IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 3, October 2025
Impact Factor: 7.67

Exploring QSAR Models for the Selection of Candidate Inhibitors in Drug Development

Nandini
M.Sc Student, Centre of Biotechnology (Bioinformatics)
Maharshi Dayanand University, Rohtak
dr.nandni2k1@gmail.com

Abstract: *A Quantitative Structure–Activity Relationship (QSAR) model correlates the biological activity of a molecule with its structural characteristics using molecular descriptors that quantitatively define key structural features. In this study, novel molecular descriptors and modelling methods were developed for QSAR analysis across six target systems associated with diseases such as cancer, neurodegenerative disorders, HIV-AIDS, and malaria. The research introduced 2D image-based descriptors derived from optimized 3D molecular structures. These descriptors were constructed using Dijkstra's algorithm and multidimensional scaling to preserve interatomic shortest path distances and partial charges in two dimensions. Principal component analysis (PCA) and support vector regression (SVR) were employed to regress the descriptors against biological activity values, though these models were found to be computationally intensive. To enhance efficiency, a new 3D pseudo-molecular field (PMF) concept was developed based on intrinsic atomic properties such as electron affinity and electronegativity, rather than conventional electrostatic fields calculated from partial atomic charges. The PMF-based partial least squares (PMF-PLS) methodology, combined with Procrustes transformation, produced QSAR models with performance comparable to existing models while being computationally lighter. Additionally, a new regression approach, Varying Component Partial Least Squares (VC-PLS), utilizing the SIMPLS variant of PLS, was proposed for QSAR modelling. Both PMF-PLS and VC-PLS models were applied to screen natural compounds structurally similar to known bioactive molecules in the target systems. The screening outcomes from both models were consistent, and subsequent molecular docking studies validated the predicted interactions, supporting the reliability of the proposed QSAR methodologies for drug discovery applications.*

Keywords: QSAR , SVR , PMF , Drug Discovery

I. INTRODUCTION

QSAR modelling using 3D molecular field descriptors have been widely used to capture the relationship between a ligand and its biological activity (Nidhi and Siddiqi, 2013; Divakar and Hariharan, 2015). In particular, comparative molecular field analysis (CoMFA) uses 3D molecular descriptors (Cramer *et al.*, 1988; Dasoondi *et al.*, 2008; Matta and Arabi, 2011) that are developed by obtaining energy minimized 3D structures of the molecules along with the partial atomic charges calculated for every atom of the molecule. The molecular structures are oriented to structurally align with each other in a box of appropriate size having a suitably chosen 3D mesh grid.

Molecular fields, such as, electrostatic and/or steric, are then calculated for all the points on the grid using coulomb potential function and Lennard-Jones potential function, respectively (Cramer *et al.*, 1988), and a 3D array of field values is obtained for every molecule. The above 3D arrays are used as molecular descriptors to develop regression models that correlate with the biological activity of the molecules. Although, the CoMFA based 3D-QSAR models relate the structural information with the activities of molecules, the structural minimization routines required for calculation of partial atomic charges are intensive (Gasteiger and Marsili, 1980). Thus, there is a need to study novel and simpler 3D molecular descriptors that provide accurate 3D-QSAR models for practical purposes. Towards this aim, here we propose and study the use of intrinsic properties of the individual atoms, namely, electronegativity and electron

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 3, October 2025
Impact Factor: 7.67

affinity values to develop and study 3D molecular field like descriptors. We term this molecular field as the pseudo-molecular field (PMF) and the molecular descriptors as pseudo-field molecular descriptors (PFMD). These descriptors would have the advantage that the atomic property values used in their calculations will be readily available and would not require to be determined for every molecule unlike partial atomic charges. Developing QSAR models based on these PFMDs would then be simpler than CoMFA based models and studying its feasibility would provide a practical and correlative way of using intrinsic atomic properties for assessing the activity of a ligand with its target. It may be noted that these PFMDs are associated with high dimensionality (similar to the CoMFA descriptors) because of the consideration of PMF values in 3D spatial coordinates. We aim to bring out and discuss here a novel methodology employing PLS, namely PMF-PLS, for efficient QSAR modelling.

