randomly selected the majority class equal to the number of the minority classes; 2) oversampling, which amplified the number of minority classes equal to the number of the majority class.

For the non-class weight balance of an imbalanced dataset, the data were randomly selected to develop the model without consideration of the ratio between major and minority classes. Figures 4 and 5 demonstrate the heat maps of MCC_{train} , MCC_{cv} , MCC_{test} , $MCC_{train-cv}$, and $MCC_{train-test}$ for each fingerprint, machine learning model, and sampling approach.

Results showed that a balanced oversampling approach yielded the best value—most of the MCC_{train} and MCC_{cv} values were more than 0.8. Moreover, most of the MCC_{test} values of oversampling were more than 0.7. The values of $MCC_{train-cv}$ in the oversampling group were lower than 0.2, whereas the values of $MCC_{train-test}$ in both balanced oversampling and imbalanced non-class weight were generally better than balanced undersampling, as the $MCC_{train-test}$ values of undersampling were mostly greater than 0.3.

As a result, we considered the oversampling approach as a good candidate to compare the performance among each model and fingerprint. Figure 4B demonstrates that Gaussian Naive Bayes and quadratic discriminant analysis did not yield acceptable MCC values (< 0.7) for all fingerprints. We further selected random forest (RF) over other machine learning methods because relevant features were able to be observed and the model was easily interpretable. As mentioned in the Methods section, RF is an ensemble method that has a root node as a starting point and splits into an N number of decision trees to learn the inherent patterns from the input data (Breiman, 2001). Following a thorough examination of all MCC values for the interpretable fingerprints—PubChem, Substructure, and Klekota–Roth—the result suggested that a model based on PubChem was a good candidate. This was demonstrated by the MCC values for PubChem in the training, cross-validation, and test sets of 1, 0.96, and 0.74, respectively, whereas the MCC values for Substructure and Klekota–Roth in the test set were 0.66 and 0.68, respectively. As a result, the RF model that was developed using the oversampling approach from the PubChem fingerprint was the best option for model interpretation. Furthermore, as indicated in Figure 6, the applicability domain was determined using the PubChem fingerprint as the input for PCA analysis. A total of 2018 compounds were split into two subsets, which consisted of internal (80 %) and external (20 %) datasets using the Kennard–Stone algorithm (Kennard and Stone, 1969). The internal set was used as the training dataset, subjected to random sampling, and the predictive model was constructed with five-fold cross-validation. The result showed that the chemical space distribution of the external dataset fits well with the internal dataset, indicating that the applicability domain was well defined for the QSAR-based classification model.

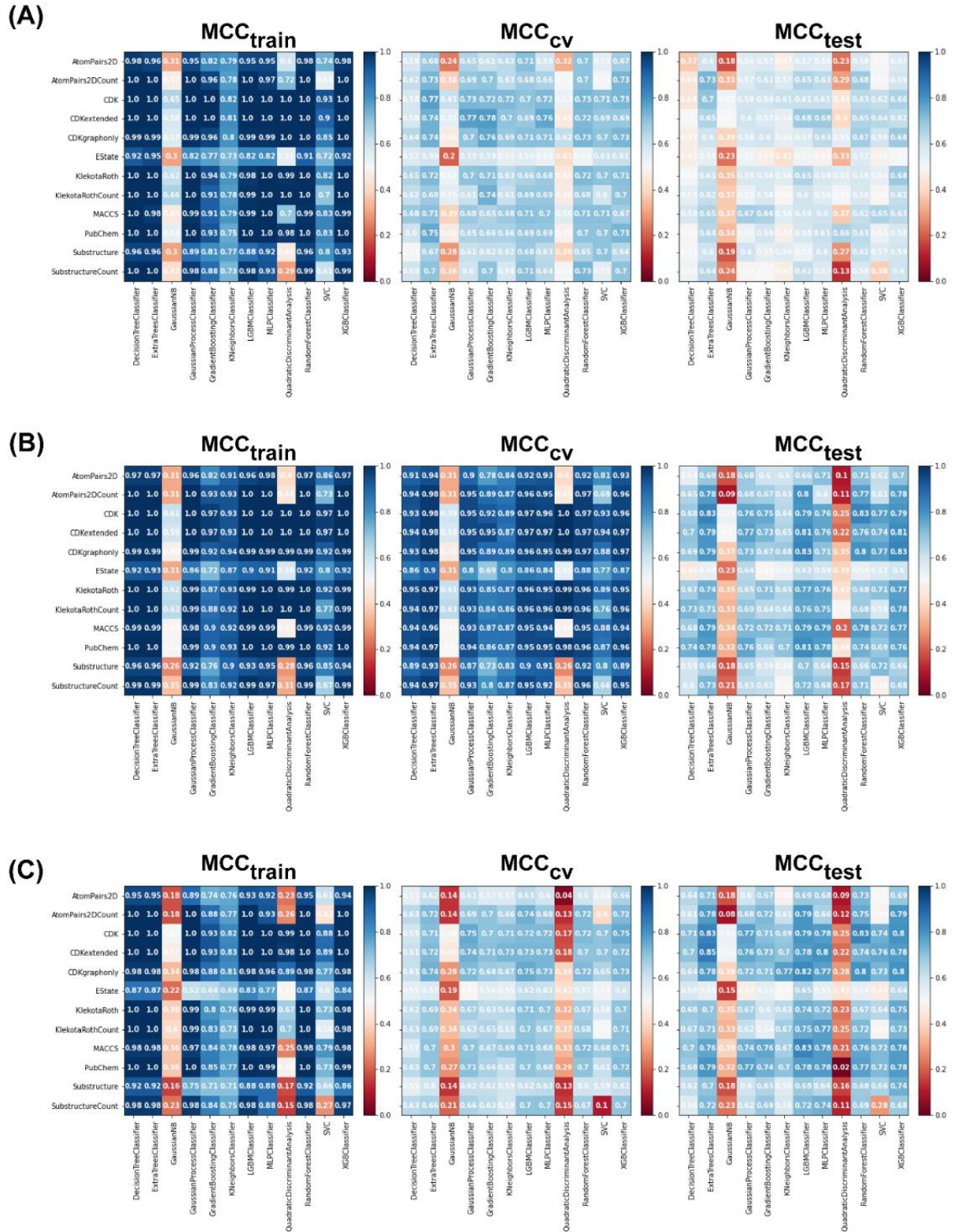


Figure 4: Heat maps of the MCC values of the training, CV, and test sets for each data sampling approach. (A) Balanced undersampling, (B) balanced oversampling, and (C) imbalanced non-class weight. Abbreviations: MCC, Matthews correlation coefficient; CV, cross-validation; gaussianNB, Gaussian Naive Bayes; LBMC, light gradient boosted machine; MLP, multi-layer perceptron; SVC, C-support vector; XGB, extreme gradient boosting

