Ai-Based QSAR Approach for Predicting Cathepsin L Inhibition

Shaheen Begum*[1]& M. Siva Parvathi[2]
[1]Institute of Pharmaceutical Technology, Sri Padmavati Mahila Visvavidyalayam, Tirupati-517502
[2]Dept. of Applied Mathematics, Sri Padmavati Mahila Visvavidyalayam, Tirupati-517502
Email. Shaheen.pharmchem@gmail.com

Abstract

Cathepsin L (CatL), belongs to cysteine protease class of enzymes that is primarily involved in proteolytic process in lysozymes. The enzyme levels are abnormally enhanced in several disease conditions. Recent studies have indicated Cathepsin L as a promising target to treat COVID-19. The entry of causative virus SARS-Co-V-2 into the host cell can be decreased by CatL inhibition. Several naturally occurring agents such as triperpenoids and Gallinamides inhibit CatL enzyme. Statistically significant and robust QSAR models establish correlation between molecular properties and biological activity of a set of chemical compounds. The present work deals with the development of a QSAR) model(s) for identifying relationship between the structural features and the biological activity of Cat L inhibitors. Classification-based (LDA) and regression-based (MLR) QSAR models were generated using different AI tools. [MLR-based QSAR model: Internal Validation Parameters: R^2 = 0.683, R^2 (adjusted)= 0.663, PRESS =119.64 Cross-validation results (Leave-One-Out) : Q^2 :0.637 External Validation Parameters: r^2 :0.604, Q^2f1/R^2(Pred) :0.61, Q^2f2 :0.601, RMSEP:0.839]. The LDA model also passed all the qualitative validation metrics computed while performing internal and external validation using the training and test sets, respectively. The derived QSAR models are significant for predicting the activity of newly designed CatL inhibitors.

Keywords: Cathepsin L, QSAR, PaDEL, SVM technique, LDA technique, MLR technique, PLS method

INTRODUCTION

Cathepsin L or CatL (EC 3.4.22.15) enzyme is primarily involved in proteolytic process in lysozymes. The enzyme levels are abnormally enhanced in several disease conditions such as diabetes, liver fibrosis, immune disorders, inflammation and invasive cancer pathogenesis. Due to the therapeutic importance of enzyme inhibition of this enzyme, a variety of irreversible and reversible inhibitors were developed in the recent years. A total of eleven Cathepsins (B, H, L, S, C, K, O, F, V, X &W) have been reported so far and they contain a cysteine and histidine pair at the active site (thiolate-imidazolium pair) in common [1]. Amongst all, CatB and L are extensively studied in the process of development of anticancer agents. Recently studies have indicated CatL as a promising target to treat COVID-19. The entry of causative virus SARS-Co-V-2 into the host cell can be decreased by CatL inhibition [2-4].

CatL can be inhibited by chemically and structurally diverse compounds such as epoxysuccinates, inhibitors having peptide scaffold (peptidyl diazomethane, peptidyl aldehyde, peptidyl hydroxylamine, peptidyl aziridines, and peptidyl methanes), azapanones, Gallinamide A derivatives, nitriles, thiosemicarbazone etc. Most of the inhibitors depend on

the potency of the warhead group that can react with the active site of the enzyme; thiolate part. Due to the high and unspecific chemical reactivity with other biological molecules, majority of these inhibitors have limited applicability as drug candidates. Hence research has been much focused on covalent-reversible inhibitors with specific activity against Cathepsin L [5-18].

QSAR approaches contribute to ligand-based drug discovery and development process. Statistically significant and robust QSAR models establish correlation between molecular properties and biological activity of a set of chemical compounds. Machine learning (ML) techniques have tremendous applicability in drug discovery projects. ML techniques provide reliable QSAR models for investigating relationship between chemical molecules and their biological effects, toxicity and off-target interactions [19-21]. In this work, all the recommended guidelines and procedures werefollowed essential for building a robust (properly validated) QSAR model.