II. LITERATURE REVIEW

Quantitative Structure–Activity Relationship (QSAR) modelling remains a core in silico strategy for prioritizing potential inhibitors in drug discovery. Early reviews and domain-specific surveys stressed descriptor selection, model validation, and applicability domain issues as foundations for reliable prediction. Jahangiri et al. (2014) reviewed QSAR work on ACE-inhibitory peptides and highlighted how descriptor choice and validation strategies determine model usefulness in lead design. Through the late 2010s and into the 2020s, QSAR studies increasingly paired with docking and molecular dynamics to improve biological relevance. Wang et al. (2020) and numerous subsequent works illustrated that combining QSAR (2D/3D descriptors) with docking enriches virtual screening and helps rationalize binding modes for prioritized hits.

Methodological evolution accelerated in the 2020s: 3D-QSAR methods (including Gaussian field approaches) and multi-dimensional descriptors were applied to design kinase and enzyme inhibitors (Singh et al., 2022; Aloui et al., 2024). These studies demonstrated that integrating structural alignment, field-based descriptors, and simulation-derived features can produce potent inhibitor hypotheses for targets such as BRAF and BTK. Concurrently, machine learning (ML) and hybrid AI workflows expanded QSAR capability. Recent reviews (Koirala et al., 2025; Evangelista et al., 2025) document the shift from classical linear QSAR/PLS to ensembles, kernel methods, tree-based ML, and deep learning (including graph neural networks and SMILES transformers), which often yield improved predictive power for diverse inhibitor datasets — provided datasets are sufficiently large and curated. Descriptor innovation has been another important theme. Studies explored novel representations that balance physics-based interpretability and computational efficiency from advanced 3D field descriptors to compact image-based encodings and pseudo-molecular fields (PMF) that use intrinsic atomic properties (electronegativity, electron affinity) instead of partial charges. These approaches aim to retain 3D spatial information while reducing computational overhead for large-scale screening. Representative PMF work and related descriptor studies have emerged in the 2023–2025 literature.

Validation and reproducibility concerns have been emphasized repeatedly. Serafim et al. (2023) and others warn that small datasets, improper cross-validation, model overfitting, and lack of external test sets lead to false positives during virtual screening; they recommend robust external validation, applicability-domain estimation, and consensus modelling.

Application studies from 2020–2024 demonstrate QSAR's practical value: (i) ML-led QSAR for TBK1 and kinase inhibitors (Ivanov et al., 2024) and (ii) 2D/3D QSAR coupled with docking for various enzyme inhibitors (Batool et al., 2024; Khairullina et al., 2024) show successful hit-identification and subsequent experimental validation in some cases. Finally, the frontier (2024–2025) shows integration of QSAR with reinforcement learning and generative models for de novo inhibitor design, promising to close the loop between prediction and molecular generation (Zavadskaya et al., 2025; Koirala et al., 2025). These approaches are nascent but rapidly maturing, contingent on rigorous benchmarking and synthetic accessibility filters.

III. RESEARCH METHODOLOGY

For ease in discussion, a schematic flowchart of steps involved in development of PFMDs and PMF-PLS QSAR modelling are shown in Figure 1 with boxes labelled as that refer to [B Fig. # . Box #]. The details of steps in the individual boxes are discussed in subsections for modularizing the algorithm. Section 3 describes the procedure to

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 3, October 2025
Impact Factor: 7.67

import molecular structures and their biological activity values from PubChemBioAssay database [B 1.1] and the procedure to identify natural molecules from SuperNatural II (Banerjee *et al.*, 2015) database having scaffolds similar to the ones used in the chosen TS but whose pIC50 values are not known [B 1.3]. Thus, the inhibitory activities of these natural compounds could be studied using the PMF-PLS QSAR modelling. Section 3 outlines two steps of preprocessing of the inhibitor structures that are obtained from the databases. Firstly, we pre-process the structures using Ligprep© module (version 2.5, 2012) in Schrodinger software to obtain scaffold based alignment of 3D structures of the chosen inhibitor molecules [B 1.4]-[B 1.6]. In the next step [B 1.7]-[B 1.8], we import the aligned data into Matlab© (version R2010b) where the atoms in the molecule are accurately positioned in a 3D mesh grid.