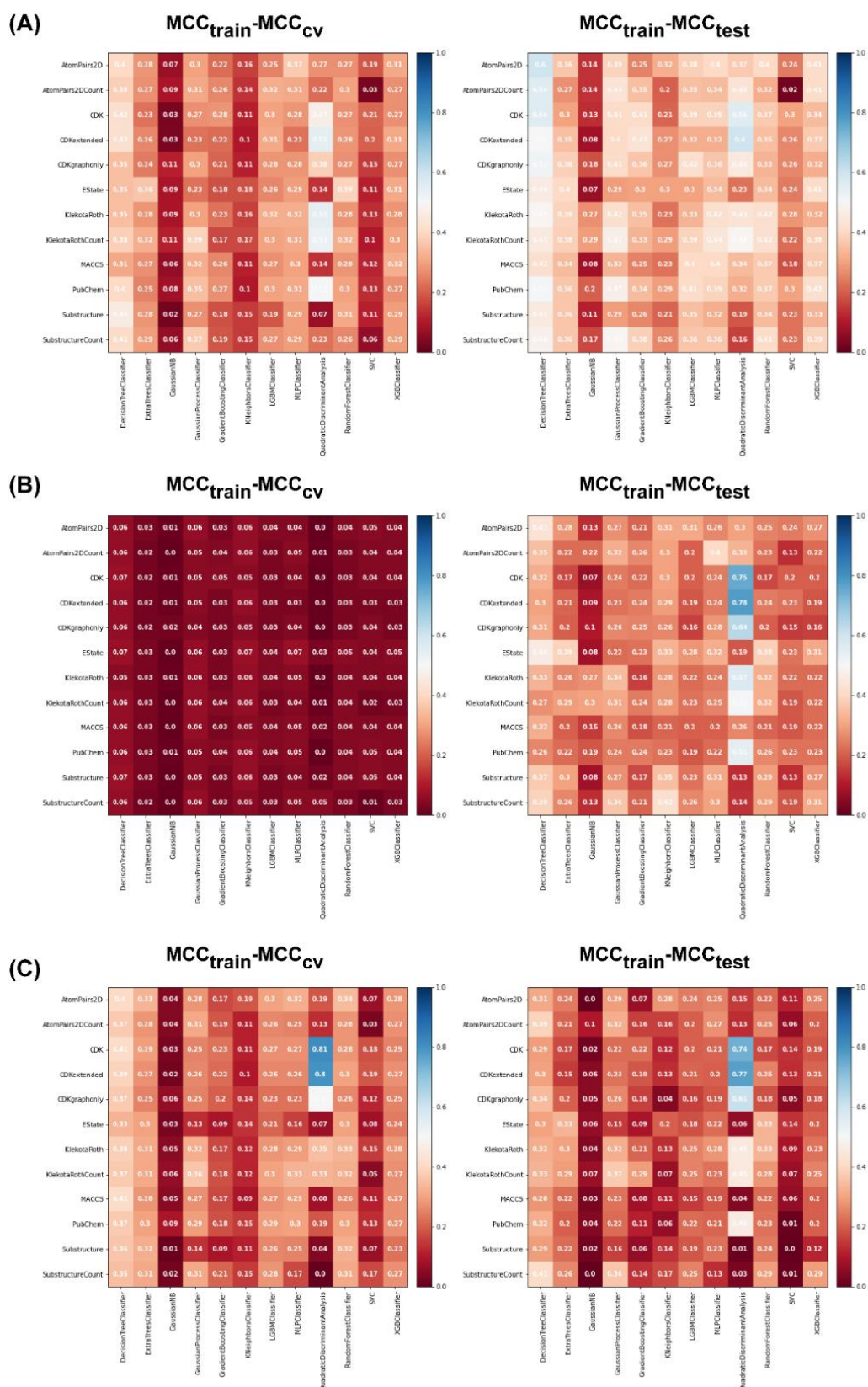


Figure 5: Heat maps of MCC_{train}-MCC_{cv} and MCC_{train}-MCC_{test} for each data sampling approach. (A) Balanced undersampling, (B) balanced oversampling, (C) imbalanced non-class weight. Abbreviations: MCC, Matthews correlation coefficient; CV, cross-validation; gaussianNB, Gaussian Naive Bayes; LBMC, light gradient boosted machine; MLP, multi-layer perceptron; SVC, C-support vector; XGB, extreme gradient boosting

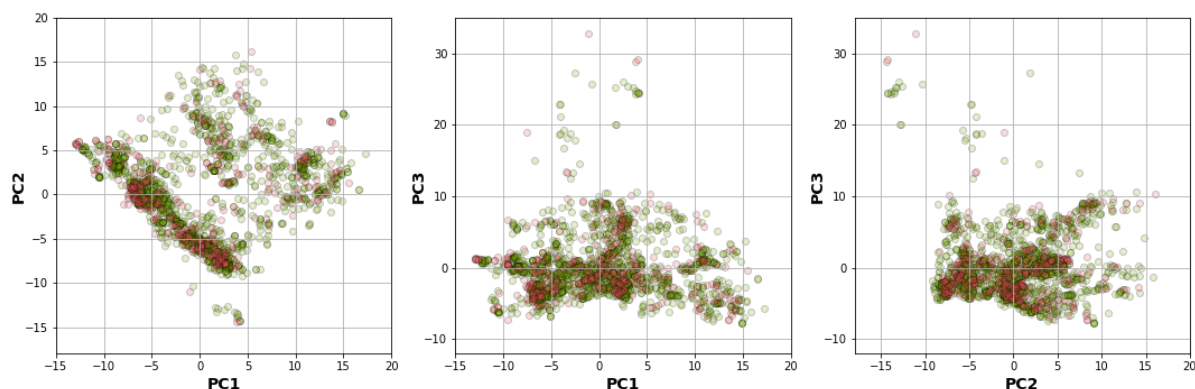


Figure 6: Plot of PCA scores for applicability domain analysis. The score plot indicates the distribution of chemical space of the internal (green) and external (red) datasets, which were used to determine the applicability domain of the PARP-1 inhibitors dataset.

Mechanistic interpretation of feature importance

To gain a better understanding of the mechanisms underlying PARP-1 activity and the significance of the features used to develop a PARP-1 activity predictability model using RF, the mean decrease of the Gini index was used to rank the importance of the PubChem feature descriptors. Measuring feature importance in RF can be evaluated by the mean decrease accuracy and the mean decrease in Gini; however, the latter gives more robust results (Calle and Urrea, 2010). Thus, we selected the top 20 PubChem substructures with the highest Gini index, illustrated in Figure 7, and their corresponding substructure descriptions are shown in Table 2. We grouped the functional groups of the PubChem fingerprints into four classes: 1) aromatic, cyclic/heterocyclic, and ring counts; 2) nitrogen-containing, consisting of hydrazine, amine, imine, and amide; 3) atom counts; and 4) ether, aldehyde, and alcohol. However, some PubChem fingerprints had more than one feature; for example, PubChemFP695 had aldehyde and amine functional groups, and PubChemFP821 had cyclic and amine functional groups.

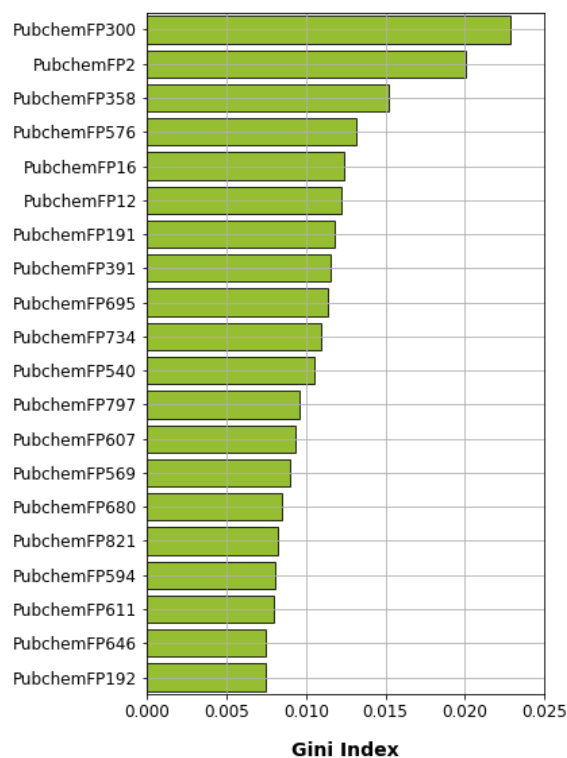


Figure 7: Feature importance plot as rationalized by Gini index obtained from random forest model using oversampling

Table 2: Descriptions of SMARTS patterns and substructures from the top 20 Gini indices