METHODOLOGY

Dataset Collection and Curation: The compounds with known experimental activity (IC_{50} values) against Cathepsin L were collected from several published articles [**5-17**]. After data collection, chemical curation was performed to confirm no structural errors in the chemicals. Chemical curation also includes removing inorganic/organometallic chemicals (**six chemicals comprising Si element were removed**), normalizing the structures and finally, optimizing the structures using MMFF94S force field. Subsequently, biological curation was performed to remove chemicals with ambiguous biological data, as well as duplicate analysis and activity cliff analyses were performed.

The chemical curation was performed using freely available Konstanz Information Miner (KNIME) workflow (https://www.knime.org/) while biological curation was performed using the small data set modeler software [22].

Descriptor Calculation and Data Pre-processing: In the next step, thousands of descriptors (structural features/identifiers) and fingerprints were calculated using the PaDEL-Descriptor software [23]. The total descriptors pool included several diverse classes of descriptors (2D, 3D, and Fingerprints) structural, topological, functional group counts, extended topochemical atom (ETA) indices, fingerprints (PubChem, MACCS, 2D atom pair), etc. After calculating the descriptors, we performed the data pre-processing to remove all the non-informative descriptors, where the non-informative descriptors include constant (variance < 0.0001) and intercorrelated (correlation coefficient > 0.99) descriptors.

Data set Division: Next, we have divided the data into training (70-80% of the whole data) and test (20-30%) sets using the random or rational (Euclidean distance-based, Kennard-Stone algorithm, etc.) approach. Here, the training set was involved in model building and selection, while the test set was exclusively used for the external validation of the selected models.

Model Development: In this step, the training set was employed to develop the QSAR models. Here, several feature selection techniques such as Genetic algorithm, Stepwise selection (forward selection and backward elimination), and best subset selection, were employed to identify the best pool of descriptors with a significant relationship with the biological activity of our interest. Further, several machine learning (ML) techniques both

linear and non-linear were utilized for developing multiple QSAR models and then compared among themselves (based on the predictive ability, simplicity in understanding, and ease in deployment, etc.) to find the best QSAR model.

Machine learning techniques employed:

Linear techniques: Multiple Linear Regression (MLR), Partial Least Square (PLS), Linear Discriminant Analysis (LDA)

Non-linear techniques: Random Forest (RF), Support Vector Machines (SVM)

Model Validation: The statistical quality and prediction quality of developed models were checked by performing both internal and external validation. The internal validation of regression models involves the calculation of several validation metrics such as determination coefficient (R^2), adjusted R^2(R_a^2), leave-one-out (*LOO*) cross-validated determination coefficient (Q^2), scaled r_m^2 *metrics*, *etc*. The internal validation metrics were utilized to evaluate the internal prediction quality as well as to select the best model among all built QSAR models. While the external validation was employed for validating the selected best model and it includes several validation metrics such as $Q^2_{ext(F1)}$ or R^2_{pred}, $Q^2_{ext(F2)}$, concordance correlation coefficient (CCC), scaled r_m^2 *metrics*, *etc*. Further, the Y-randomization test was performed to confirm whether the model is developed by chance or not. In the Y-randomization test, we have developed 50 random models by shuffling the response columns only, and then compared the R^2 and Q^2 values of random models with the original model. Also, response and residual plots were generated and reported to confirm the quality of predictions (based on the response plot) and to check the presence of systematic errors (based on the residual plot), respectively. For classification models, qualitative validation metrics, Mathews correlation coefficient, F-measure, etc. computed using a confusion matrix were employed for the internal and external validation. Here, receiver operating characteristics (ROC) plots were generated and reported to confirm the discriminatory power of the developed model.

Applicability Domain Determination

In the final step, the domain of applicability of the selected best model was defined using the standardization (for regression and classification model) and/or confidence-based (for classification model) approach [24]**.**

Notably, all the QSAR modeling steps (starting from data pre-processing to model development and validation) were performed using DTC-QSAR v1.0.5 software [25]. MLR calculations were also performed using MLRplusValidation tool v1.3 [25].