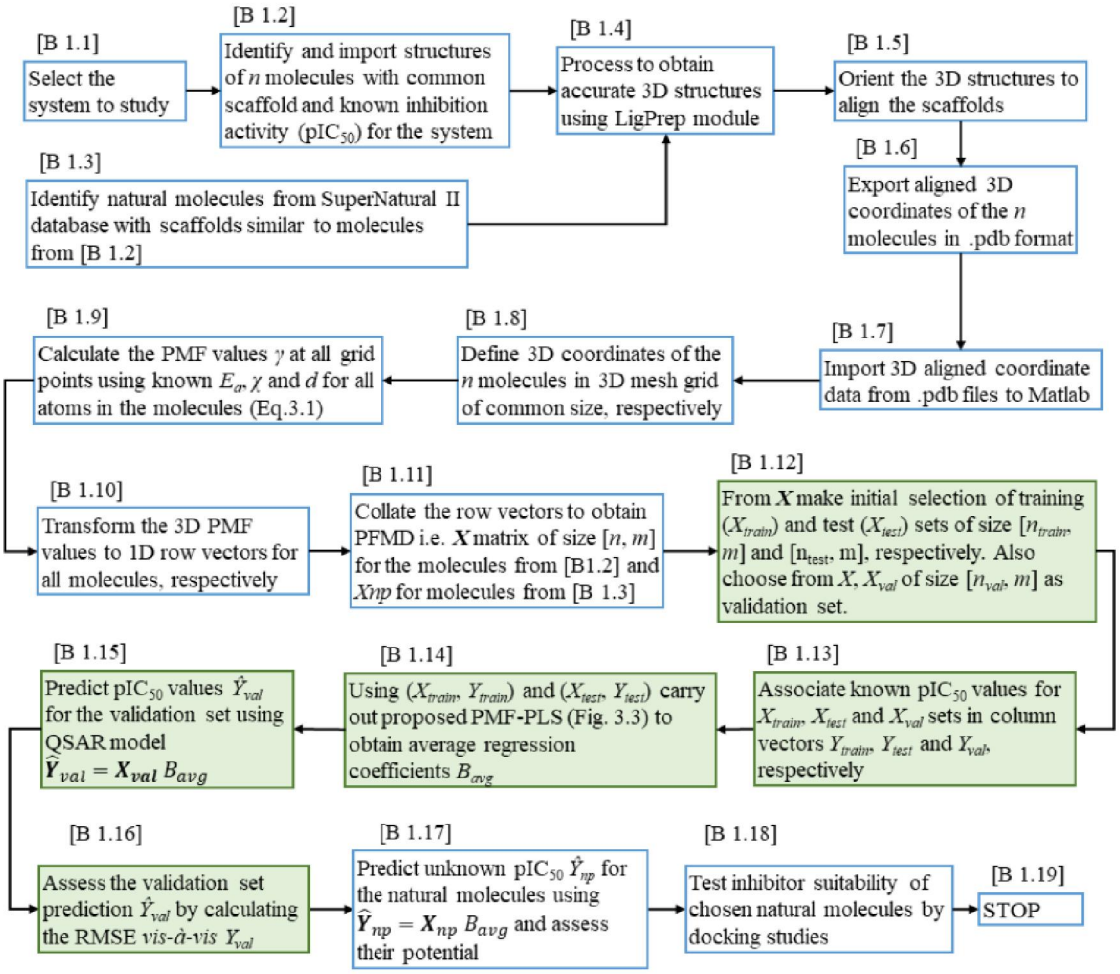

Figure 1: General flowchart to study the proposed PMF-PLS approach for QSAR Modelling

This section describes the calculation of PMF values at the mesh grid points [B 1.9] using electron affinity and electronegativity values of atoms to obtain the PFMDs [B 1.10]-[B 1.11]. Subsequently, Section 3.2.4 elucidates the steps in PMF-PLS algorithm that are used to develop the QSAR model and its validation [B 1.12]-[B 1.16].

We next use this model to calculate the pIC50 values of the natural molecules obtained from the Super Natural II database (Banerjee *et al.*, 2015) [B 1.17]. To further confirm the potential inhibitory actions of natural molecules with the calculated pIC50 values, it is proposed to carry out docking studies of these molecules to confirm that they indeed bind to the selected targets. Successful docking along with the prediction of high pIC50 value by the QSAR model would suggest that the molecule has a good potential for inhibiting the target [B 1.18].

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 3, October 2025
Impact Factor: 7.67

IV. RESULTS AND DISCUSSION

The 3D arrays of PMF values were converted into 1D PFMD arrays [B 1.10] and regressed with ***Y*** values using PMF-PLS methodology.

Regression models built using a single training set tend to have a bias for the training set used which can result in problems arising due to the overfitting of the QSAR model. A way to reduce the model bias is to use multiple training sets that yield average values of the regression coefficients to build the final QSAR model (Wold, 1978).