Rank	Features	SMARTS pattern	Substructure description
1	PubChemFP300	N-N	Hydrazine
2	PubChemFP2	>= 16 H	Greater than or equal to 16 hydrogen atoms
3	PubChemFP358	C(~C):N	Ethanamine
4	PubChemFP576	N=C-C:C-[#1]	Propan-1-imine
5	PubChemFP16	>= 4 N	Greater than or equal to 4 nitrogen atoms
6	PubChemFP12	>= 16 C	Greater than or equal to 16 carbon atoms
7	PubChemFP191	>= 2 unsaturated non-aromatic heteroatom-containing ring size 6	Greater than or equal to 2 unsaturated non-aromatic heteroatom-containing ring size 6
8	PubChemFP391	N(~C)(~C)(~C)	N,N-dimethylmethanamine
9	PubChemFP695	O=C-C-C-C-C-N	5-aminopentanal
10	PubChemFP734	Cc1cc(C)ccc1	1,3-Xylene
11	PubChemFP540	C-N-C-[#1]	N-methylmethanamine
12	PubChemFP797	CC1CC(C)CCC1	1,3-Dimethylcyclohexane
13	PubChemFP607	N-C-C-C:C	Butan-1-amine
14	PubChemFP569	N-C-C-N	Ethane-1,2-diamine
15	PubChemFP680	O-C-C-C-C-C	Pentan-1-ol
16	PubChemFP821	CC1C(N)CCCC1	2-Methylcyclohexan-1-amine
17	PubChemFP594	C-O-C-C=C	3-Methoxyprop-1-ene
18	PubChemFP611	N-C-C-N-C	N'-methylethane-1,2-diamine
19	PubChemFP646	O=C-N-C-[#1]	N-methylformamide
20	PubChemFP192	>= 3 any ring size 6	Greater than or equal any ring size 6

Aromatic, cyclic/heterocyclic, and ring count functional groups

The fingerprints belonging to these groups consisted of PubChem191, PubChem734, PubChem797, PubChem821, and PubChem192. PubChem192 was on the lowest rank of the top 20 and it was not specified whether it was aromatic- or heteroatom-containing, but it must have a ring size of six for at least three rings. Thus, the aromatic (PubChemFP734) and cyclic (PubChem191, PubChem797, and PubChem821) moieties overlapped with PubChem192. Based on aromatic, cyclic/heterocyclic, and ring counts, PubChem191, PubChem797, and PubChem821 were at the 7th, 12th, and 16th positions of the top 20. Taking a closer look at our post-processing dataset (2018 compounds), there were 35 compounds in total containing all three fingerprints, of which 34 compounds were considered active. Moreover, a total of

33 compounds contained both cyclic (PubChem191, PubChem797, and PubChem821) and aromatic moieties (PubChemFP734), and all of them were active. This meant that aromatic and cyclic/heterocyclic functional groups with a ring size equal to six or more than two were the important features of the active compounds. The first generation of PARP-1 inhibitors was designed to mimic the benzamide scaffold of NAD⁺ (Steffen et al., 2013). Later the efficacy was improved by using quinazolinone as a scaffold to synthesize PARP-1 inhibitors (Malyuchenko et al., 2015). Inhibitors derived from those two scaffolds contain both the aromatic and cyclic/heterocyclic moieties and play an important role in the NAD⁺ binding pocket. The aromatic ring forms π - π interactions with the tyrosine residues in the NAD⁺ binding pocket, and both the aromatic ring and cyclic/hetero-

cyclic moieties form hydrophobic interactions with the hydrophobic residues in the NAD⁺ binding pocket. The crystal structure of human PARP-1 revealed a hydrophobic interaction between the quinazolinone part of the FR257517 inhibitor and the phenyl ring of Tyr907 and a CH- π interaction with C β of Tyr869 (Kinoshita et al., 2004). Moreover, docking analysis between PARP-1 and tricyclic compounds containing a non-aromatic A-ring demonstrated the fit within the NAD⁺ binding pocket, even though the non-aromatic A-ring was not flat (Park et al., 2010). Most of the active compounds reported herein had IC₅₀ values ranging from 0.013–0.695 μ M. It should be noted that PubChem191 was in the highest rank among the aromatic, cyclic/heterocyclic, and ring counts groups. This could be explained by the nitrogen in the non-aromatic moiety of the inhibitors contributing to hydrogen bonds forming with the glycine in the NAD⁺ binding pocket. The crystal structure of PARP-1 conjugated with FR257517 revealed three hydrogen bonds, one from the NH of the quinazolinone part of FR257517 to Gly863-C=O (Kinoshita et al., 2004). In addition, cyclic benzamide derivatives increased potency in PARP-1 and led to the optimization of novel PARP-1 inhibitors. Steinhagen and colleagues (2002) reported that core variations within the cyclohexene moiety of PubChem191 affected the potency of inhibitors (Steinhagen et al., 2002). Moreover, the study demonstrated that substitution of the 3,6-dihydro-2-thiopyrane subunit yielded a three- to tenfold increase in potency compared with the cyclohexenyl moiety.

Nitrogen-containing functional groups, including hydrazine, amine, imine, and amide

This class of functional groups possessed the largest number of fingerprints, including hydrazine (PubChemFP300), amine (PubChemFP358, PubChemFP391, PubChemFP695, PubChemFP540, PubChemFP607, PubChemFP569, PubChemFP821, and PubChemFP611), amide (PubChemFP646), and imine (PubChemFP576). There were two fingerprints in

this group, PubChemFP695 and PubChemFP821, also containing aldehyde and cyclic functional groups, respectively.

PubChemFP300 was in the first rank of important fingerprints based on all features. This is because PubChemFP300 is part of the basic scaffold during PARP-1 inhibitor development (Ferraris, 2010). Banasik and colleagues (1992) introduced pthalazine derivatives and analogues as part of the development of PARP-1 inhibitors (Banasik et al., 1992). Moreover, Xu and colleagues (2014b) synthesized a series of compounds which contained tetraaza phenalen-3-one as a main scaffold to inhibit PARP-1 (Xu et al., 2014b). The compounds sensitized tumor cells to ionizing radiation and temozolomide. Ji and colleagues (2015) used phthalic hydrazide as a pharmaceutical scaffold to synthesize novel PARP-1 inhibitors (Ji et al., 2015). Another study produced novel PARP-1 inhibitors by fusing a pyrazolo pyridin-2-one to a non-aromatic heterocycle or carbocycle. These resulted in a vast variety of IC₅₀ values, ranging from 0.002 to >10 μ M (Moree et al., 2008).

As well as PubChemFP300, another four fingerprints were within the top ten important features: PubChemFP358 (3rd rank), PubChemFP576 (4th rank), PubChemFP391 (8th rank), and PubChemFP695 (9th rank). PubChemFP358 is part of the benzamide scaffold, thus making it critical for PARP-1 inhibitor synthesis because this scaffold mimics the NAD⁺ substrate. This scaffold has been maintained through all generations of PARP-1 synthesis (Malyuchenko et al., 2015). As previously mentioned, the crystal structures revealed that NH in the quinazolinone scaffold of FR257517 forms a hydrogen bond with the Gly863-C=O that is required for the inhibitor to remain in the NAD⁺ binding pocket (Kinoshita et al., 2004). Moreover, PubChemFP358 is part of the pendant fluorobenzyl group that participates in the adenine-ribose binding pocket within the NAD⁺ binding site (Pescatore et al., 2010).

