ECOLOGIA BALKANICA

2025, Vol. 17, Issue 1 June 2025 pp. 216-224

Tautomerism influence on QSAR/QSPR modelling of ecotoxicity and physicochemical properties of chemical compounds

Nikolay Kochev[1,*], Vesselina Paskaleva[1], Nina Jeliazkova[2]

[1]University of Plovdiv "Paisii Hilendarski", Faculty of Chemistry, Department of analytical chemistry and computer chemistry, Tsar Assen Str. 24, Plovdiv, BULGARIA
[2]Ideaconsult Ltd., Sofia, BULGARIA
*Corresponding author: nick@uni-plovdiv.net

Abstract. The OECD guidelines for QSAR/QSPR (Quantitative Structure Activity/Property Relationship) modelling and ecotoxicity testing play a significant role in ecological studies by providing standardised, scientifically validated methods to assess the environmental impact of chemicals. Laboratory and terrain tests are combined with QSAR/QSPR models, ensuring consistency, reliability and support for regulatory acceptance, enabling screening and prioritisation, global harmonisation and promotion of alternatives to animal testing. Tautomerism is a fundamental structural phenomenon that can significantly influence the predictive reliability of QSAR/QSPR models used for assessing ecotoxicity and physicochemical properties of chemical compounds. We examine how tautomeric variation affects molecular descriptors, data curation, and model performance. We propose practical incorporation of tautomer information into the model development and cheminformatics pipelines to enhance predictive accuracy and regulatory applicability. By generating exhaustive tautomeric ensembles, our approach supports QSAR/QSPR modelling in line with OECD guidelines for regulatory ecotoxicity endpoints, contributing to the development of robust in silico predictions. By addressing the challenges posed by tautomerism, this work advances the use of computational methods in sustainable chemical safety assessment and supports innovation in non-testing approaches.

Key words: ecotoxicity, tautomers, QSAR, QSPR, modelling, molecular descriptors.

Introduction

The OECD Harmonised Templates (OHTs) are standardised data formats used for reporting study results on chemical substances, including ecotoxicological data (Harmonised Templates for Reporting Chemical Test Summaries (OHTs), 2025). The ecotoxicity endpoints covered by OHTs include those relevant for regulatory submissions under REACH (Registration, Evaluation, Authorisation and Restriction of Chemicals) (Regulation - 1907/2006 - EN - REACH - EUR-Lex, n.d.), OECD guidelines, and other international frameworks. REACH regulation requires a large set of endpoints to be reported in the context of ecotoxicity, e.g. Acute toxicity to fish (OHT 41), Toxicity to aquatic algae and cyanobacteria (OHT 46), Chronic toxicity to fish (OHT 42), Toxicity to soil macroorganisms (OHT 50), Biodegradation and environmental fate endpoints (OHT 24-40) and many others. In line with Safe and Sustainable by Design (SSbD) EU recommendations (Caldeira et al., 2023) and the ongoing shift toward alternative testing methods, with implications for REACH-compliant chemical assessments and regulatory data workflows, special attention is given to the facilitation of animal-free testing strategies based on quality QSAR/QSPR models.

Tautomerism presents a significant challenge in the processing of structural information. Due to differences in their topological representations, tautomers are often treated as distinct molecular entities. This distinction affects core structure representation and, in turn, impacts the entire cheminformatics workflow. As a result, tautomerism

should not be overlooked (Warr, 2010)—especially in the context of QSAR modelling and, more specifically, in ecotoxicity assessments. While the generation of large volumes of high-quality chemical data related to tautomers is a valuable step forward, it raises an important question: how can this wealth of data be effectively utilized?

This paper aims to explore the impact of tautomerism on data variability across different contexts and to examine how this variability affects tasks related to chemical information processing and QSAR modelling. We investigate how incorporating multiple tautomeric forms influences various cheminformatics operations. To the best of our knowledge, the use of tautomeric information has so far been limited to a few specific problem domains, with only isolated examples of its application reported in the literature.

Materials and methods

Tautomer generation

We utilised our in-house developed software, AMBIT-Tautomer, to exhaustively and automatically generate all tautomeric forms of a given chemical compound (Kochev et al., 2013). The underlying algorithm has been theoretically validated and benchmarked against leading tautomer generation tools, demonstrating the efficiency and reliability of AMBIT-Tautomer in producing high-quality chemical information. In this study, all tautomeric forms of the test compounds were generated using the AMBIT-Tautomer software, which employs the IA-DFS algorithm—an incremental approach based on the depth-first search algorithm. It incorporates tautomeric transformation rules for 1,3- and 1,5-hydrogen shifts, along with post-processing steps such as the removal of topologically equivalent structures and filtering of allene structures.