RESULTS AND DISCUSSION

Initially, the collected data set comprised 375 compounds with known experimental activity (IC_{50}) against Cathepsin L. After data curation (both chemical and biological data), the final curated data set consisted of 347 compounds. For your information, the chemicals with ID 10, 49, 122 were removed as their activity values were found to be ambiguous and the chemical with ID 321 was removed later during the descriptor calculation step as PaDEL-Descriptor software was unable to calculate some descriptors for this chemical (especially 3D descriptors). Other chemicals removed were either found to be duplicates/with structure errors or comprise of Si element (for more details see ***ActivityAndCurationDetails.xlsx*** in the supplementary material). All the activity values against Cathepsin L reported in IC_{50} (nM,

μM, or M) were first converted into IC_{50} (M) and then converted into pIC_{50} (negative logarithm of IC_{50}) for QSAR modeling. Further for classification modeling, the biological activity data were categorized into two classes: **active** and **inactive** using a threshold value (i.e., active ≤ 1000 nM and inactive > 1000 nM; see ***CuratedDataSet.xlsx***).

Descriptor Calculation and Data Preprocessing:

A total of **14,816** structural features including 2D & 3D descriptors as well as fingerprints, were calculated using PaDel-Descriptor software. After data pre-processing (i.e., removing constant and inter-correlated descriptors), the final set comprises 2708 structural features. All the information about the calculated descriptors is available in the supporting material.

Data set Division

Though both random and rational approaches were employed for dividing the data set into training and test sets. However, better models were found when rational approaches were used. Note that rational approaches ensure proper distribution of chemicals into training and test sets, such that the test set is representative of the training set, as well as, there is no loss of information from the training set. In the present study, for the classification model, we have employed the Kennard-stone algorithm, while for the regression model the division was performed using the Euclidean distance-based division.

Model development and Validation results

After the data set division, the training set was employed for model development, while the test set was kept aside for model validation. Here, we have provided all the resultant information obtained from the performed QSAR study. This study resulted in two QSAR models, i.e., a classification-based LDA model and a regression-based MLR model. Both the models were developed using linear techniques and are easy to interpret.

Classification-based QSAR model

In this study, we have utilized a genetic algorithm technique for feature selection (available in DTC-QSAR software), which helped in identifying top features with a significant relationship with our activity of interest. Further, using the best pool of descriptors, the classification model was finally developed using both LDA and the random forest technique. However, the best model was found to be LDA with 6 descriptors only. The model descriptors, their standard coefficient values with a brief description are reported below.The standardized coefficient values indicate the contribution of each descriptor towards the inactivity class of the chemicals. For instance, the SIC0_2D with a negative coefficient value (-1.97023) is negatively contributing towards the inactive class and thus contributing positively towards the active class9 [Table1].

Table1. Details of the derived LDA model

LDA Model Descriptors	Standardized Coefficients	Description
SIC0_2D	-1.97023	Structural information content index (neighborhood symmetry of 0-order)
APC2D9_C_Br	0.50051	Count of atom pair (C-Br) at topological distance 9
AATS3v_2D	0.58450	Average Broto-Moreau autocorrelation - lag 3 / weighted by van der Waals volumes
SubFP200	0.59736	Presence of Sulfon (fingerprint descriptor)
APC2D8_C_C	-1.58526	Count of atom pair C-C at topological distance 8
AATSC3v_2D	0.58366	Average centered Broto-Moreau autocorrelation - lag 3 / weighted by van der Waals volumes

Validation results for classification-based (LDA) model: The LDA model also passed all the qualitative validation metrics computed while performing internal and external validation using the training and test sets, respectively. The validation results are shown below [Table 2].

