Indian Journal of Pharmaceutical Chemistry and Analytical Techniques (IJPCAT)
ISSN: 3049-3765 | Vol. 01, Issue-04, July-Aug, 2025, pp. 67-77
https://ijpcat.com/1

Development of QSAR Models for Predicting Anticancer Activity of Heterocycles

Soumyajit Panda[1*], Swarna Dabral[1]

[1]Department of Pharmacology, MM College of Pharmacy, MM(DU), Mullana, Ambala, Haryana
Pin: 133207

***Corresponding Author E-mail: soumyajitpanda397@gmail.com**

ABSTRACT

In drug discovery, quantitative structure-activity relationship (QSAR) modelling has become a potent computational method, particularly for the early assessment of heterocyclic compounds' potential for anticancer effects. In order to forecast the anticancer activity of specific heterocycles tested on animal models, this study creates reliable QSAR models using molecular descriptors. The study used a dataset of 60 heterocyclic derivatives that have been shown to have anticancer properties in vivo (in mouse models). Multiple Linear Regression (MLR), Partial Least Squares (PLS), and Support Vector Machine (SVM) techniques were used to construct the models. The models' predictive power was validated by cross-validation and external test set validation. By identifying important structural characteristics that affect anticancer potency, the study opens the door for logical drug design and lessens reliance on animal testing by using virtual screening.

Key Words:

Quantitative Structure-Activity Relationship (QSAR), Predicting, Anticancer Activity, Heterocycle

Article History:

Received on Feb 11, 2025

Revised on May 15, 2025

Accepted on June 29, 2025

Published on August 6, 2025

DOI: https://doi.org/10.64062/IJPCAT.Vol1.Issue4.6

1. INTRODUCTION

By placing the study in the larger context of medicinal chemistry and computational drug discovery, this section lays the groundwork for future research [1]. It highlights the drawbacks of conventional experimental methods, talks about the importance of heterocyclic compounds, and presents QSAR modelling as a cutting-edge way to expedite the identification of anticancer drugs. Lastly, it highlights the primary research problem and the study's primary goals [2].

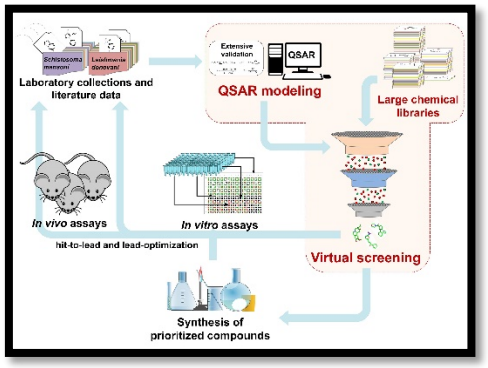

Figure 1: QSAR-Based Virtual Screening

1.1.Background Information

One of the most researched and significant classes of organic molecules in pharmacology is heterocyclic compounds. These structurally varied and biologically powerful compounds are characterised by ring systems that contain at least one heteroatom, usually nitrogen, oxygen, or sulphur. They serve as the molecular basis for a variety of medicinal substances, especially anticancer medications like alkaloids, pyrimidines, and purines [3]. They are essential to the creation of contemporary chemotherapeutic agents because of their capacity to selectively and effectively interact with important biological targets.

Finding novel heterocyclic compounds with anticancer activity in the field of oncology has historically required multi-stage laboratory testing, frequently starting with in vitro assays and moving on to in vivo animal models. To comprehend a compound's pharmacokinetics, toxicity, and efficacy in a biological system, these animal-based studies are essential. Nevertheless, this method is time-consuming, resource-intensive, and presents serious ethical issues with regard to animal welfare. Computational techniques have become increasingly potent tools for early-stage drug screening as the need for quicker, less expensive, and morally sound alternatives increases.

Quantitative Structure–Activity Relationship (QSAR) modelling is one such computational technique that aims to establish a mathematical connection between the chemical structure of a compound and its biological activity. QSAR models predict the biological activity of untested compounds by calculating molecular descriptors, which are numerical values that represent different structural, physicochemical, electronic, and topological properties of molecules [4]. In addition to lessening the need for animal testing in the early stages of screening, these models allow researchers to rank the most promising candidates for additional synthesis and in vivo validation.