Cheminformatics processing

In this study, we used three large data sets, as summarised in Table 1, with the full datasets available in the Supplementary materials. The structures of generated tautomeric forms were preprocessed using ChemAxon Standardizer (version 5.12.2) including: extraction of SMILES (Weininger, 1988) linear notation from sdf files, kekulization of aromatic structures, conversion of explicit hydrogen atoms to implicit ones and removal of stereo information. To reduce computational time during tautomer generation, additional structural filtering was applied to the DrugBank (Knox et al., 2011) and NCI (NCI Database, n.d.) datasets. Specifically, all molecules containing more than 60 heavy atoms or more than four rings were excluded. As a result, the size of the DrugBank dataset was reduced from 6,477 to 5,550 structures. The same preprocessing approach was applied to the NCI dataset. The original NCI dataset contained 203,576 compounds, which initially yielded 3,767,644 tautomeric forms. After filtering, a subset of 70,878 structures was selected for this study, resulting in a total of 1,379,518 tautomeric forms used in subsequent analyses.

Table 1. List of the used data sets and their main characteristics.

Dataset Name	Number of structures	Number of generated tautomeric forms	Average number of tautomers per structure
Drug Bank	5550	174,777	32
NCI	70878	1,379,518	14
Ames	5451	73,028	13

Comparison and Testing

This study presents an in-depth analysis of the variance in descriptor and fingerprint values induced by tautomerism. Molecular descriptors and fingerprints for all datasets and their corresponding sets of tautomeric forms were calculated using PaDEL-Descriptor (Yap, 2011).

The impact of tautomerism on a given molecular descriptor was assessed using the mean relative standard deviation (RSD), calculated across all compounds with at least two tautomeric forms. The relative standard deviation, RSD(D_{ij}), for the j-th descriptor of structure, i, with n_i tautomers (k=1, 2, …, n_i), was computed as follows:

$$\boldsymbol{RSD(D_{ij})} = \sqrt{\frac{1}{n_i}\sum_{k=1}^{n_i}\left(\boldsymbol{D_{ij}^k} - \boldsymbol{mean(D_{ij})}\right)^2 / \boldsymbol{mean(D_{ij})}} \qquad (1)$$

The mean RSD for descriptor *j* across a data set containing *n* structures (i=1, 2, …, n) is calculated using the following formula:

$$\boldsymbol{mean - RSD(D_j)} = \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{RSD(D_{ij})} \qquad (2).$$

QSAR model building

We utilized two existing QSPR models for prediction of the partition coefficient LogP – the XlogP and Crippen LogP models, as implemented in the Chemistry Development Kit (CDK) (Steinbeck et al., 2006) and integrated into the PaDEL-Descriptor software.

In addition, we developed two new QSAR models as part of this study:

1. An AMES mutagenicity model, trained on a dataset of 5451 compounds, was built using the Random Forest algorithm with 80 trees. A total of 623 fingerprint bits (selected from 2212 fingerprints calculated with the PaDEL software) were used as input features.

2. An aquatic toxicity QSAR model for *Tetrahymena pyriformis* was constructed using a training set of 644 structures and an external validation set of 110 compounds.

Two principal variants of the aquatic toxicity model were created: a conventional model (CM) that does not incorporate tautomeric forms and a weighted model (WM) that uses tautomer-weighted descriptor values to account for tautomerism. All model variations were based on PaDEL descriptors (0D, 1D and 2D) and developed using WEKA software (version 3.7.9) (Frank et al., 2016). Descriptor selection was carried out using the CfsSubsetEval evaluator in combination with the Best First search strategy. Several machine learning algorithms were tested: KNN (k = 10) with inverse distance weighting and Manhattan distance, ExtraTree, and REPTree classification algorithms. For the conventional model, 27 descriptors were selected out of 729 total PaDEL descriptors, while 24 descriptors were chosen for the weighted model. The list of selected descriptors and their corresponding mean RSD values is presented in Table 2.

The mean absolute error (MAE), a key statistical measure of model performance, was calculated (see Eq. 3) for different subsets of compounds, grouped by the number of tautomeric forms they possess:

$$\boldsymbol{MAE} = \frac{\mathbf{1}}{\boldsymbol{n}}\sum_{\boldsymbol{i}=\mathbf{1}}^{\boldsymbol{n}}|\boldsymbol{P_i} - \boldsymbol{F}(\boldsymbol{S_i})| \qquad (3).$$

Table 2. Selected descriptors for the Conventional Model (CM) and the Weighted Model (WD) with the corresponding mean RSD values.