PubChemFP576 is part of the pyridine and pyrimidine moieties. Moree and colleagues (2008) fused a pyrazolo pyridin-2-

one to a non-aromatic heterocycle or carbocycle to generate novel PARP-1 inhibitors (Moree et al., 2008). The fused structures were designed based on the observation that pyrazolo pyridin-2-one showed a similar binding mode between chicken PARP-1 (PDB: 1PAX) and the Parke–Davis/Pfizer inhibitor. Ferraris and colleagues (2003) synthesized a series of aza-5[*H*]-phenanthridine-6-inhibitors where nitrogen atoms were introduced to the 5[*H*]-phenanthridin-6-one core at different positions to compare the potency (Ferraris et al., 2003b). Moreover, this fingerprint was part of the tetraaza phenalen-3-one (Xu et al., 2014b), 4-benzyl-2*H*-phthalazin-1-one (Mear et al., 2008), and 4-[4'-fluoro-3'-(piperazine-1'-carbonyl)benzyl]-2*H*-phthalazin-1-one cores (Zmuda et al., 2015). Torrisi and colleagues (2010) demonstrated that introduction of 3-pyridyl to a hexahydrobenzophthyrinone pharmacophore resulted in metabolic stability (Torrisi et al., 2010).

PubChemFP391 represents the tertiary amines that Ferraris and colleagues (2003a) added to the partially saturated aza-5[*H*]-phenanthridine-6-ones to increase aqueous solubility (Ferraris et al., 2003a). Moreover, it is part of the optimal nitrogen substituent of the hexahydrobenzophthyrinone pharmacophore to synthesize diverse ranges of PARP-1 inhibitors that was synthesized by Torrisi and colleagues (2010). Pescatore and colleagues (2010) synthesized a series of pyrrolo[1,2-*a*]pyrazin-1(2*H*)-one to inhibit PARP-1 (Pescatore et al., 2010). Additionally, the same study revealed that the pyrrolo[1,2-*a*]pyrazin-1(2*H*)-one scaffold exhibited good potency and inhibited *BRCA*-deficient tumor cells. Rhee and colleagues (2009) used isoquinolinone-based tetracycles as the main scaffold to develop PARP-1 inhibitors (Rhee et al., 2009). Based on this fingerprint, some of the compounds from this study exhibited an IC_{50} lower than 1 μ M. Zhou and colleagues (2017) made a group of compounds called fused tetra- or penta-cyclic compounds, in which one part of the ring had

a tertiary amine as a spacer to link other substituents, that showed diverse ranges of enzymatic activity (Zhou et al., 2017).

PubChemFP695 overlapped with both PubChemFP358 and PubChemFP191, which are important for the NAD^+ binding pocket. Moreover, PubChemFP695 was part of tricyclic derivative PARP-1 inhibitor synthesis (Myung-Hwa et al., 2014), and substituents participated in the adenine-ribose (AD) binding site within the NAD^+ binding pocket (Scarpelli et al., 2010). PubChemFP695 is a component of proline derivatives and contributes to lipophilicity, which is necessary for cell permeability (Dunn et al., 2012). This was confirmed by introducing the polar carboxylic acid moiety to proline derivatives, resulting in less cell-based activity. Moreover, PubChemFP695 also overlapped with PubChemFP391, making this fingerprint part of the AD binding site.

Collectively, this suggests that nitrogen-containing fingerprints are important in model construction.

Ether, aldehyde, and alcohol functional groups

One fingerprint, PubChem695, which contained both aldehyde and amine functional groups, is categorized in this class and has been discussed previously. The remaining fingerprints falling into this class, PubChem680 (15th rank) and PubChem594 (17th rank), were not ranked in the top ten important features. Based on our curated dataset ($n = 2018$), few compounds contained these fingerprints: PubChem680, $n = 714$ (18th rank); and PubChem594, $n = 468$ (18th rank). PubChem680 is composed of alkane and alcohol functional groups and participates in the nicotinamide-ribose (NI) and AD binding sites within the NAD^+ binding pocket. The study led by Ferraris and colleagues (2003b) replaced the C=O of the amide group from the benzamide scaffold with C-OH, which resulted in IC_{50} values ranging from 14–0.042 μ M (Ferraris et al., 2003b). This suggests that OH could be able to maintain a hydrogen bond within the

NAD⁺ binding pocket. Additionally, this fingerprint served as an o-linked spacer between two distinct pharmacophores, one of which was responsible for the NI binding site and the other for the AD binding site, as demonstrated by Park and colleagues (2010) via the synthesis of a series of 1,2-dihydro-4H-thiopyrano[3,4-c]quinolin-5(6H)-one derivatives (Park et al., 2010). As part of the AD binding site, this fingerprint also overlapped with PubChemFP695, which contributes to aqueous solubility and cellular permeability, as previously mentioned.

PubChemFP594 is part of the pyran and was found to play roles in both the NI and AD binding sites within the NAD⁺ binding pocket. Several studies have used pyran as part of the scaffold. For example, introducing a dihydropyran to the A-ring caused the derivatives to be more polar but less potent toward PARP-1 inhibition (Shultz et al., 2013). Xu and colleagues (2014a) filed the patent on the synthesis of diazabenz[de]anthracen-3-one derivatives that contain pyran as part of the tri-cyclic ring (Xu et al., 2014a). All the compounds reported in this study were categorized as active compounds. Conversely, the patent filed by Cheung and colleagues (2015) revealed mostly inactive compounds against PARP-1 (Cheung et al., 2015). For the AD binding site, this fingerprint participated in phenyl derivative substituents, as demonstrated by Orvieto and colleagues (2009) when they introduced methyl groups to the aromatic ether (Orvieto et al., 2009). They found that this improved the inhibitory effect compared with its parental phenyl. As previously mentioned, PubChemFP594 also functionally overlapped with PubChem680, as part of the o-linked spacer between two binding modes of pharmacophores.

Structural interpretation

PARP-1 has three important domains: 1) the DNA binding domain, 2) the catalytic domain, and 3) the nuclear acceptor protein (Ferraris, 2010). The catalytic domain is subdivided into: 1) the helical domain (HD), and

2) the ADP-ribosyl transferase (ART) domain, as illustrated in Figure 8 (Patel et al., 2012). Most of the compounds were synthesized to inhibit the catalytic domain that consists of three subsites: 1) the nicotinamide-ribose binding site (NI), 2) the phosphate binding site (PH), and 3) the adenine-ribose binding site (AD), and the inhibitors were designed to mimic the nicotinamide scaffold of NAD⁺ (Kinoshita et al., 2004). Thus, all generations of PARP-1 inhibitors have maintained the basal chemical interaction network between the inhibitors and the key amino acids within the NI binding site (Malyuchenko et al., 2015). These key amino acids include Gly863 (nitrogen of the α -amine) and Ser904 (oxygen of the R-group) forming hydrogen bonds with either C=O or C-OH of inhibitors. The oxygen of the carboxyl group of Gly863 forms a hydrogen bond with either the nitrogen-containing ring of inhibitors or the NH group of the nicotinamide scaffold, whereas the hydrogen of the amino group of Gly863 forms a hydrogen bond with either the C=O or C-OH of the inhibitors, as illustrated in Figure 8. Additionally, π - π and hydrophobic interactions between the side chain of Tyr896 and Tyr907 in PARP-1 and either the cyclic or aromatic ring of inhibitors contribute to the NI binding site, as shown in Figure 8. These interactions were shown by the co-crystallization of chicken PARP-1, which is highly conserved with human PARP-1 (sequence identity and similarity, 79 % and 89 %, respectively), with three different inhibitors: 6-amino-benzo[de]isoquinoline-1,3-dione (4ANI), 3-methoxybenzamide (3MBA), and 8-hydroxy-2-methyl-3-hydro-quinazolin-4-one (NU1025) (Kinoshita et al., 2004; Ruf et al., 1998). The importance of the chemical interaction network has been confirmed through site-directed mutagenesis on human PARP-1. Ruf and colleagues (1998) demonstrated that G863A, Y896N, and Y907N reduced PARP-1 activity to 70 %, 15 % and 1.1 %, respectively, compared with wildtype (Ruf et al., 1998).