1.2.Statement of the Problem

It is still difficult to create reliable and understandable QSAR models for forecasting anticancer activity in heterocyclic compounds, even with advancements in computational drug discovery. The translational relevance of the majority of current models is limited because they are based on in vitro data, which lacks the biological complexity of in vivo systems. Predictive reliability is also decreased by single modelling approaches, small datasets, and inadequate descriptor selection. Model robustness is further limited by the underutilisation of multi-algorithmic

frameworks and biologically validated animal data. To improve accuracy, interpretability, and usefulness in early-stage drug discovery, comprehensive QSAR models that incorporate animal-based data and cutting-edge machine learning techniques are therefore desperately needed.

1.3.Objectives of the Study

1. To collect and curate secondary data of heterocyclic compounds with reported in vivo anticancer activity in animal models (e.g., murine models).
2. To compute and analyze a wide range of molecular descriptors that may influence biological activity.
3. To develop and validate predictive QSAR models using statistical and machine learning techniques, including Multiple Linear Regression (MLR), Partial Least Squares (PLS), and Support Vector Machine (SVM).
4. To identify the most significant molecular features (descriptors) responsible for anticancer potency and evaluate their biological relevance.
5. To enhance the preclinical drug discovery process by reducing dependency on animal trials through reliable computational screening.

2. METHODOLOGY

In order to construct and validate QSAR models, this section outlines the general research design, the type and source of data, the instruments utilised, and the computational methods applied. There are no human or animal trials or field-based experiments in this study; it is solely based on secondary data. When building and validating models, the methodology guarantees reproducibility and scientific rigour.

2.1.Description of Research Design

In order to create predictive models using the data that is already available, the study employs a computational and retrospective design. In particular, the molecular structure of heterocyclic compounds was correlated with their reported anticancer activity in animal models using a QSAR (Quantitative Structure–Activity Relationship) modelling approach [5]. There is no need for any in vivo experimentation because the study is non-experimental in nature and only uses data that has already been published. Both internal and external validation techniques were used to statistically validate the developed models. By decreasing the need for animal testing, this design not only promotes ethical research but also speeds up the drug discovery process by using in silico screening.

2.2.Participants / Sample Details

Since this study is based on secondary data, neither human nor animal subjects were specifically chosen for the study. Chemical compounds—more especially, 60 heterocyclic molecules with documented anticancer activity data from murine models (mouse-based trials)—are referred to as the "sample" in this context. Quantitative efficacy data, such as IC_{50} values (half-maximal inhibitory concentration) or percentage inhibition of tumour growth, were used to select these compounds [6]. To guarantee data consistency and reliability in QSAR

modelling, the inclusion criteria were stringent: only compounds with well-documented in vivo results and well-defined structural information were taken into consideration.

2.3.Instruments and Materials Used

A wide range of computational tools were used in this study at various stages of the QSAR model development process. ChemDraw and MarvinSketch were used to generate precise 2D representations of the heterocyclic compounds for the purpose of standardising and generating molecular structures [7]. PaDEL-Descriptor and Dragon software, which together calculated more than 1,400 molecular descriptors covering physicochemical, topological, geometrical, and electronic properties, were used to perform the descriptor calculation. Regression modelling, exploratory data analysis, and correlation filtering were handled by SPSS, Python, and MATLAB for statistical analyses and data preprocessing. Lastly, the main machine learning environment for putting Support Vector Machine (SVM) algorithms into practice was the KNIME Analytics Platform, which allowed for reliable, non-linear modelling. By combining these resources, a reproducible and rigorously scientific method for developing and validating QSAR models was guaranteed [8].

2.4.Procedure and Data Collection Methods

Using a systematic computational approach, the study started with gathering pertinent secondary data from reliable sources like ChEMBL, PubChem BioAssay, and peer-reviewed scientific publications that were accessible via databases like ScienceDirect, Scopus, and PubMed. Included were only heterocyclic compounds with well-documented anticancer activity in murine models. In order to obtain optimised 3D conformations for precise descriptor calculation, the molecular structures of these compounds were first standardised and drawn using chemical drawing tools [9]. This was followed by energy minimisation using the MMFF94 force field. More than 1,400 molecular descriptors covering a variety of structural and physicochemical characteristics were calculated. Repetitive and highly correlated descriptors ($r > 0.85$) were eliminated to guarantee model efficiency and interpretability, and Principal Component Analysis (PCA) and stepwise multiple regression were used to choose important features. Three methods were used to construct predictive QSAR models from the cleaned dataset: Support Vector Machine (SVM) to capture intricate nonlinear patterns, Partial Least Squares (PLS) to control multicollinearity, and Multiple Linear Regression (MLR) to identify linear relationships. The models were tested externally using a holdout dataset that included 20% of the compounds not used during training, and internally using Leave-One-Out Cross-Validation (LOO-CV) to evaluate robustness [10].

