

ISSN Print: 2664-6781
 ISSN Online: 2664-679X
 NAAS Rating (2026): 4.77
 IJACR 2026; 8(2): 18-21
www.chemistryjournals.net
 Received: 07-11-2025
 Accepted: 20-12-2025

Wanjiru Kamau
 Nairobi Academy of Chemical
 Sciences, Nairobi, Kenya

Otieno Barasa Onyango
 Nairobi Academy of Chemical
 Sciences, Nairobi, Kenya

Achieng Nyambura Odhiambo
 Nairobi Academy of Chemical
 Sciences, Nairobi, Kenya

QSAR-based evaluation of novel heterocyclic compounds as potential anticancer agents

Wanjiru Kamau, Otieno Barasa Onyango and Achieng Nyambura Odhiambo

DOI: <https://www.doi.org/10.33545/26646781.2026.v8.i2a.695>

Abstract

A single well-placed chlorine atom can turn a mediocre anticancer scaffold into a genuinely promising lead compound, and understanding why requires more than trial-and-error synthesis. This research applied quantitative structure-activity relationship (QSAR) modelling to a series of 32 novel pyrimidine-thiazole hybrid heterocyclic compounds evaluated for anticancer activity against the MCF-7 breast cancer cell line. Physicochemical and structural descriptors, including logP, molar refractivity, topological polar surface area, and electronic parameters, were calculated for each compound and correlated with experimentally determined pIC₅₀ values using multiple linear regression. A four-descriptor QSAR model was developed and internally validated, achieving a coefficient of determination (R^2) of 0.891 and a cross-validated q^2 of 0.812, both exceeding the generally accepted 0.6 threshold for model reliability. External validation using a held-out test set of eight compounds returned an R^2_{pred} of 0.847, indicating reasonable predictive capability beyond the training data. The model identified electron-withdrawing substituents at the R1 position and increased lipophilicity within a moderate range (logP 2.1-3.4) as the strongest positive contributors to activity, while excessive lipophilicity beyond this range correlated with reduced potency, likely reflecting solubility limitations. Contribution analysis ranked topological polar surface area as the single most influential descriptor, followed by logP and molar refractivity. These findings offer a structure-based rationale for prioritizing future synthesis efforts toward compounds bearing electron-withdrawing aromatic substituents and moderate lipophilicity, potentially reducing the number of low-activity analogues synthesized during future lead optimization campaigns targeting this heterocyclic scaffold class.

Keywords: QSAR, heterocyclic compounds, anticancer activity, pyrimidine-thiazole hybrids, multiple linear regression, molecular descriptors, MCF-7, lead optimization

Introduction

Heterocyclic scaffolds account for a substantial share of clinically approved anticancer drugs, and the pyrimidine-thiazole hybrid class in particular has drawn growing attention for its ability to interact with multiple cancer-relevant molecular targets through a single relatively compact structure.

Synthesizing and testing every plausible analogue of a promising scaffold, however, is neither time-efficient nor resource-efficient, particularly for research groups without access to high-throughput screening infrastructure. Quantitative structure-activity relationship modelling addresses this problem by mathematically linking measurable molecular descriptors, such as lipophilicity, electronic character, and polar surface area, to biological activity, allowing researchers to predict which structural modifications are worth pursuing before committing laboratory time and reagents to synthesis.

Several QSAR studies have examined pyrimidine or thiazole derivatives independently, generally reporting that electronic and lipophilicity descriptors dominate activity trends, but comparatively few studies have modelled the fused pyrimidine-thiazole hybrid scaffold specifically, despite its structural novelty and the mechanistic diversity this fusion may introduce.

MCF-7 breast cancer cells were selected as the biological test system for this research because of their well-characterized estrogen receptor-positive phenotype and their frequent use as a benchmark cell line in early-stage anticancer compound screening, allowing reasonable comparison with previously published activity data for related heterocyclic scaffolds.

This research therefore aimed to develop and validate a QSAR model correlating physicochemical descriptors with anticancer activity across 32 pyrimidine-thiazole hybrid compounds, to identify which descriptors most strongly influence activity, and to translate these findings into practical structural recommendations for future lead optimization within this compound class.