CM Descriptors	Mean RSD	WM Descriptors	Mean RSD
ALogp2	0.221	ALogp2	0.221
CrippenLogP	0.152	ATSc3	0.642
ETA_Alpha	0	CrippenLogP	0.152
ETA_AlphaP	0	ETA_Alpha	0
ETA_dAlpha_A	0	ETA_AlphaP	0
ETA_dEpsilon_A	0.001	ETA_dAlpha_A	0
ETA_Epsilon_1	0	ETA_dEpsilon_A	0.001
maxdS	0.003	ETA_dEpsilon_B	0.182
maxHCsats	0.169	ETA_dEpsilon_D	0.301
MDEN-11	0	ETA_Epsilon_1	0.0004
MDEN-33	0	maxdS	0.003
mindCH2	0.078	maxHCsats	0.169
mindS	0.003	MDEC-11	0
minHBint7	0.002	MDEN-11	0
minHsSH	0.0004	MDEN-33	0.08
minsNH2	0.068	mindS	0.003
MLFER_BH	0.08	MLFER_BO	0.123
MLFER_BO	0.123	n5Ring	0
n5Ring	0	n7Ring	0
n7Ring	0	ndssS	0.001
ndssS	0.001	nT7Ring	0
nssssC	0	SddC	0
nT7Ring	0	SHCHnX	0.106
SddC	0	SHssNH	0.021
SHCHnX	0.106		
SHsSH	0.004		
SHssNH	0.021		

Results

Fig. 1 illustrates the variation of two different molecular descriptors for the compound tacrine, which has 22 tautomeric forms generated with the AMBIT-Tautomer software. As shown, the descriptor values exhibit notable variability, with relative standard deviations (RSD) as follows: TIE (E-state topological parameter) - RSD_{TIE} = 0.07 (7%) and DELS (molecular electrotopological variation) - RSD_{DELS} = 0.11 (11%). These RSD values serve as localised measures of data variability for this specific molecule and highlight the intrinsic uncertainty introduced by tautomerism. Such information could be leveraged to assess or weight the reliability of a chemical structure within QSAR modelling and related cheminformatics tasks.

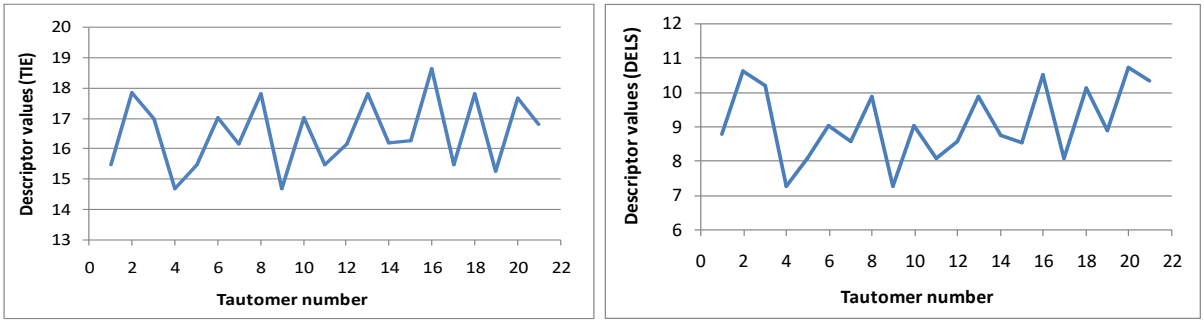


Fig. 1. Descriptor variation for the molecule of tacrine due to 22 tautomeric forms.

Fig. 2 presents histograms of the mean relative standard deviation (mean-RSD) values for PaDEL descriptors, calculated for the NCI and DrugBank datasets. The results from both datasets are consistent and reveal that approximately 60% of the descriptors show no variation or only statistically insignificant variation due to tautomerism. These descriptors are primarily atom counts, other constitutional descriptors, and autocorrelation descriptors (ATS) based on properties unaffected by tautomeric shifts, such as atomic mass, ring counts, or functional group counts that do not involve tautomerizable atoms.