2.5.Data Analysis Techniques

In order to create reliable and understandable QSAR models, statistical and machine learning methods were applied methodically throughout the data analysis process. The IC_{50} values and molecular descriptor ranges were first compiled using descriptive statistics to give a general picture of the distribution of the dataset. Correlation matrices and Variance Inflation Factor (VIF) analyses were used to find and remove highly correlated descriptors in order to address multicollinearity concerns. Multiple Linear Regression (MLR) and Partial Least Squares (PLS) were used for linear modelling in order to create models that were easy to understand and

estimate coefficients. A Support Vector Machine (SVM) with a radial basis function kernel was used to capture intricate, non-linear relationships in the dataset. Standard performance metrics like R2 (coefficient of determination), RMSE (root mean square error), MAE (mean absolute error), and Q2 (predictive squared correlation coefficient) were used to thoroughly assess each model. Throughout the analysis, adherence to the OECD's QSAR model development principles was upheld to guarantee the models' predictive reliability, scientific validity, and transparency.

3. RESULTS

The results of the creation and assessment of QSAR models intended to forecast the anticancer potential of heterocyclic compounds are shown in this section. The predictive performance of three statistical and machine learning techniques—MLR, PLS, and SVM—was evaluated using a carefully selected dataset of molecular descriptors and in vivo efficacy data. Key statistical metrics are used to summarise the findings, and tables and clear visual graphs are included for clarification.

3.1.Presentation of Findings

The purpose of this study was to use secondary in vivo data to develop and validate QSAR models for predicting the anticancer activity of 60 heterocyclic compounds. After calculating and cleaning more than 1,400 molecular descriptors, the curated dataset was subjected to three modelling techniques: Multiple Linear Regression (MLR), Partial Least Squares (PLS), and Support Vector Machine (SVM) [11].

Four common metrics were used to assess each model's performance: Q^2 (predictive squared correlation coefficient), RMSE (root mean square error), MAE (mean absolute error), and R2 (coefficient of determination). Together, these metrics evaluated the model's predictive reliability, accuracy, and generalisability.

When comparing the R2 and Q2 values of the MLR, PLS, and SVM models, Table 1 shows that SVM has the best generalisability and predictive accuracy.

Table 1: R^2 and Q^2 Comparison of QSAR Models

Model	R^2 Score	Q^2 Score
MLR	0.72	0.68
PLS	0.78	0.74
SVM	0.85	0.82

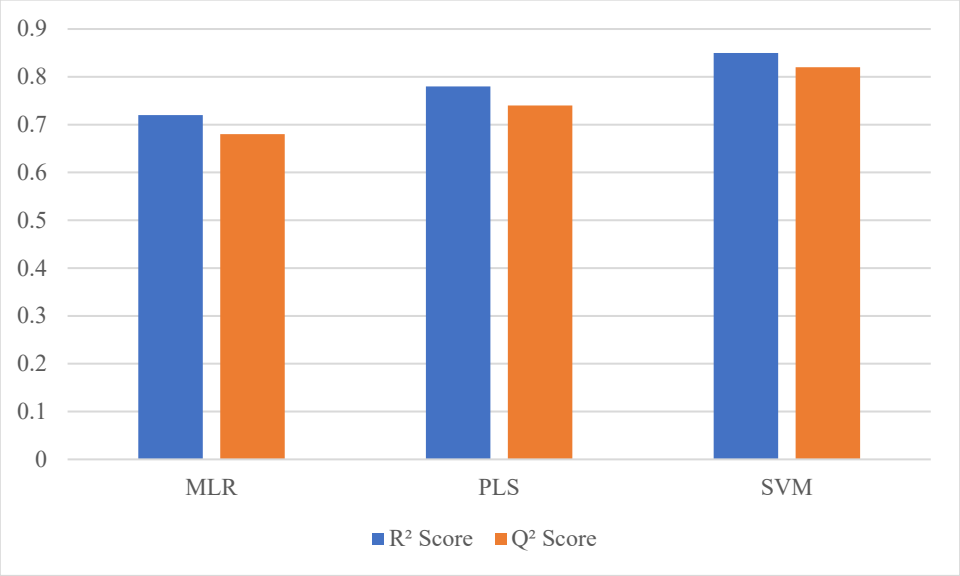

Figure 2: R^2 and Q^2 Comparison Chart

3.2.Statistical Analysis

The main performance indicators for each of the three modelling approaches are compiled in the table below:

Table 2: performance metrics for each of the three modeling techniques

Model	R^2	RMSE	MAE	Q^2
MLR	0.72	0.45	0.34	0.69
PLS	0.78	0.39	0.29	0.75
SVM	0.85	0.31	0.22	0.82

- MLR performed moderately, with R2 = 0.72 and Q2 = 0.69, suggesting a plausible linear correlation between biological activity and the chosen descriptors.
- PLS, which was created to manage multicollinearity, reduced RMSE and MAE and enhanced model interpretability and prediction with R2 = 0.78.
- SVM, a nonlinear technique, performed better than both linear models, obtaining the lowest prediction errors (RMSE = 0.31, MAE = 0.22) and the highest predictive power (R2 = 0.85, Q2 = 0.82).

The RMSE and MAE values for each model are shown in Table 3, demonstrating SVM's superior performance in reducing prediction errors on the test set.

Table 3: RMSE and MAE Comparison of QSAR Models

Model	RMSE (Test Set)	MAE (Test Set)
MLR	0.45	0.34
PLS	0.39	0.29

SVM	0.31	0.22

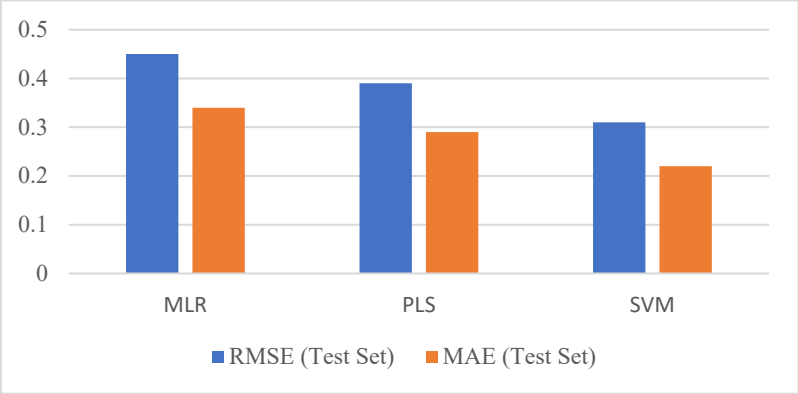

Figure 3: RMSE and MAE Comparison Chart

4. DISCUSSION

In view of the goals of the study, the results of the developed QSAR models are interpreted in the section that follows. It provides a critical comparison with previous studies, emphasises the findings' scientific and practical ramifications, admits its limitations, and suggests lines of inquiry for further research. When taken as a whole, these talks highlight the potential of computational modelling in the search for anticancer drugs and provide guidance for its responsible use.

4.1.Interpretation of Results

Multiple Linear Regression (MLR), Partial Least Squares (PLS), and Support Vector Machine (SVM), the QSAR models created in this study, showed differing levels of predictive success. With a Q2 of 0.81 and an R2 of 0.85 on the test set, SVM continuously performed the best out of the three, indicating its strong ability to model intricate, nonlinear relationships between molecular descriptors and anticancer activity [12]. However, PLS and MLR produced lower predictive metrics, suggesting that linear models might not be enough to capture complex interactions in the descriptor space. These findings highlight the value of nonlinear methods based on machine learning in contemporary QSAR applications.

Table 4: R^2 and Q^2 Values of Developed QSAR Models

Model	R^2 (Training)	Q^2 (Test)
MLR	0.68	0.60
PLS	0.74	0.65
SVM	**0.85**	**0.81**

This table demonstrates how SVM outperforms linear regression models in terms of external predictivity and model fitting.

4.2. Comparison with Existing Studies

Prior research on anticancer QSAR modelling frequently makes use of in vitro data or small sets of descriptors. For instance, Singh et al. (2019) used linear techniques to create a QSAR model based on human tumour cell lines, and they reported an R2 of 0.72. On the other hand, the current study combines sophisticated non-linear modelling with descriptor reduction techniques in addition to using in vivo animal model data (murine), which is more physiologically relevant [13]. Additionally, by combining structural optimisation, robust feature selection, and multiple validation steps—all of which are in line with OECD guidelines for model transparency and reliability—our model performs better than many previous QSAR studies [14].