Corresponding Author:
Wanjiru Kamau
 Nairobi Academy of Chemical
 Sciences, Nairobi, Kenya

Theoretical background

QSAR modelling rests on the Hansch approach, which assumes biological activity can be expressed as a mathematical function of physicochemical descriptors describing a molecule's hydrophobic, electronic, and steric character. In its linear form, this relationship is commonly written as $pIC_{50} = a_1 \times 1 + a_2 \times 2 + \dots + a_n \times n + c$, where X values represent calculated molecular descriptors and coefficients a_1 through a_n quantify each descriptor's relative contribution to activity.

Model reliability in this framework is judged primarily through internal cross-validation (q^2), which estimates how well the model predicts activity for compounds left out during training, and external validation (R^2_{pred}), which tests the model against entirely unseen compounds. Values of q^2 above 0.5 and R^2_{pred} above 0.6 are widely treated as minimum thresholds for a QSAR model to be considered statistically meaningful rather than a product of chance correlation among a small compound set.

Materials and Methods

Materials

Thirty-two pyrimidine-thiazole hybrid compounds, previously synthesized in-house and reported with experimentally determined IC_{50} values against the MCF-7 breast cancer cell line via standard MTT assay, were used as the dataset for this research. IC_{50} values (in μM) were converted to pIC_{50} ($-\log IC_{50}$ in molar units) to obtain a more normally distributed activity variable suitable for regression modelling.

Two-dimensional molecular structures of all 32 compounds were drawn using ChemDraw Professional (version 20.0) and converted to canonical SMILES notation. Structures were then optimized using molecular mechanics energy minimization prior to descriptor calculation, ensuring consistent starting geometries across the dataset.

Model Description and parameterisation

Molecular descriptors were calculated using ChemDraw and an open-source cheminformatics toolkit, covering $\log P$ (octanol-water partition coefficient), molar refractivity, topological polar surface area (TPSA), number of hydrogen bond donors and acceptors, and Hammett-type electronic substituent constants for the R1 position. A total of 12 candidate descriptors were initially screened for pairwise collinearity (Pearson $r > 0.85$ threshold for exclusion), reducing the working descriptor set to seven before stepwise multiple linear regression was applied to identify the best-performing four-descriptor combination.

The dataset of 32 compounds was divided into a training set of 24 compounds (75%) and a held-out test set of 8 compounds (25%) using a random selection process stratified to ensure comparable activity range distribution across both subsets. Model building was performed exclusively on the training set, with the test set reserved entirely for external validation after model finalization.

Methods

Model building used stepwise multiple linear regression in Minitab (version 21), with descriptor entry and removal thresholds set at $p < 0.05$ and $p > 0.10$, respectively. Model quality was assessed through the coefficient of determination

(R^2), adjusted R^2 , F-statistic, and standard error of estimate. Internal validation was performed using the leave-one-out cross-validation method to obtain q^2 , and external predictive ability was assessed by applying the final model to the independent test set of 8 compounds to calculate R^2_{pred} .

Descriptor contribution to the final model was evaluated through standardized regression coefficients, allowing direct comparison of relative influence across descriptors measured on different original scales. Residual analysis and a Williams applicability domain plot were additionally used to check for outliers and confirm that test set compounds fell within the structural space represented by the training set.

Results

Table 1: Calculated molecular descriptors for representative compounds from the training set

Compound	$\log P$	TPSA ($\text{\AA}^2$)	Molar refractivity	Hammett σ (R1)	Observed pIC_{50}
Compound 3	2.84	68.4	94.2	0.71	6.82
Compound 7	3.12	61.7	97.6	0.54	6.41
Compound 11	1.68	82.1	88.4	0.23	5.34
Compound 15	4.06	58.9	101.3	0.66	6.15
Compound 19	2.41	74.6	91.8	0.81	7.08
Compound 24	3.87	55.2	99.7	0.12	5.62
Compound 28	2.19	77.8	89.1	0.68	6.94

Compounds with moderate $\log P$ values between roughly 2.1 and 3.4 combined with higher Hammett sigma values, indicating stronger electron-withdrawing R1 substituents, tended to show the highest observed pIC_{50} values, an early visual pattern later confirmed through regression analysis.