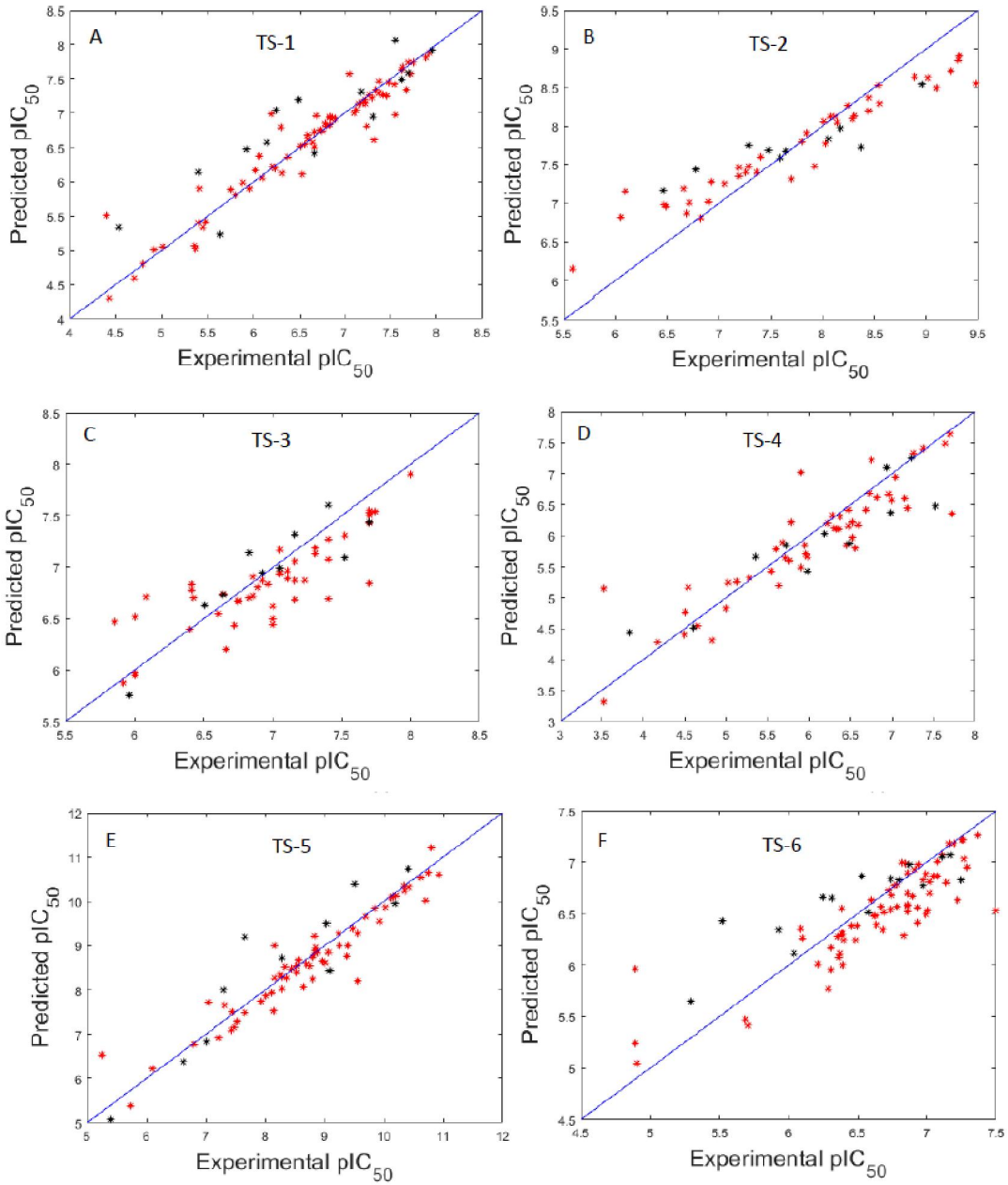

Figure 2: Plots of actual pIC50 values (***Y***) vs. the predicted values ($\hat{Y}$) for cross-validation using PMFPLS QSAR model. (A) TS-1, (B) TS-2, (C) TS-3, (D) TS-4, (E) TS-5 and (F) TS-6 inhibitors. The training set compounds are marked in red and test set compounds in black as specified in the Appendix, Tables A15 to A20, respectively

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 3, October 2025
Impact Factor: 7.67