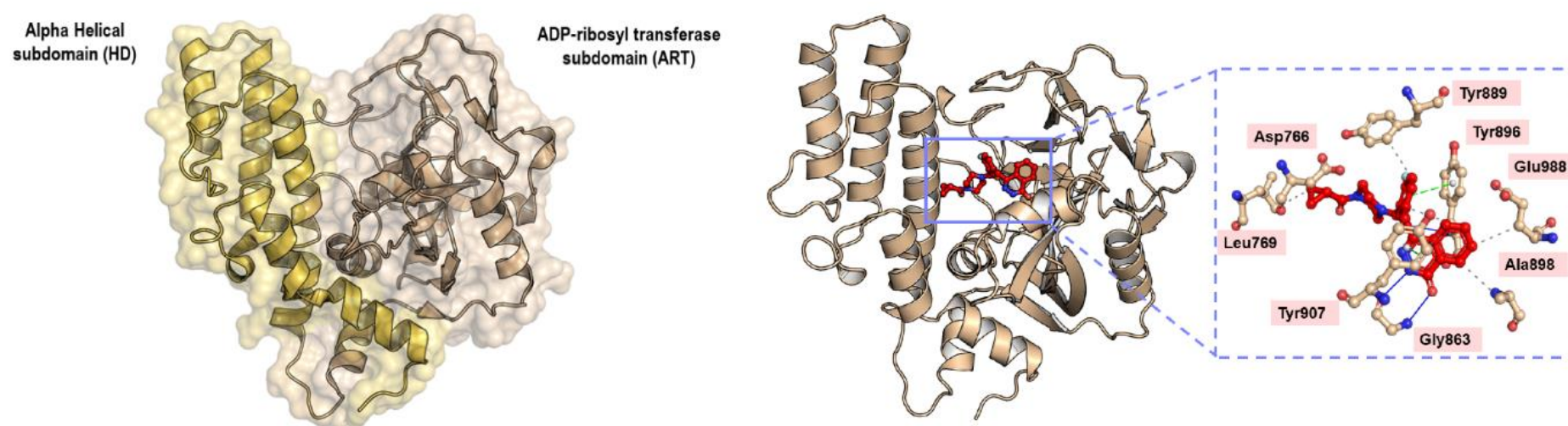


Figure 8: Crystal structure of the catalytic domain of PARP-1 (PDB ID 1UK0) and the interaction network between PARP-1 and olaparib (PDB ID 7KK4). The alpha-helical subdomain (HD) is shown in light orange color while the ADP-ribosyl transferase subdomain (ART) is shown in wheat color. Hydrogen forming network (blue solid line), π - π (green dashed line), and hydrophobic (grey dashed line) interactions between key amino acids within the nicotinamide binding site and olaparib

To improve the potency of PARP-1 inhibitors, because the NI binding site is found in other NAD⁺ binding proteins, the development of PARP-1 inhibitors was extended to use the AD binding site to increase the selectivity of PARP-1 inhibition. In particular, this helps to differentiate between PARP-1 and PARP-2, which share very high similarity at the active site, and double knockout of PARP-1 and PARP-2 is lethal during embryogenesis (Ménissier de Murcia et al., 2003). PARP-2 knockout in mice also demonstrated a role in maintaining the genetic integrity of hematopoietic stem/progenitor cells (Farrés et al., 2013). Cross-reactivity of inhibitors with PARP-2 could therefore have significant side-effects.

The amino acids making up the AD binding site include Glu763, Asp766, Asn767, Leu769, Asp770, His862, Ser864, Asn868, Ile872, Gly876, Ile877, Arg878, and Ala880, as defined by several co-crystal structures (Kinoshita et al., 2004; Patel et al., 2012, 2014). Glu763, Asp766, Asn767, and Asp770 are part of the helical domain which uncoils upon DNA-binding activation, thus enabling inhibitors to insert into the catalytic pocket (van Beek et al., 2021). Ishida and colleagues (2006) used structure-based drug design to understand the different interactions of inhibitors between PARP-1 and PARP-2 (Ishida et al., 2006). They discovered that two chemical frameworks, quinazolinone and quinoxaline derivatives, fit the AD binding site differently and inhibit PARP-1 and PARP-2, respectively. Zhao and colleagues (2017) modified the spacer and the *N*-Boc-pyrrolidin-3-yl subunit of a quinazoline-2,4(1*H*,3*H*)-dione derivative to adjust the interaction within both the spacer and the AD binding site (Zhao et al., 2017). Moreover, Zhou and colleagues (2021) exploited the unique AD binding site between PARP-1 and PARP-2 to generate a series of quinazoline-2,4(1*H*,3*H*)-dione derivatives with a variety of substituted cyclic amines (Zhou et al., 2021). They reported that compound 24, which had an (*R*)-3-ethyl piperazine ring, showed high enzymatic potency

and selectivity toward PARP-1. This compound also demonstrated an acceptable pharmacokinetic profile and reduced tumor growth in xenograft and orthotopic models of breast cancer and glioblastoma, respectively. Co-crystallization of PARP-1 with compounds 4 (PDB ligand ID 6WZ) and 6 (PDB ligand ID 6X2) demonstrated a favorable hydrophobic interaction of either the methyl or ethyl substituent on the piperazine ring with the key amino acids His862 and Leu877. Additionally, the substituents on the piperazine nitrogen projected onto a key subpocket consisting of Asp766, Leu769, and Asp770 in PARP-1. Leu769 is replaced by Gly338 in PARP-2, and so this was used as rational for PARP-1 selectivity. Johannes and colleagues (2021) attached various aryl piperazines to an 8-chloroquinazolinone core and found that the interactions between 1) the piperazine moiety and His862 through water molecules and 2) the imidazole moiety and Asp770 via a hydrogen bond resulted in selectivity toward PARP-1 (Johannes et al., 2021). Yu and colleagues (2022) used the key amino acid differences between PARP-1 (Gln759, Glu763, and Asp766) and PARP-2 (Gln324, Ser328, and Gln332) and further modified rucaparib to obtain increased selectivity of PARP-1 inhibitors (Yu et al., 2022). They discovered that Y49 showed excellent selectivity (IC₅₀ of PARP-1 and PARP-2, 0.96 nM and 61.90 nM, respectively). Molecular docking demonstrated hydrogen bond formation between the amino group of 4-aminopiperidine-1-yl with Glu763 and Asp766 in PARP-1, whereas 4-aminopiperidine-1-yl caused steric hindrance in PARP-2. Thus, they suggested that nitrogen-containing basic substituents were required to fit into the hydrophilic pocket formed by acidic amino acids around the AD site.

Model deployment as web server

To facilitate accessibility for non-chemoinformatic scientists who intend to determine whether their compounds have PARP-1 inhibitory activity, a public web server was created. Thus, the predictive

model, PARP1pred, is available at <https://parp1pred.streamlitapp.com>.

Briefly, the PARP1pred web server uses SMILES as the input for the query compound. PadelPy is used to convert SMILES to PubChem fingerprints, which are then used as an input to trained classification models whose outputs are reported as active or inactive (Figure 9).

CONCLUSION

In the era of precision medicine, targeting of DNA repair is effective in killing cancer cells. PARP-1 plays a role in DNA damage and repair, and is a well-known target for cancers with *BRCA1/2* mutations. Several drugs targeting PARP-1 have been FDA approved;

however, accessing such targeted drugs is problematic because of their high cost, particularly in middle- and low-income countries. Thus, advancements in drug development would contribute to the alleviation of such access constraints. In this study, computer-aided drug design was used to understand the relationship between the chemical structures of inhibitors and PARP-1 through the QSAR building model. Understanding such relationships will facilitate rational drug design to effectively target PARP-1. Our study retrieved a set of biological activities from the ChEMBL database that contained 2018 non-redundant compounds. A PubChem fingerprint-based random forest classification model from an oversampling approach was built to predict PARP-1 activity. Gini index

(A) PARP1pred app

PARP1pred allow users to predict whether a query molecule is active/inactive towards the PARP1 target protein.