Table 2. Details of the validation parameters for LDA method

Qualitative validation metrics	Training set (internal validation)	Test set (external validation)
True Positive	77	37
False Positive	13	6
True Negative	135	52
False Negative	18	9
Accuracy (%)	87.24	85.58
Precision (%)	85.56	86.05
Sensitivity (%)	81.05	80.43
Specificity (%)	91.22	89.65
F1-score	0.832	0.832
MCC	0.730	0.707
Area under ROC curve	0.891	0.881

Receiver operation characteristics (ROC Plot): ROC plot (figure 1) helps us to understand the discriminatory power of the developed LDA model. The area under the curve (AUC) should be higher than 0.50 to consider the developed model better than a random classifier. In our case, the AUC for training (0.8913) and test set (0.8812) were found to be significantly greater than 0.5, suggesting a good discriminatory power of the model.

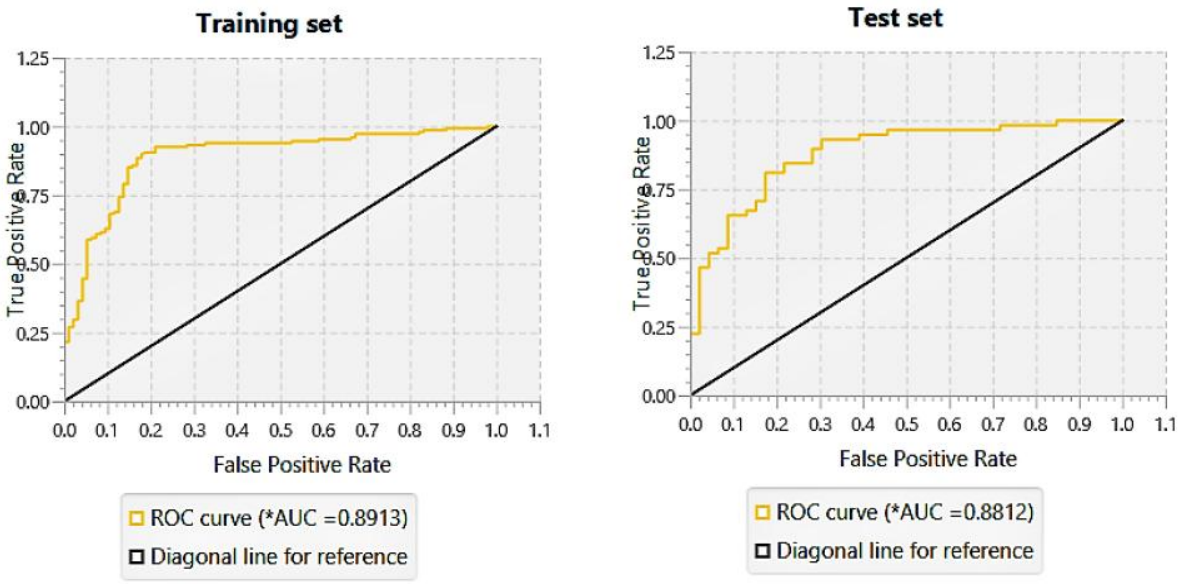

Figure.1 ROC plot of the derived LDA model

Regression-based QSAR model

For regression modeling, we have utilized both genetic algorithm and best subset selection techniques (available in DTC-QSAR software) for identifying top features with a significant relationship with our activity of interest. Finally, the regression model was finally developed using both MLR and SVM (using Weka software) techniques. However, the best model was found to be an MLR model with 14 descriptors. The following is the MLR model equation showing all 14 descriptors with their coefficient values:

pIC50 = 0.65051 + 0.79296×AD2D314 + 1.19877×KRFP4520 - 1.31101×MACCSFP38 + 1.12428× SpMax4_Bhm_2D + 1.18615×PubchemFP663 - 4.514×MATS3v_2D - 0.92655×AD2D636 + 1.09507×PubchemFP536 + 11.77526×VCH-4_2D + 0.94022×GATS4s_2D - 1.61651×KRFP1300 - 0.00662×TDB1m_3D - 0.32124×APC2D10_C_Br + 0.38072×nX_2D