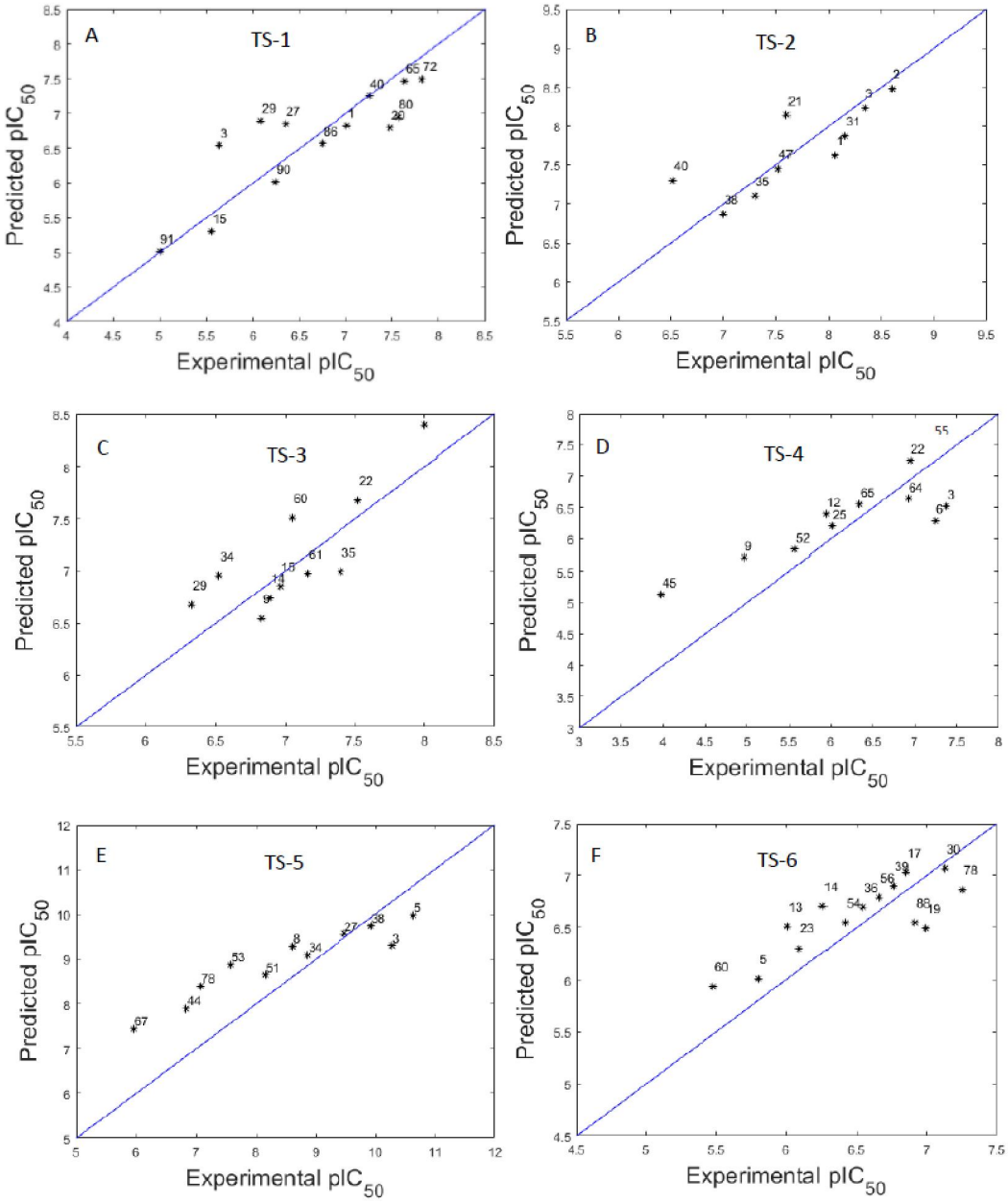

Figure 3: Plots for actual pIC50 values (***Yval***) vs. the predicted values (***Ŷval***) using PMF-PLS QSAR model for validation sets of (A) TS-1, (B) TS-2, (C) TS-3, (D) TS-4, (E) TS-5 and (F) TS-6 inhibitors.

The PMF-PLS QSAR model quality was further assessed by applying the mean absolute error (MAE) based criteria for the validation set predictions (Roy *et al.*, 2016).