Main About What is PARP1? Dataset Model performance Python libraries Citing us

Predict PARP1 inhibitory activity

Enter SMILES notation

Example SMILES

Submit

(B) PARP1pred app

PARP1pred allow users to predict whether a query molecule is active/inactive towards the PARP1 target protein.

Main About What is PARP1? Dataset Model performance Python libraries Citing us

Predict PARP1 inhibitory activity

Enter SMILES notation

O=C(c1cc(Cc2n[nH]c(=O)c3ccccc23)ccc1F)N1CCN(C(=O)C2CC2)CC1

Example SMILES

O=C(c1cc(Cc2n[nH]c(=O)c3ccccc23)ccc1F)N1CCN(C(=O)C2CC2)CC1

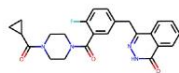
Submit

Input molecule:

Show SMILES

O=C(c1cc(Cc2n[nH]c(=O)c3ccccc23)ccc1F)N1CCN(C(=O)C2CC2)CC1

Show chemical structures



Descriptors

Show full set of descriptors as calculated for query molecule

Show subset of descriptors as used in trained model

Predictions

Active

Figure 9: Screenshot of the PARP1pred webserver before (A) and after (B) entering the SMILES input. Notice that after submission of the SMILES notation the corresponding molecular fingerprints are computed whereby the trained predictive model is applied to classify the query molecule as active or inactive. In this case, the query molecule is classified to be active.

calculation revealed the important features in the random forest model, which included aromatic/cyclic/heterocyclic moieties and nitrogen-containing fingerprints, and ether/aldehyde/alcohol moieties. Additionally, a detailed examination of the structure–activity relationship revealed that hydrophobic interactions and hydrogen bonding networks with nitrogen-containing scaffolds are critical for developing PARP-1 inhibitors. As a result, this insight provides a framework for data-driven PARP-1 inhibitor design.

Conflict of interests

All authors declare that there are no conflicts of interest.

Acknowledgments

We thank Dr. Patipark Kueanjinda for useful discussion on machine learning. We thank Catherine Perfect, MA (Cantab), from Edanz (www.edanz.com/ac), for editing a draft of this manuscript. This project is funded by the National Research Council of Thailand (NRCT) and Mahidol University (NRCT5-TRG63009-04).

REFERENCES

- Abbasi-Radmoghaddam Z, Riahi S, Gharaghani S, Mohammadi-Khanaposhtanai M. Design of potential anti-tumor PARP-1 inhibitors by QSAR and molecular modeling studies. *Mol Diversity*. 2021;25:263-77.
- Armstrong JF, Faccenda E, Harding SD, Pawson AJ, Southan C, Sharman JL, et al. The IUPHAR/BPS Guide to PHARMACOLOGY in 2020: extending immunopharmacology content and introducing the IUPHAR/MMV Guide to MALARIA PHARMACOLOGY. *Nucleic Acids Res*. 2020;48(D1):D1006-21.
- Balasubramaniam S, Beaver JA, Horton S, Fernandes LL, Tang S, Horne HN, et al. fda approval summary: Rucaparib for the treatment of patients with deleterious *BRCA* mutation-associated advanced ovarian cancer. *Clin Cancer Res*. 2017;23:7165-70.
- Banasik M, Komura H, Shimoyama M, Ueda K. Specific inhibitors of poly(ADP-ribose) synthetase and mono(ADP-ribosyl)transferase. *J Biol Chem*. 1992;267:1569-75.
- Baudino TA. Targeted cancer therapy: the next generation of cancer treatment. *Curr Drug Discov Technol*. 2015;12:3-20.
- Beck C, Robert I, Reina-San-Martin B, Schreiber V, Dantzer F. Poly(ADP-ribose) polymerases in double-strand break repair: focus on PARP1, PARP2 and PARP3. *Exp Cell Res*. 2014;329:18-25.
- Breiman L. Random forests. *Machine Learning*. 2001;45(1):5-32.
- Brown JS, O'Carrigan B, Jackson SP, Yap TA. Targeting DNA repair in cancer: beyond PARP inhibitors. *Cancer Discov*. 2017;7:20-37.
- Calle ML, Urrea V. Letter to the Editor: Stability of Random Forest importance measures. *Brief Bioinform*. 2010;12(1):86-9.
- Carracedo-Reboredo P, Liñares-Blanco J, Rodríguez-Fernández N, Cedrón F, Novoa FJ, Carballal A, et al. A review on machine learning approaches and trends in drug discovery. *Comput Struct Biotechnol J*. 2021;19:4538-58.
- Cheung AK, Chin DN, Fan J, Miller-Moslin KM, Shultz MD, Smith TD, et al., inventors. 2-piperidin-1-yl-acetamide compounds for use as tankyrase inhibitors. US patent, US9181266B2. 2015 Nov 10.
- Cortes-Ciriano I, Bender A, Malliavin T. Prediction of PARP inhibition with proteochemometric modelling and conformal prediction. *Mol Inform*. 2015;34:357-66.
- Davies M, Nowotka M, Papadatos G, Dedman N, Gaulton A, Atkinson F, et al. ChEMBL web services: streamlining access to drug discovery data and utilities. *Nucleic Acids Res*. 2015;43(W1):W612-20.
- de Bono J, Mateo J, Fizazi K, Saad F, Shore N, Sandhu S, et al. Olaparib for metastatic castration-resistant prostate cancer. *N Engl J Med*. 2020;382:2091-102.
- Dunn D, Husten J, Ator MA, Chatterjee S. Novel poly(ADP-ribose) polymerase-1 inhibitors. *Bioorg Med Chem Lett*. 2012;22:222-24.
- Farrés J, Martín-Caballero J, Martínez C, Lozano JJ, Llacuna L, Ampurdanés C, et al. Parp-2 is required to maintain hematopoiesis following sublethal γ -irradiation in mice. *Blood*. 2013;122:44-54.
- Ferraris DV. Evolution of poly(ADP-ribose) polymerase-1 (PARP-1) inhibitors. From concept to clinic. *J Med Chem*. 2010;53:4561-84.
- Ferraris D, Ficco RP, Pahutski T, Lautar S, Huang S, Zhang J, et al. Design and synthesis of poly(ADP-ribose)polymerase-1 (PARP-1) inhibitors. Part 3: In vitro evaluation of 1,3,4,5-Tetrahydro-benzo[c][1,6]- and [c][1,7]-naphthyridin-6-ones. *Bioorg Med Chem Lett*. 2003a;13:2513-18.