About 30% of the descriptors exhibit moderate variability, with mean-RSD values in the range of 0.05 to 0.15. In our view, these descriptors should be considered carefully in various cheminformatics applications. Approximately 10% of the descriptors demonstrate high variability, with mean-RSD values exceeding 0.20, and in some cases, even greater than 1.00 (i.e., 100%). Notably, the descriptors with the highest variability include: hydrophilic and hydrophobic factors (including various models for LogP), hydrogen bond donors count, topological autocorrelations based on partial charges etc. Importantly, many of these highly variable descriptors are essential for QSAR modelling, including LogP, H-bonding features, and charge-related information. This underscores the importance of accounting for tautomerism in cheminformatics workflows that rely on such descriptors, particularly in the development of robust and reliable QSAR models.

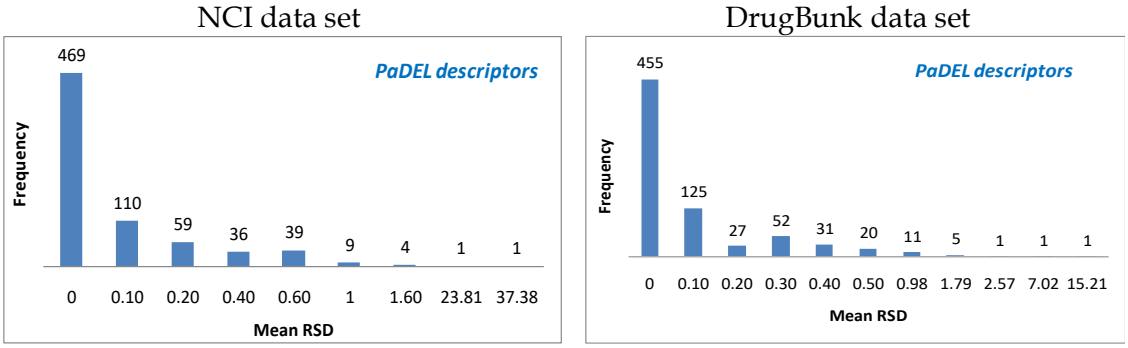


Fig. 2. Distribution of mean-RSD values for PaDEL descriptors (statistics are performed for NCI and DrugBunk data sets).

Hashed fingerprints are calculated using a dynamic set of fragments, uniquely generated for each molecular structure. These fragments are derived by exhaustively searching all topological paths within the molecule, typically considering all paths up to a specified maximum length. Each identified path (i.e. fragment) is then encoded by applying a hash function, which maps the fragment to one or more bit positions in the fingerprint vector. Our results indicate that hashed fingerprints are significantly affected by tautomerism, primarily due to changes in topological paths caused by the redistribution of double bonds. This structural variation alters the set of generated fragments, leading to noticeable differences in the resulting MACCS fingerprint representations of tautomeric forms, as shown in Table 3.

Table 3. Hashed fingerprints for the tautomers of methimazole.

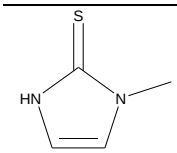
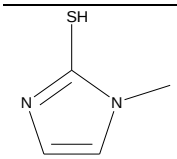
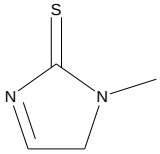
Structure	Fingerprint bits equal to 1	N_{FP}
	9,17,18,24,30,36,41,57,75,83,90,111,134,178,202,274,309,346,357,374,402,468,544,558,560,561,577,589,600,615,651,657,659,660,664,712,722,742,743,745,751,802,815,822,846,847,856,873,877,921,931,935,949,953,954,972	56
	2,12,13,18,32,56,88,137,146,148,152,156,178,195,215,219,222,228,229,238,284,287,318,345,352,374,375,380,413,449,485,490,505,528,533,543,575,592,615,620,655,681,736,742,743,745,798,850,855,879,883,886,908,912,913,921,934,952,953,964,970,972,973,978,984,986,990,994	68
	18,41,72,73,75,106,130,141,201,232,234,239,259,262,263,265,309,328,348,353,361,389,392,429,456,506,511,514,538,544,600,601,611,622,638,651,658,660,685,687,712,720,742,743,745,913,921,935,949,953,954,955,972,975,1017,1020,1023	57

Table 4 presents the variation in XLogP values (Wang et al., 1997), calculated using the CDK software library (version 1.4.18) across all 15 tautomeric forms of violuric acid. The XLogP values range from -1.26 to 1.23, indicating that some tautomers are predicted to be hydrophilic, while others are considered hydrophobic. Similarly, the predicted Ames mutagenicity results show variation: while most tautomers are estimated to be mutagenic, two are predicted to be non-mutagenic.