Table 3 :The descriptors and their brief description

AD2D314 *=> 'Positive Contribution' => Presence of C-N at topological distance 5*
KRFP4520 *=> 'Positive Contribution' => Presence of this chemical substructure: O=CCCc1c[nH]c2ccccc12*
MACCSFP38 *=> 'Negative Contribution' => Presence of NC(C)N substructure*
SpMax4_Bhm_2D *=> 'Positive Contribution' => (BurdenModifiedEigenvaluesDescriptor) Largest absolute eigenvalue of Burden modified matrix - n 4 / weighted by relative mass*

PubchemFP663 *=> 'Positive Contribution' => Presence of substructure (O-C-C-O-[#1])*
MATS3v_2D *=> 'Negative Contribution' => Moran autocorrelation - lag 3 / weighted by van der Waals volumes*
AD2D636 *=> 'Negative Contribution' => Presence of C-X at topological distance 9*
PubchemFP536 *=> 'Positive Contribution' => Presence of substructure (O=CCN)*
VCH-4_2D *=> 'Positive Contribution' => (ChiChain descriptor) Valence chain, order 4*
GATS4s_2D *=> 'Positive Contribution' Geary autocorrelation - lag 4 / weighted by I-state*
KRFP1300 *=> 'Negative Contribution' =>Presene of this substructure [!#1]C(=O)[NH][NH]C(=O)[!#1]*
TDB1m_3D *=> 'Negative Contribution' => (Autocorrelation 3D descriptor) 3D topological distance based autocorrelation - lag 1 / weighted by mass*
APC2D10_C_Br *=> 'Negative Contribution' => Count of atom pair C-Br at topological distance 10*
nX_2D *=> 'Positive Contribution' => Number of halogen atoms (F, Cl, Br, I, At, Uus)*

Validation results for regression-based (MLR) model: All the important internal and external validation metrics were showing statistically acceptable values (shown below).

Internal Validation Parameters:

R^2 = 0.683, R^2 (adjusted)= 0.663, *PRESS* =119.64

Cross-validation results (Leave-One-Out) : Q^2 :0.637

External Validation Parameters:

r^2 :0.604, $Q^2_{f1}/R_{2(Pred)}$:0.61, Q^2_{f2} :0.601, *RMSEP*:0.839

Based on the validation metrics (internal and external), the prediction quality of selected MLR model was found to be moderate. More details about the validation metrics are available in the supplementary material.

Response Plot (for both training and test set): The response plot (Y-observed versus Y-predicted values) was plotted to observe the quality of predictions in both training and test sets. As observed in the Figure 2 below, most of the values are near the diagonal line confirming that the predictions are considerably good.