- Ferraris D, Ko Y-S, Pahutski T, Ficco RP, Serdyuk L, Alemu C, et al. Design and synthesis of poly ADP-ribose polymerase-1 inhibitors. 2. Biological evaluation of Aza-5[H]-phenanthridin-6-ones as potent, aqueous-soluble compounds for the treatment of ischemic injuries. *J Med Chem.* 2003b;46:3138-51.
- Fong PC, Boss DS, Yap TA, Tutt A, Wu P, Mergui-Roelvink M, et al. Inhibition of poly(ADP-ribose) polymerase in tumors from BRCA mutation carriers. *N Engl J Med.* 2009;361:123-34.
- Fundytus A, Sengar M, Lombe D, Hopman W, Jalink M, Gyawali B, et al. Access to cancer medicines deemed essential by oncologists in 82 countries: an international, cross-sectional survey. *Lancet Oncol.* 2021;22:1367-77.
- Gilson MK, Liu T, Baitaluk M, Nicola G, Hwang L, Chong J. BindingDB in 2015: A public database for medicinal chemistry, computational chemistry and systems pharmacology. *Nucleic Acids Res.* 2016;44:D1045-53.
- Golan T, Hammel P, Reni M, Van Cutsem E, Macarulla T, Hall MJ, et al. Maintenance olaparib for germline BRCA-mutated metastatic pancreatic cancer. *N Engl J Med.* 2019;381:317-27.
- Gupte R, Liu Z, Kraus WL. PARPs and ADP-ribosylation: recent advances linking molecular functions to biological outcomes. *Genes Dev.* 2017;31:101-26.
- Halder AK, Saha A, Saha KD, Jha T. Stepwise development of structure-activity relationship of diverse PARP-1 inhibitors through comparative and validated in silico modeling techniques and molecular dynamics simulation. *J Biomol Struct Dyn.* 2015;33:1756-79.
- Helleday T, Petermann E, Lundin C, Hodgson B, Sharma RA. DNA repair pathways as targets for cancer therapy. *Nat Rev Cancer.* 2008;8:193-204.
- Hoy SM. Talazoparib: first global approval. *Drugs.* 2018;78:1939-46.
- Ishida J, Yamamoto H, Kido Y, Kamijo K, Murano K, Miyake H, et al. Discovery of potent and selective PARP-1 and PARP-2 inhibitors: SBDD analysis via a combination of X-ray structural study and homology modeling. *Bioorg Med Chem.* 2006;14:1378-90.
- Ji J, Guo N, Xue T, Kang B, Ye X, Chen X, et al., inventors: Poly (ADP-ribose) polymerase inhibitor. US patent, US9187430B2, 2015 Nov 17.
- Johannes JW, Balazs A, Barratt D, Bista M, Chuba MD, Cosulich S, et al. Discovery of 5-{4-[(7-Ethyl-6-oxo-5,6-dihydro-1,5-naphthyridin-3-yl)methyl]piperazin-1-yl}-N-methylpyridine-2-carboxamide (AZD5305): A PARP1–DNA trapper with high selectivity for PARP1 over PARP2 and other PARPs. *J Med Chem.* 2021;64:14498-512.
- Kanan T, Kanan D, Al Shardoub EJ, Durdagi S. Transcription factor NF- κ B as target for SARS-CoV-2 drug discovery efforts using inflammation-based QSAR screening model. *J Mol Graph Model.* 2021;108:107968.
- Kennard RW, Stone LA. Computer aided design of experiments. *Technometrics.* 1969;11:137-48.
- Kim G, Ison G, McKee AE, Zhang H, Tang S, Gwise T, et al. FDA approval summary: Olaparib monotherapy in patients with deleterious germline BRCA-mutated advanced ovarian cancer treated with three or more lines of chemotherapy. *Clin Cancer Res.* 2015;21:4257-61.
- Kim S, Thiessen PA, Bolton EE, Chen J, Fu G, Gindulyte A, et al. PubChem substance and compound databases. *Nucleic Acids Res.* 2016;44:D1202-13.
- Kinoshita T, Nakanishi I, Warizaya M, Iwashita A, Kido Y, Hattori K, et al. Inhibitor-induced structural change of the active site of human poly(ADP-ribose) polymerase. *FEBS Letters.* 2004;556:43-6.
- Ledermann J, Harter P, Gourley C, Friedlander M, Vergote I, Rustin G, et al. Olaparib maintenance therapy in platinum-sensitive relapsed ovarian Cancer. *N Engl J Med.* 2012;366:1382-92.
- Li J, Zhou N, Cai P, Bao J. In silico screening identifies a novel potential PARP1 inhibitor targeting synthetic lethality in cancer treatment. *Int J Mol Sci.* 2016;17(2):258.
- Li N, Bu H, Liu J, Zhu J, Zhou Q, Wang L, et al. An open-label, multicenter, single-arm, phase ii study of fluzoparib in patients with germline BRCA1/2 mutation and platinum-sensitive recurrent ovarian cancer. *Clin Cancer Res.* 2021;27:2452-8.
- Lipinski CA, Lombardo F, Dominy BW, Feeney PJ. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv Drug Deliv Rev.* 2001;46:3-26.
- Malik AA, Phanus-Umporn C, Schaduengrat N, Shoombuatong W, Isarankura-Na-Ayudhya C, Nantasenamat C. HCVpred: A web server for predicting the bioactivity of hepatitis C virus NS5B inhibitors. *J Comput Chem.* 2020;41:1820-34.

- Malyuchenko NV, Kotova EY, Kulaeva OI, Kirpichnikov MP, Studitskiy VM. PARP1 inhibitors: antitumor drug design. *Acta Naturae*. 2015;7(3):27-37.
- Mateo J, Carreira S, Sandhu S, Miranda S, Mossop H, Perez-Lopez R, et al. DNA-repair defects and olaparib in metastatic prostate cancer. *N Engl J Med*. 2015;373:1697-708.
- Mateo J, Lord CJ, Serra V, Tutt A, Balmana J, Castroviejo-Bermejo M, et al. A decade of clinical development of PARP inhibitors in perspective. *Ann Oncol*. 2019;30:1437-47.
- 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(D1):D930-40.
- Menear KA, Adcock C, Boulter R, Cockcroft X-I, Copsey L, Cranston A, et al. 4-[3-(4-Cyclopropanecarbonylpiperazine-1-carbonyl)-4-fluorobenzyl]-2H-phthalazin-1-one: A novel bioavailable inhibitor of poly(ADP-ribose) polymerase-1. *J Med Chem*. 2008;51:6581-91.
- Ménissier de Murcia J, Ricoul M, Tartier L, Niedergang C, Huber A, Dantzer F, et al. Functional interaction between PARP-1 and PARP-2 in chromosome stability and embryonic development in mouse. *EMBO J*. 2003;22:2255-63.
- Mirza MR, Monk BJ, Herrstedt J, Oza AM, Mahner S, Redondo A, et al. Niraparib maintenance therapy in platinum-sensitive, recurrent ovarian cancer. *N Engl J Med*. 2016;375:2154-64.
- Moree WJ, Goldman P, Demaggio AJ, Christenson E, Herendeen D, Eksterowicz J, et al. Identification of ring-fused pyrazolo pyridin-2-ones as novel poly(ADP-ribose)polymerase-1 inhibitors. *Bioorg Med Chem Lett*. 2008;18:5126-29.
- Myung-Hwa K, Seung-Hyun K, Sae-Kwang K, Chun-Ho P, Bo-Young J, Kwang-Woo C et al., inventors. Tricyclic derivative or pharmaceutically acceptable salts thereof, preparation method thereof, and pharmaceutical composition containing the same. US patent, US8815891B2. 2014 Aug 26.
- Nantasenamat C, Prachayasittikul V. Maximizing computational tools for successful drug discovery. *Exp Opin Drug Discov*. 2015;10:321-9.
- Nantasenamat C, Isarankura-Na-Ayudhya C, Naenna T, Prachayasittikul V. A practical overview of quantitative structure-activity relationship. *EXCLI J*. 2009;8:74-88.
- Nantasenamat C, Worachartcheewan A, Mandi P, Monnor T, Isarankura-Na-Ayudhya C, Prachayasittikul V. QSAR modeling of aromatase inhibition by flavonoids using machine learning approaches. *Chem Papers*. 2014;68:697-713.
- Ocran Mattila P, Ahmad R, Hasan SS, Babar Z-U-D. Availability, affordability, access, and pricing of anti-cancer medicines in low- and middle-income countries: a systematic review of literature. *Front Public Health*. 2021;9:628744.
- OECD. Guidance document on the validation of (Quantitative) Structure-Activity Relationship [(Q)SAR] models. Paris: OECD Publ., 2014.
- Orvieto F, Branca D, Giomini C, Jones P, Koch U, Ontoria JM, et al. Identification of substituted pyrazolo[1,5-a]quinazolin-5(4H)-one as potent poly(ADP-ribose)polymerase-1 (PARP-1) inhibitors. *Bioorg Med Chem Lett*. 2009;19:4196-200.
- Park C-H, Chun K, Joe B-Y, Park J-S, Kim Y-C, Choi J-S, et al. Synthesis and evaluation of tricyclic derivatives containing a non-aromatic amide as inhibitors of poly(ADP-ribose)polymerase-1 (PARP-1). *Bioorg Med Chem Lett*. 2010;20:2250-3.
- Patel MR, Pandya KG, Lau-Cam CA, Singh S, Pino MA, Billack B, et al. Design and synthesis of N-substituted indazole-3-carboxamides as poly(ADP-ribose)polymerase-1 (PARP-1) inhibitors(†). *Chem Biol Drug Des*. 2012;79:488-96.
- Patel MR, Bhatt A, Steffen JD, Chergui A, Murai J, Pommier Y, et al. Discovery and structure-activity relationship of novel 2,3-dihydrobenzofuran-7-carboxamide and 2,3-dihydrobenzofuran-3(2h)-one-7-carboxamide derivatives as poly(ADP-ribose)polymerase-1 inhibitors. *J Med Chem*. 2014;57:5579-601.
- Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: machine learning in Python. *J Mach Learn Res*. 2011;12:2825-30.
- Pescatore G, Branca D, Fiore F, Kinzel O, Bufl LL, Muraglia E, et al. Identification and SAR of novel pyrrolo[1,2-a]pyrazin-1(2H)-one derivatives as inhibitors of poly(ADP-ribose) polymerase-1 (PARP-1). *Bioorg Med Chem Lett*. 2010;20:1094-9.
- Revathi P, Kanth SS, Gururaj S, Chander OS, Rajender PS. Understanding structural characteristics of PARP-1 inhibitors through combined 3D-QSAR and molecular docking studies and discovery of new inhibitors by multistage virtual screening. *Struct Chem*. 2021;32:2035-50.