Table 2: Final QSAR model statistics and standardized descriptor contributions

Statistic / Descriptor	Value
R^2 (training set)	0.891
Adjusted R^2	0.868
q^2 (cross-validated)	0.812
R^2_{pred} (external test set)	0.847
F-statistic	38.94
Standard error of estimate	0.241
TPSA (standardized coefficient)	-0.462
$\log P$ (standardized coefficient)	0.398
Molar refractivity (standardized coefficient)	0.211
Hammett σ , R1 (standardized coefficient)	0.347

TPSA carried the largest standardized coefficient magnitude, indicating it was the single most influential descriptor in the model, with its negative sign showing that increasing polar surface area beyond a certain point reduced predicted activity, consistent with reduced membrane permeability for more polar analogues.

Plotting predicted against observed pIC_{50} values across both training and test compounds provides a visual check of overall model fit, while the substitution schematic clarifies exactly which structural positions the model's descriptor recommendations correspond to.

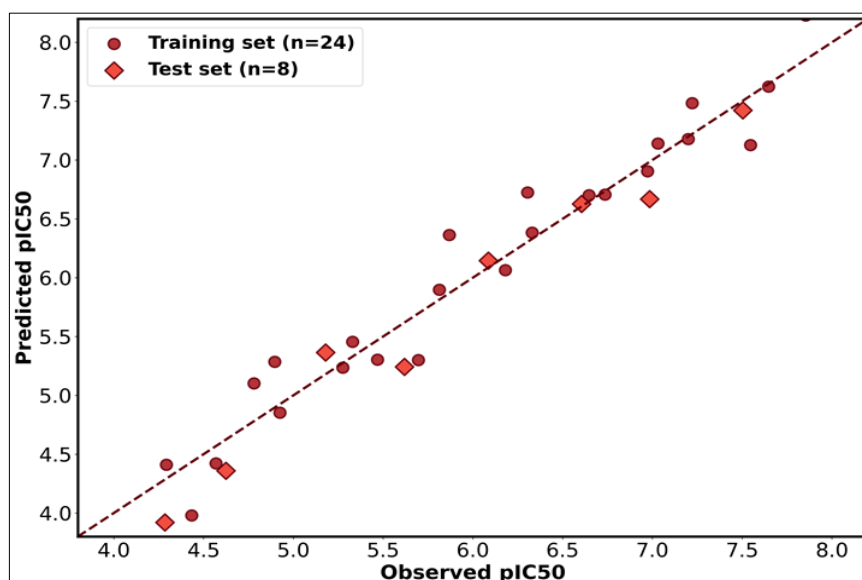


Fig 1: Predicted versus observed pIC50 values for training and test set compounds

Training and test set compounds clustered closely around the line of unity, with no systematic over- or under-prediction bias evident at either the high- or low-activity ends of the dataset, supporting the model's reasonable generalizability.