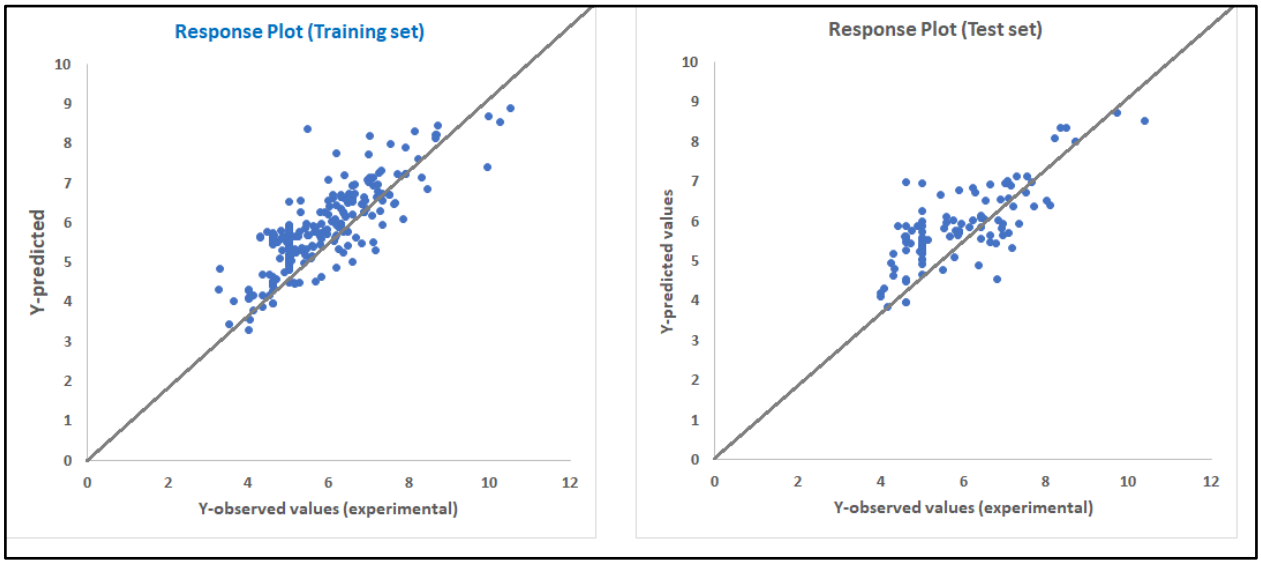

Figure 2 . The response plot obtained for derived MLR model

Residual Plot (for training set): This plot (residual versus Y-observed values of the training set) assists us to identify the presence of any systematic error if any. Ideally, the points should be distributed randomly above and below the horizontal line passing through zero, which is also observed in the residual plot obtained for the developed MLR model (see below).

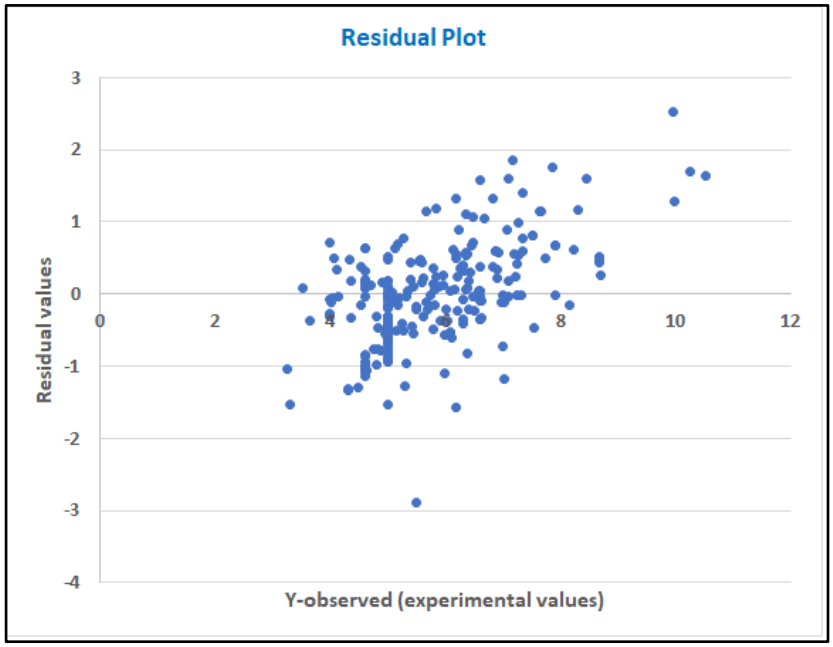

Figure 3. The residual plot of the derived MLR method

Y-randomization test: The selected MLR model passed the Y-randomization test, which confirms that the model was not developed by chance. The R^2 (0.0716) and Q^2 (-0.059) values of random models were found inferior when compared with the R^2 (0.6824) and Q^2 (0.637) values of the original MLR model.

Applicability domain results: For the regression model, the applicability domain was determined using the standardization approach, while for the classification model the applicability domain was determined using the standardization method as well as using the confidence estimation approach. All the relevant supporting information (including input and output files) are provided as supplementary material.