- Rhee HK, Lim SY, Jung MJ, Kwon Y, Kim MH, Choo HY. Synthesis of isoquinolinone-based tetracycles as poly (ADP-ribose) polymerase-1 (PARP-1) inhibitors. *Bioorg Med Chem*. 2009;17:7537-41.
- Ruf A, Rolli V, de Murcia G, Schulz GE. The mechanism of the elongation and branching reaction of Poly(ADP-ribose) polymerase as derived from crystal structures and mutagenesis¹¹ Edited by R. Huber. *J Mol Biol*. 1998;278:57-65.
- Sahin K, Durdagi S. Identifying new piperazine-based PARP1 inhibitors using text mining and integrated molecular modeling approaches. *J Biomol Struct Dyn*. 2021;39:681-90.
- Scarpelli R, Boueres JK, Cerretani M, Ferrigno F, Ontoria JM, Rowley M, et al. Synthesis and biological evaluation of substituted 2-phenyl-2H-indazole-7-carboxamides as potent poly(ADP-ribose) polymerase (PARP) inhibitors. *Bioorg Med Chem Lett*. 2010;20:488-92.
- Schaduangrat N, Malik AA, Nantasenamat C. ERpred: a web server for the prediction of subtype-specific estrogen receptor antagonists. *PeerJ*. 2021;9:e11716.
- Scott LJ. Niraparib: First global approval. *Drugs*. 2017;77:1029-34.
- Shibata A, Jeggo PA. DNA double-strand break repair in a cellular context. *Clin Oncol (R Coll Radiol)*. 2014;26:243-9.
- Shultz MD, Cheung AK, Kirby CA, Firestone B, Fan J, Chen CH, et al. Identification of NVP-TNKS656: the use of structure-efficiency relationships to generate a highly potent, selective, and orally active tankyrase inhibitor. *J Med Chem*. 2013;56:6495-511.
- Srivastava M, Raghavan SC. DNA double-strand break repair inhibitors as cancer therapeutics. *Chem Biol*. 2015;22:17-29.
- Steffen JD, Brody JR, Armen RS, Pascal JM. Structural implications for selective targeting of PARPs. *Front Oncol*. 2013;3:301.
- Steinhagen H, Gerisch M, Mittendorf J, Schlemmer K-H, Albrecht B. Substituted uracil derivatives as potent inhibitors of poly(ADP-ribose)polymerase-1 (PARP-1). *Bioorg Med Chem Lett*. 2002;12:3187-90.
- Torrisi C, Bisbocci M, Ingenito R, Ontoria JM, Rowley M, Schultz-Fademrecht C, et al. Discovery and SAR of novel, potent and selective hexahydrobenzonaphthyridinone inhibitors of poly(ADP-ribose)polymerase-1 (PARP-1). *Bioorg Med Chem Lett*. 2010;20:448-52.
- Tutt ANJ, Garber JE, Kaufman B, Viale G, Fumagalli D, Rastogi P, et al. Adjuvant olaparib for patients with BRCA1- or BRCA2-mutated breast cancer. *N Engl J Med*. 2021;384:2394-405.
- van Beek L, McClay É, Patel S, Schimpl M, Spagnolo L, Maia de Oliveira T. PARP power: A structural perspective on PARP1, PARP2, and PARP3 in DNA damage repair and nucleosome remodelling. *Int J Mol Sci*. 2021;22(10):5112.
- van de Waterbeemd H. Physicochemical approaches to drug absorption. In: van de Waterbeemd H, Testa B (eds): *Drug bioavailability: estimation of solubility, permeability, absorption and bioavailability*, Vol. 40. 2nd ed. (pp 69-99). New York: Wiley, 2008.
- Wildman SA, Crippen GM. Prediction of physicochemical parameters by atomic contributions. *J Chem Inf Comput Sci*. 1999;39:868-73.
- Worachartcheewan A, Nantasenamat C, Isarankura-Na-Ayudhya C, Prachayasittikul V. QSAR study of H1N1 neuraminidase inhibitors from influenza A virus. *Lett Drug Des Discov*. 2014;11:420-7.
- Xu B, Yin Y, Dong M, Song Y, Li W, Huang X, et al. Pamiparib dose escalation in Chinese patients with non-mucinous high-grade ovarian cancer or advanced triple-negative breast cancer. *Cancer Med*. 2021;10(1):109-18.
- Xu W, Delahanty G, Zhang J, inventors. Diazabenz[de] anthracen-3-one compounds and methods for inhibiting PARP. US Patent, US8470825B2. 2014a Nov 11.
- Xu W, Delahanty G, Wei L, Zhang J, inventors. PARP inhibitor compounds, compositions and methods of use. US patent, US8894989B2. 2014b Nov 25.
- Yap CW. PaDEL-descriptor: an open source software to calculate molecular descriptors and fingerprints. *J Comput Chem*. 2011;32:1466-74.
- Yu J, Luo L, Hu T, Cui Y, Sun X, Gou W, et al. Structure-based design, synthesis, and evaluation of inhibitors with high selectivity for PARP-1 over PARP-2. *Eur J Med Chem*. 2022;227:113898.
- Zhao H, Ji M, Cui G, Zhou J, Lai F, Chen X, et al. Discovery of novel quinazoline-2,4(1H,3H)-dione derivatives as potent PARP-2 selective inhibitors. *Bioorg Med Chem*. 2017;25:4045-54.
- Zhou C, Ren B, Wang H, inventors. Fused tetra or penta-cyclic dihydrodiazepinocarbazolones as parp inhibitors. US patent, US9617273B2. 2017 Apr 11.

Zhou J, Ji M, Wang X, Zhao H, Cao R, Jin J, et al. Discovery of quinazoline-2,4(1H,3H)-dione derivatives containing 3-substituted piperazines as potent PARP-1/2 inhibitors—design, synthesis, in vivo antitumor activity, and X-ray crystal structure analysis. *J Med Chem.* 2021;64:16711-30.

Zmuda F, Malviya G, Blair A, Boyd M, Chalmers AJ, Sutherland A, et al. Synthesis and evaluation of a radioiodinated tracer with specificity for poly(ADP-ribose) polymerase-1 (PARP-1) in vivo. *J Med Chem.* 2015; 58:8683-93.