Table 4. Predicted Ames mutagenicity and XLogP values for all tautomers of violuric acid.

Violuric acid tautomers /SMILES notations/	Ames Mutagenicity (model)	XLogP
O=C1NC(=O)C(=NO)C(=O)N1	mutagenic	0.135
O=C1N=C(O)N=C(O)C1(=NO)	non-mutagenic	-0.086
O=C1N=C(O)C(=NO)C(O)=N1	non-mutagenic	0.267
O=C1N=C(O)C(=NO)C(=O)N1	mutagenic	0.041
O=C1N=C(O)NC(=O)C1(=NO)	mutagenic	0.361
O=NC1=C(O)N=C(O)N=C1(O)	mutagenic	-0.102
O=NC=1C(=O)NC(O)=NC=1(O)	mutagenic	-0.084
O=NC=1C(=O)N=C(O)NC=1(O)	mutagenic	1.230
O=NC=1C(O)=NC(=O)NC=1(O)	mutagenic	0.698
O=NC=1C(=O)NC(=O)NC=1(O)	mutagenic	0.363
O=NC1C(O)=NC(=O)N=C1(O)	mutagenic	-0.277
O=NC1C(=O)N=C(O)N=C1(O)	mutagenic	-1.056
O=NC1C(=O)NC(=O)N=C1(O)	mutagenic	-0.932
O=NC1C(=O)N=C(O)NC1(=O)	mutagenic	-1.038
O=NC1C(=O)NC(=O)NC1(=O)	mutagenic	-1.267

Fig. 3 illustrates how the mean absolute error (MAE) varies with the number of tautomers. Applying a weighting scheme, based on average values (AV), improves the CrippenLogP model by approximately 0.15 logarithmic units for structures with 4 to 50 tautomers. However, this weighting scheme negatively impacted model performance for structures with a very high number of tautomers. The sample size for these structures is very limited, and the statistics are less representative. This decline can be attributed to the simple averaging method used, which assigns equal weight to all tautomers—including many chemically irrelevant ones—when calculating the modified model, thereby degrading the results. This effect could likely be mitigated by employing a more refined weighting approach, such as averaging only the top five lowest-energy tautomers.

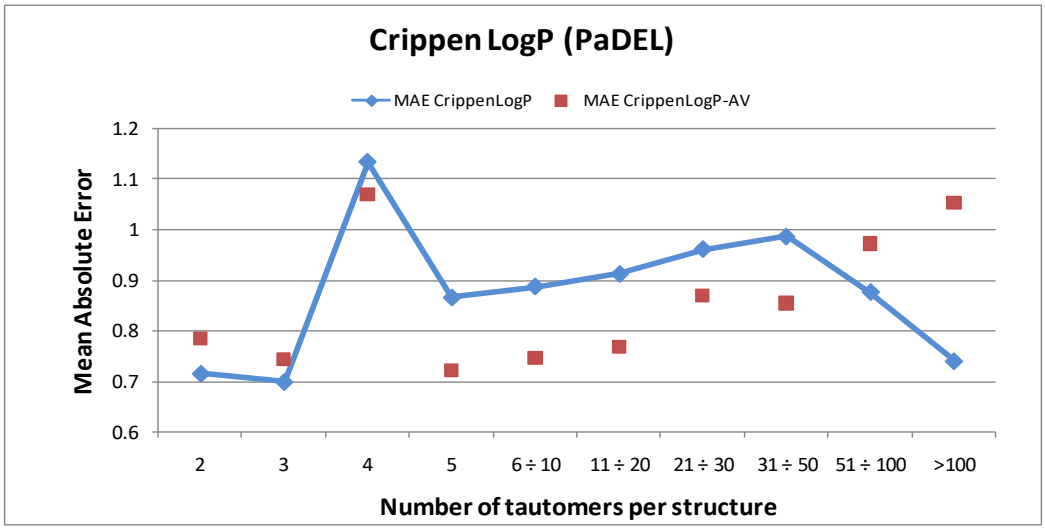


Fig. 3. Mean absolute error (MAE) of the Crippen LogP model (lines connected) and its weighted model modification across all tautomeric forms.

Discussion

The redistribution of double bonds alters the topological structure, making the 2D (topological) structure the primary representation level affected by tautomerism. As a result, tautomerism also impacts the 3D structure due to changes in atom hybridisation and bond orders. Some 1D descriptors are influenced as well, since these constitutional descriptors incorporate partial topological information (e.g., counts of specific small groups). Generally, no significant effect of tautomerism is expected on 0D descriptors, as these are purely constitutional and do not consider any topological information. It is worth noting that the classification of 0D and 1D descriptors can vary between software packages, and some literature sources do not differentiate between these levels at all (Fig. 4).