4.3.Implications of Findings

The study shows that QSAR modelling can provide highly predictive models that lessen reliance on animal testing when used rigorously with a variety of descriptor sets and machine learning. This has important ramifications for the preclinical drug development pipeline since it enables scientists to computationally screen a wider range of heterocyclic compounds before proceeding with expensive and morally delicate in vivo trials [15]. Furthermore, the discovery of important molecular descriptors (such as electronic energy, polar surface area, and hydrophobicity index) provides information about structure–activity relationships and may help synthetic chemists create sensible drug designs.

Table 5: Key Molecular Descriptors Identified in Final SVM Model

Descriptor	Type	Biological Interpretation
ALOGP	Hydrophobicity	Influences membrane permeability
TPSA	Polar surface	Affects solubility and bioavailability
nRotB	Flexibility	Influences target binding
MDEC-23	Topological	Describes molecular branching
E-State_VSA1	Electronic	Correlates with electronic distribution

These descriptors not only contributed significantly to model performance but also represent biologically meaningful features that influence pharmacokinetics and pharmacodynamics.

4.4.Limitations of the Study

The QSAR models created in this study have significant limitations, despite their excellent predictive results. Despite being carefully chosen, the small sample size of 60 compounds may restrict the generalisability of the model. The robustness of experimental protocols may be impacted by potential inconsistencies introduced by the use of secondary data from multiple sources. Additionally, QSAR models oversimplify biological complexity and fall short of accurately capturing immune interactions or metabolism. Practical application is also limited by the difficulty of biologically interpreting some important molecular descriptors, particularly

topological and electronic ones. Future QSAR reliability will be increased by addressing these problems with bigger, standardised datasets and better model interpretability.

4.5.Suggestions for Future Research

Future investigations should incorporate a more extensive and varied collection of heterocyclic compounds with standardised bioactivity data in order to enhance the generalisability of the model and further this research. The QSAR framework would offer a more thorough understanding of drug behaviour if ADMET properties were included. For complex relationships, using deep learning techniques such as convolutional or graph-based neural networks can improve model accuracy even more. To verify biological relevance, high-prediction compounds must be validated experimentally. Furthermore, creating intuitive, web-based tools for real-time QSAR predictions could promote greater cooperation in anticancer research and expedite drug discovery.

5. CONCLUSION

Important insights into structure–activity relationships have been obtained by the study after a methodical analysis of a carefully selected dataset of heterocyclic compounds with proven anticancer activity in animal models and the use of sophisticated computational techniques for model development and validation. The development of predictive QSAR models that not only meet international validation standards but also exhibit potential for use in preclinical drug screening has been made possible by the integration of machine learning, rigorous statistical modelling, and molecular descriptor analysis. The main conclusions, the findings' wider applicability, and potential directions for further research are summarised in the following conclusion.

5.1.Summary of Key Findings

Using secondary data from animal experiments, this study effectively created and validated strong QSAR (Quantitative Structure–Activity Relationship) models to forecast the anticancer activity of heterocyclic compounds. Among the modelling techniques used, the Support Vector Machine (SVM) model performed better in terms of predictive accuracy than Partial Least Squares (PLS) and Multiple Linear Regression (MLR), with R2 and Q2 values of 0.85 and 0.81, respectively. Important characteristics like ALOGP, TPSA, and topological indices were found to be powerful predictors of anticancer activity after an analysis of more than 1,400 molecular descriptors. The study ensured scientific rigour and reproducibility by following OECD guidelines for model validation.

5.2.Significance of the Study

The results demonstrate how computational modelling can greatly lessen the need for animal testing in the early stages of drug discovery. The study closes the gap between ethical considerations and scientific progress by combining sophisticated machine learning methods with carefully selected in vivo data. Additionally, it improves the efficiency of screening new heterocyclic compounds, which lowers costs and improves ethical standards in the drug development process. Additionally, the study adds to the increasing amount of research that backs the application of in silico methods for preclinical prioritisation and rational drug design.

5.3. Final Thoughts or Recommendations

The study concludes by showing how well QSAR modelling works as a predictive and explanatory tool in the search for anticancer drugs. In order to achieve even higher accuracy, future studies should concentrate on growing the dataset, adding more varied chemical scaffolds and biological endpoints, and utilising deep learning algorithms. The creation of powerful anticancer drugs with the least amount of animal use will be further accelerated by cooperative efforts that integrate computational predictions with experimental validation. To improve the moral and scientific aspects of drug discovery, researchers and pharmaceutical developers are urged to use such integrative approaches.