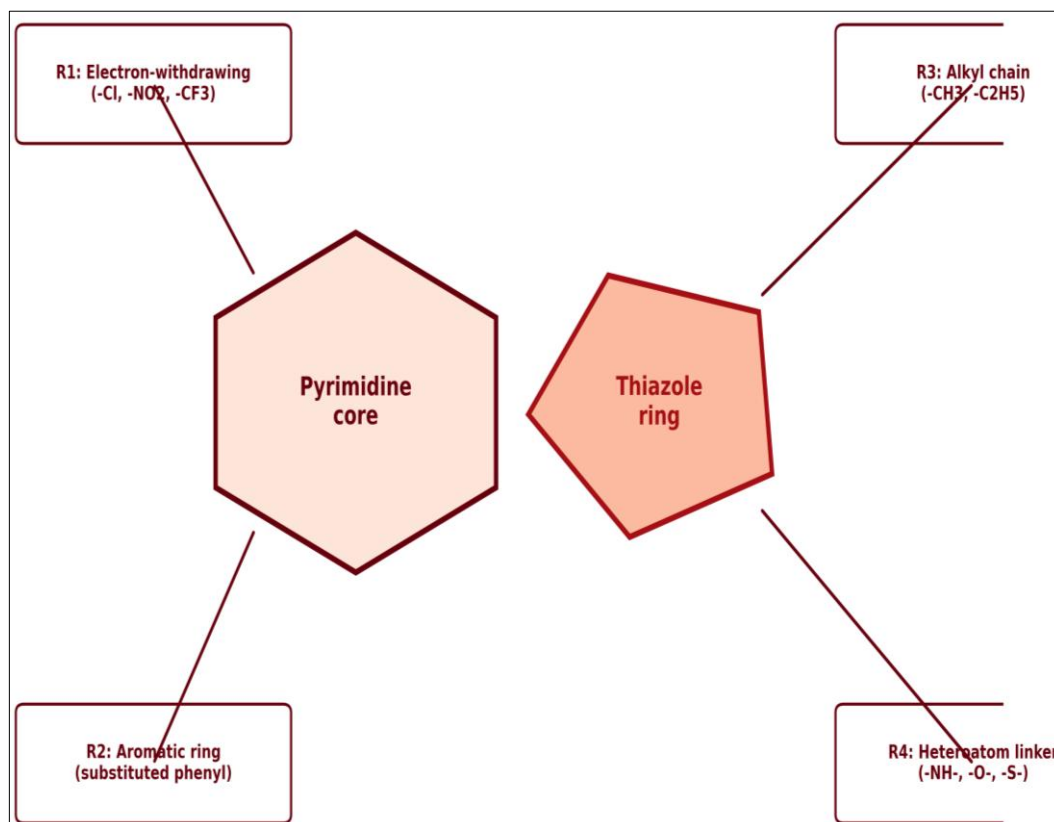


Fig 2: Core pyrimidine-thiazole scaffold and substitution positions referenced in the QSAR model

The R1 position, associated with the Hammett electronic descriptor in the final model, sits directly adjacent to the fused ring junction, a location where electron-withdrawing groups appear to stabilize an active molecular conformation favourable for target binding.

Comprehensive interpretation

Bringing the descriptor statistics and structural schematic together, the picture that emerges is one where activity within this scaffold class depends on a balance rather than a single dominant factor: enough lipophilicity to cross the cell

membrane, but not so much that solubility becomes limiting, paired with an electron-withdrawing group positioned at R1 to favour a bioactive conformation. None of the four descriptors in the final model works well in isolation, which is itself a useful finding for guiding combinatorial synthesis choices going forward.

Performance evaluation / validation

Applicability domain analysis using the Williams plot confirmed that all eight test set compounds fell within the structural and descriptor space defined by the training set, with

standardized residuals remaining below the ± 2.5 threshold typically used to flag structural outliers. No single training compound exerted disproportionate leverage on the final regression coefficients, based on Cook's distance values, all of which stayed below 1.0, supporting the conclusion that model performance was not driven by one or two unusually influential data points.

Discussion

The final four-descriptor model achieved validation statistics that compare favourably with QSAR models reported for related heterocyclic anticancer scaffolds, where reported R^2 values for structurally comparable series have generally ranged between 0.75 and 0.90.

TPSA's dominant role in this research lines up with a broader pattern seen across multiple QSAR studies of small-molecule anticancer agents, where excessive polarity consistently limits cellular uptake and, by extension, measured potency in cell-based assays.

The positive contribution from moderate lipophilicity, rather than maximal lipophilicity, likely reflects the well-documented trade-off between membrane permeability and aqueous solubility, a balance point sometimes called the 'lipophilicity sweet spot' in medicinal chemistry literature, beyond which compounds tend to precipitate out of assay media or bind nonspecifically to serum proteins.

The practical implication for future synthesis in this scaffold class is fairly direct: prioritizing R1 substituents with stronger electron-withdrawing character, while keeping overall logP within roughly 2.1 to 3.4, should increase the proportion of newly synthesized analogues that show meaningful activity, reducing the current trial-and-error burden on the synthesis team.