CONCLUSION

Cathepsin L is an emerging biological target for several diseases including COVID-19. Literature indicated that this enzyme can be inhibited by varied chemical structures. The present work involving both classification-based and regression-based QSAR model development provided suitable set of descriptors that can be significantly correlated with inhibition potency. The developed QSAR models can be utilized for predicting the activity of newly designed compounds.

FUNDING

The work is funded by DST -CURIE-AI Centre, Sri Padmavati Mahila Visvavidyalayam, Tirupati.

REFERENCES

1. Turk, V., Stoka, V., Vasiljeva, O., Renko, M., Sun, T., Turk, B., Turk, D.(2012). Cysteine cathepsins: From structure, function and regulation to new frontiers, *Biochimica et Biophysica Acta (BBA) - Proteins and Proteomics*, 1824(1) , 68-88.
2. Sudhan, D. R., & Siemann, D. W. (2015). Cathepsin L targeting in cancer treatment. Pharmacology & therapeutics, 155, 105–116.
3. Gondi, C. S., & Rao, J. S. (2013). Cathepsin B as a cancer target. *Expert opinion on therapeutic targets*, 17(3), 281–291.
4. Gomes, C. P., Fernandes, D. E., Casimiro, F., da Mata, G. F., Passos, M. T., Varela, P., Mastroianni-Kirsztajn, G., &Pesquero, J. B. (2020). Cathepsin L in COVID-19: From Pharmacological Evidences to Genetics. *Frontiers in cellular and infection microbiology*, 10, 589505.
5. Chavarria, G. E., Horsman, M. R., Arispe, W. M., Kumar, G. K., Chen, S. E., Strecker, T. E., & Trawick, M. L. (2012). Initial evaluation of the antitumour activity of KGP94, a functionalized benzophenone thiosemicarbazone inhibitor of cathepsin L. *European Journal of Medicinal Chemistry*, 58, 568-572.
6. Falgueyret, J. P., Oballa, R. M., Okamoto, O., Wesolowski, G., Aubin, Y., Rydzewski, R. M., & Percival, M. D. (2001). Novel, nonpeptidic cyanamides as potent and reversible inhibitors of human cathepsins K and L. *Journal of medicinal chemistry*, *44*(1), 94-104.
7. Kumar, G. K., Chavarria, G. E., Charlton-Sevcik, A. K., Arispe, W. M., MacDonough, M. T., Strecker, T. E., Pinney, K. G. (2010). Design, synthesis, and biological evaluation of potent thiosemicarbazone based cathepsin L inhibitors. *Bioorganic &Medicinal chemistry Letters*, 20(4), 1415-1419.
8. Kumar, G. K., Chavarria, G. E., Charlton-Sevcik, A. K., Yoo, G. K., Song, J., Strecker, T. E., Pinney, K. G. (2010). Functionalized benzophenone, thiophene, pyridine, and fluorene thiosemicarbazone derivatives as inhibitors of cathepsin L. *Bioorganic &Medicinal Chemistry Letters*, 20(22), 6610-6615.
9. Liu, Z., C Myers, M., P Shah, P., Pat Beavers, M., A Benedetti, P., L Diamond, S., M Huryn, D. (2010). Design, synthesis and biological evaluation of a library of thiocarbazates and their activity as cysteine protease inhibitors. Combinatorial *chemistry & high throughput screening*, 13(4), 337-351.
10. Marquis, R. W., James, I., Zeng, J., Trout, R. E. L., Thompson, S., Rahman, A., Veber, D. F. (2005). Azepanone-based inhibitors of human cathepsin L. *Journal of medicinal chemistry*, 48(22), 6870-6878.
11. Dana, D., Pathak, S. K. (2020). A review of small molecule inhibitors and functional probes of human cathepsin L. *Molecules*, *25*(3), 698.
12. Myers, M. C., Shah, P. P., Beavers, M. P., Napper, A. D., Diamond, S. L., Smith III, A. B., & Huryn, D. M. (2008). Design, synthesis, and evaluation of inhibitors of cathepsin L: Exploiting a unique thiocarbazate chemotype. *Bioorganic & medicinal chemistry letters*, *18*(12), 3646-3651.