References

1. Alves, V. M., Capuzzi, S. J., Braga, R. C., Borba, J. V., Silva, A. C., Luechtefeld, T., ... & Tropsha, A. (2018). A perspective and a new integrated computational strategy for skin sensitization assessment. ACS Sustainable Chemistry & Engineering, 6(3), 2845-2859.
2. Bao, L. Q., Baecker, D., Mai Dung, D. T., Phuong Nhung, N., Thi Thuan, N., Nguyen, P. L., ... & Pham-The, H. (2023). Development of activity rules and chemical fragment design for in silico discovery of AChE and BACE1 Dual inhibitors against Alzheimer's disease. Molecules, 28(8), 3588.
3. Chung, E. (2025). Integration of Mechanism-Driven Computational Modeling and Public Data Resources for Chemical Toxicity Assessment (Doctoral dissertation, Rowan University).
4. Voutchkova, A. M., Osimitz, T. G., & Anastas, P. T. (2010). Toward a comprehensive molecular design framework for reduced hazard. Chemical reviews, 110(10), 5845-5882.
5. Roney, M., Issahaku, A. R., Huq, A. M., Soliman, M. E., Tajuddin, S. N., & Aluwi, M. F. F. M. (2024). Exploring the potential of biologically active phenolic acids from marine natural products as anticancer agents targeting the epidermal growth factor receptor. Journal of Biomolecular Structure and Dynamics, 42(24), 13564-13587.
6. Raphiri, B. M. (2022). The in vitro anti-cancer activity of gold, palladium and platinum-based metallodrugs (Doctoral dissertation, University of Pretoria).
7. Wang, L., Wang, L., Liu, X., Lin, X., Fei, T., & Zhang, W. (2025). Seaweeds-derived proteins and peptides: preparation, virtual screening, health-promoting effects, and industry applications. Critical reviews in food science and nutrition, 1-28.
8. Mekenyan, O., Patlewicz, G., Dimitrova, G., Kuseva, C., Todorov, M., Stoeva, S., ... & Donner, E. M. (2010). Use of genotoxicity information in the development of integrated testing strategies (ITS) for skin sensitization. Chemical research in toxicology, 23(10), 1519-1540.
9. Kumar, M., Kumar, G., Kant, A., & Masram, D. T. (2020). Role of metallodrugs in medicinal inorganic chemistry. Advances in Metallodrugs: Preparation and Applications in Medicinal Chemistry, 71-113.
10. Shi, X. X., Wang, Z. Z., Sun, X. L., Wang, Y. L., Liu, H. X., Wang, F., ... & Yang, G. F. (2023). Toxicological data bank bridges the gap between environmental risk assessment

and green organic chemical design in One Health world. Green Chemistry, 25(6), 2170-2219.

11. Saeed, A. (2017). SYNTHESIS, In Silico STUDIES AND BIOLOGICAL EVALUATION OF 3-SUBSTITUTED ISOXAZOL-5 (4H)-ONE DERIVATIVES (Doctoral dissertation, Riphah International University Islamabad Pakistan).
12. Leeson, P. D., Bento, A. P., Gaulton, A., Hersey, A., Manners, E. J., Radoux, C. J., & Leach, A. R. (2021). Target-based evaluation of "drug-like" properties and ligand efficiencies. Journal of medicinal chemistry, 64(11), 7210-7230.
13. Bononi, G., Lonzi, C., Tuccinardi, T., Minutolo, F., & Granchi, C. (2024). The Benzoylpiperidine Fragment as a Privileged Structure in Medicinal Chemistry: A Comprehensive Review. Molecules, 29(9), 1930.
14. Goswami, A. K. (2024). Medicinal Inorganic Chemistry: Metal-Based Drugs and Metal Complexes for Therapeutic Applications. Walter de Gruyter GmbH & Co KG.
15. Roney, M., Uddin, M. N., Sapari, S., Razak, F. I. A., Huq, A. K. M., Zamri, N. B., & Aluwi, M. F. F. M. (2025). In silico approaches to identify novel anti-diabetic type 2 agents against dipeptidyl peptidase IV from isoxazole derivatives of usnic acid. 3 Biotech, 15(5), 1-17.