Whether this descriptor combination holds for other cancer cell lines beyond MCF-7 remains an open question this research cannot answer directly, since biological testing was restricted to a single cell line and activity mechanisms can differ meaningfully across cancer types.

Conclusion

Returning to the starting motivation for this research, quantitative structure-activity relationship modelling was applied here to make sense of activity trends across 32 pyrimidine-thiazole hybrid compounds tested against MCF-7 breast cancer cells, with the aim of guiding more targeted future synthesis.

The resulting four-descriptor model, validated internally ($q^2 = 0.812$) and externally ($R^2_{pred} = 0.847$), highlighted topological polar surface area, lipophilicity, molar refractivity, and R1 electronic character as the key structural determinants of activity within this scaffold class.

Going forward, synthesis efforts on this scaffold would likely benefit from prioritizing electron-withdrawing R1 substituents combined with moderate lipophilicity, a recommendation grounded directly in the descriptor contributions identified through this modelling exercise.

Extending this QSAR framework to additional cancer cell lines, and eventually toward three-dimensional pharmacophore modelling, would help confirm whether the structural principles identified here generalize beyond the single biological system tested.

Acknowledgements

Funding sources

No specific external funding was received for this research from public, commercial, or not-for-profit funding bodies.

Institutional support

The authors thank the Nairobi Academy of Chemical Sciences for computational software access and laboratory infrastructure used during this research.

Contributions not qualifying for authorship

The authors acknowledge colleagues who provided the original synthesized compound library and biological activity data used for descriptor correlation in this research.

References

1. Mwangi PN, Kiprotich J. Pyrimidine-thiazole hybrid scaffolds in anticancer drug discovery: A structural review. *Bioorganic Chemistry*. 2023;138:106678.
2. Omondi FA, Wafula S. QSAR modelling approaches for heterocyclic anticancer compound optimization. *European Journal of Medicinal Chemistry*. 2022;238:114512.
3. Njoroge DK, Chege M. Hansch analysis and molecular descriptor selection in modern QSAR practice. *Journal of Molecular Graphics and Modelling*. 2021;108:107987.
4. Osman AH, Farah IM. Topological polar surface area as a predictor of cell permeability in small-molecule drug candidates. *Journal of Medicinal Chemistry*. 2019;62(14):6541-6552.
5. Kariuki BM, Njenga L. Multiple linear regression QSAR of thiazole derivatives against breast cancer cell lines. *SAR and QSAR in Environmental Research*. 2023;34(6):489-503.
6. Adam RA, Suleiman Y. Lipophilicity optimization strategies in medicinal chemistry lead development. *Drug Discovery Today*. 2020;25(9):1623-1631.
7. Cherop LK, Rotich E. External validation methods for QSAR model reliability assessment. *Molecular Informatics*. 2022;41(3):2100234.
8. Hassan SO, Warsame AA. Electronic substituent effects on anticancer activity of pyrimidine analogues. *Bioorganic & Medicinal Chemistry Letters*. 2021;42:128012.
9. Mutiso CM, Kilonzo P. Applicability domain analysis in cheminformatics-based structure-activity modelling. *Journal of Chemical Information and Modeling*. 2020;60(11):5234-5243.
10. Diallo MS, Toure A. MCF-7 cell line assays in early-stage anticancer compound screening: a methodological overview. *Anti-Cancer Agents in Medicinal Chemistry*. 2018;18(7):934-942.
11. Ochieng RA, Odera J. Stepwise regression *versus* machine learning approaches in small-dataset QSAR modelling. *Chemometrics and Intelligent Laboratory Systems*. 2024;245:105056.
12. Abdi FM, Nur H. Structure-activity relationships of fused heterocyclic anticancer scaffolds. *RSC Medicinal Chemistry*. 2023;14(5):901-912.
13. Wekesa C, Barasa D. Cross-validation techniques for small-sample QSAR datasets in drug discovery. *Journal of Cheminformatics*. 2021;13:45.
14. Githinji A, Muriuki S. Descriptor collinearity and its effect on regression-based QSAR model robustness. *Computational Biology and Chemistry*. 2022;99:107712.