13. Parker, E. N., Song, J., Kumar, G. K., Odutola, S. O., Chavarria, G. E., Charlton-Sevcik, A. K., Pinney, K. G. (2015). Synthesis and biochemical evaluation of benzoylbenzophenone thiosemicarbazone analogues as potent and selective inhibitors of cathepsin L. *Bioorganic & medicinal chemistry*, *23*(21), 6974-6992.
14. Song, J., Jones, L. M., Kumar, G. K., Conner, E. S., Bayeh, L., Chavarria, G. E., Pinney, K. G. (2012). Synthesis and biochemical evaluation of thiochromanone thiosemicarbazone analogues as inhibitors of cathepsin L. *ACS medicinal chemistry letters*, *3*(6), 450-453.
15. Song, J., Jones, L. M., Chavarria, G. E., Charlton-Sevcik, A. K., Jantz, A., Johansen, A., Pinney, K. G. (2013). Small-molecule inhibitors of cathepsin L incorporating functionalized ring-fused molecular frameworks. *Bioorganic & medicinal chemistry letters*, *23*(9), 2801-2807.
16. Altmann, E., Cowan-Jacob, S. W., Missbach, M. (2004). Novel purine nitrile derived inhibitors of the cysteine protease cathepsin K. *Journal of medicinal chemistry*, *47*(24), 5833-5836.
17. Yasuma, T., Oi, S., Choh, N., Nomura, T., Furuyama, N., Nishimura, A., Sohda, T. (1998). Synthesis of peptide aldehyde derivatives as selective inhibitors of human cathepsin L and their inhibitory effect on bone resorption. *Journal of medicinal chemistry*, *41*(22), 4301-4308.
18. Ramalho, S. D., De Sousa, L. R., Nebo, L., Maganhi, S. H., Caracelli, I., Zukerman- Schpector, J., Vieira, P. C. (2014). Triterpenoids as novel natural inhibitors of human cathepsin L. *Chemistry & biodiversity*, *11*(9), 1354-1363.
19. Yu-Chen, Lo., Stefano, E. R., Wen, T., Russ, B. Altman, Machine learning in chemoinformatics and drug discovery, Drug Discovery Today, 23 (8), 2018, 1538-1546.
20. Rim, K, T. In silico prediction of toxicity and its applications for chemicals at work. *Toxicology and Environmental Health Sciences,* 12, 191–202 (2020).
21. Zakharov A. V, Varlamova E.V., Lagunin A.A., Dmitriev A.V., Muratov E.N., Fourches D., Kuz'min V.E., Poroikov V.V., Tropsha A., Nicklaus M.C. (2016). QSAR Modeling and Prediction of Drug-Drug Interactions. *Molecular Pharmaceutics*. 13(2):545-56.
22. Ambure, P., Gajewicz-Skretna, A., Cordeiro, M. N. D. S., Roy, K. (2019). New workflow for QSAR model development from small data sets: small dataset curator
and small dataset modeler. Integration of data curation, exhaustive double cross-validation, and a set of optimal model selection techniques. *The Journal of Chemical Information and Modeling*., 59(10), 4070-4076.
23. Yap, C. H. (2011). PaDEL-descriptor: open-source software to calculate moleculardescriptors and fingerprints. *Journal of Computational Chemistry* 32(7): 1466-74.
24. Roy, K., Kar, S., Ambure, P. (2015) On a simple approach for determining applicability domain of QSAR models, *Chemometrics and Intelligent Laboratory Systems*, 145, 22-29.
25. https://dtclab.webs.com/software-tools