V. CONCLUSION

The methodology of PMF-PLS is seen to offer a simpler way of QSAR modelling that uses an effective correlative descriptor in terms of the intrinsic properties of atoms, namely, the electron affinity and electronegativity values. This is in contrast to CoMFA where the descriptors are obtained using the partial atomic charges which are calculated

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 3, October 2025
Impact Factor: 7.67

separately for every molecule. We apply the PMF-PLS methodology to six target systems, namely, 4-phenylpyrrolocarbazole derivative inhibitors of WEE1 as anti-cancer compounds, benzylpiperidine derivative inhibitors of AChE against neurological disorders, 2-substituted dipyridodiazepinone derivatives and 2-pyridinone derivatives as HIV-1 RT inhibitors, cyclic urea derivatives as HIV-1 PR inhibitors and azilide derivatives as anti-malarial compounds. The QSAR models showed good prediction statistics for all six TSs and it brings out the viability of the PMF-PLS approach. It takes care of many practical situations encountered in QSAR modelling. Thus, the high dimensionality of the descriptor data could be reduced drastically by projection to a lower dimensional latent subspace. The practical problem of overfitting of model could then be addressed. The usefulness of Procrustes transformation in modifying the descriptor data for better optimization of PLS scores and loadings has been proposed which gave improved predictions. A comparison of the PMF-PLS QSAR modelling results with the QSAR models reported in the literature for the same set of inhibitors shows that the former yields comparable results. Additionally, PMF-PLS QSAR models were used to predict pIC50 values for natural compounds with unknown biological activities. The time taken for the PMF-PLS algorithm to arrive at the reference training and test sets (first part of the algorithm) was in the order of 6-8 hours. However, the second and third part of the algorithm took about 1-2 minutes to complete.

REFERENCES

[1]. Tarasova, O. A., et al. (2015). *Case study for HIV-1 reverse transcriptase inhibitors: suitability of public and commercial databases for QSAR modelling.* Journal of Chemical Information and Modelling.

[2]. Mena-Ulecia, K., Tiznado, W., & Caballero, J. (2015). *Study of the differential activity of thrombin inhibitors using docking, QSAR, molecular dynamics, and MM-GBSA.* PLOS ONE.

[3]. Singh, H., et al. (2015). *QSAR based model for discriminating EGFR inhibitors and non-inhibitors using a large dataset.* Biology Direct.

[4]. Ghanbarzadeh, S., et al. (2015). *2D-QSAR study of some 2,5-diaminobenzophenone farnesyltransferase inhibitors by different chemometric methods.* EXCLI Journal.

[5]. Lee, S., et al. (2015). *Development of 3D-QSAR model for acetylcholinesterase inhibitors combining docking and pharmacophore features*. (3D-QSAR study). PubMed.

[6]. Nongonierma, A. B., & FitzGerald, R. J. (2016). *Learnings from quantitative structure–activity relationship (QSAR) modelling applied to bioactive peptides.* RSC Advances.

[7]. Abuhammad, A. (2016). *QSAR studies in the discovery of novel antidiabetic agents: strategies and successful applications.* (Review). PubMed.

[8]. (2017) *2D- and 3D-QSAR analyses for EGFR inhibitors: approaches combining descriptors and modelling for inhibitor design.* (Study, 2017).

[9]. Sebastián-Pérez, V., et al. (2019). *QSAR modelling to identify LRRK2 inhibitors: multi-objective feature selection and machine-learning approaches.* (Case study).

[10]. Lahyaoui, M., et al. (2023). *QSAR modelling, molecular docking and ADMET assessment of phosphorus-substituted quinoline derivatives as potential inhibitors.* (Application study).

[11]. Aloui, M., et al. (2023). *QSAR, docking and dynamics studies for design of BTK inhibitors derived from pyrrolopyrimidine.* (Design & prediction study).

[12]. Jin, S., et al. (2024). *Structure-based 3D-QSAR modelling of RET kinase inhibitors: insights for design of potent inhibitors.* ACS Omega (or equivalent journal).

[13]. Khairullina, V., et al. (2024). *QSAR modelling and biological testing of 15-LOX inhibitors: descriptor selection and model validation.* (Experimental + QSAR).

[14]. Ancuceanu, R., et al. (2024). *QSAR regression models for predicting HMG-CoA reductase inhibitory activity and virtual screening for new inhibitors.* (Machine-learning QSAR).

[15]. Wang, Q., et al. (2024). *QSAR modelling and virtual screening for inhibitors of SARS-CoV-2 non-structural protein Nsp14: a first report.* (Targeted QSAR study).