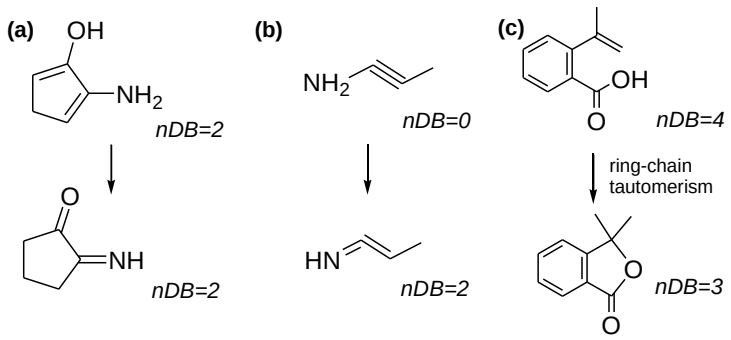


Fig. 4. Variation of constitutional-0D descriptor, nDB, for different types of tautomeric transformations.

Typically, constitutional 0D descriptors - such as nO (number of oxygen atoms), nC (number of carbon atoms), MW (molecular weight), nSB (number of single bonds), and nDB (number of double bonds) - are not affected by tautomerism, as they represent atom counts or sums of elemental properties without considering molecular topology. Usually, the nDB descriptor remains unchanged during tautomerism because the total number of double bonds is conserved despite their redistribution. However, if the tautomer transformations involve a triple-bonded carbon converting to an allene-like structure (with two double bonds), some constitutional 0D descriptors - like nDB and nTB (number of triple bonds) - can change (see Fig. 4b). Similar changes may occur in more complex ring-chain tautomerism cases (see Fig. 4c). Ring-chain tautomerism poses a significant challenge and is typically not handled systematically by most software tools. Excluding ring-chain transformations, the impact of tautomerism on constitutional 0D descriptors can generally be considered negligible.

Impact of tautomerism on fingerprint calculation

Molecular fingerprints are bit-vectors that compactly encode information from a molecule's topological structure. They are widely used in key cheminformatics tasks such as database screening, similarity searching, QSAR/QSPR modelling, and identifying biological activity cliffs and ecotoxicity anomalies. Fingerprints can be generated using various computational techniques, including key(fragment)-based methods, hashed fingerprints, or by binning continuous descriptor values into discrete intervals, each assigned to a specific bit. The primary advantage of fingerprint-based approaches is their exceptional speed; as bitwise operations are among the fastest computations on a computer. This efficiency enables rapid processing of very large chemical datasets, which explains their widespread use. Table 3 illustrates the significant impact that tautomerism can have on fingerprint values. The primary reason for this outcome aligns with previous discussions: tautomeric forms significantly alter the topological representation. For example, the three tautomers of methimazole (see Table 3) have CDK hashed fingerprints with 56, 68, and 57 bits set to 1, respectively, out of 1024 bits based on molecular graph paths up to eight atoms in length. Moreover, the first and second tautomers share only 10 common bits, the first and third share 10 as well, and the second and third share just 8. In this case, more than 80% of the fingerprint bits are distinct across the tautomers. Such variation can significantly affect structure similarity measures and, consequently, the outcomes of similarity searches. Most machine learning approaches for QSAR / QSPR modelling—including modern generative AI—rely on the "similarity principle", which assumes that structurally similar molecules exhibit similar biological activities and toxicities. For instance, QSAR models based on the k-nearest neighbors (kNN) method depend heavily on a defined similarity measure.

Impact of tautomerism on QSAR/QSPR modelling

In the previous sections, we demonstrated that tautomerism affects all fundamental cheminformatics tasks that precede QSAR/QSPR modelling, including topological (2D) and geometrical (3D) structure representations, descriptor calculations, and fingerprint generation. Since these foundational steps are significantly influenced by tautomerism, it is reasonable to expect that the final QSAR/QSPR model outputs will be affected as well. Table 4 illustrates the variance in model predictions across different tautomeric forms when tautomerism is not accounted for during model construction, highlighting this variability as a post-modelling artefact due to tautomerism. Although this represents a post-modelling scenario (i.e., tautomers were not included during model training or validation), the results in Table 4 still offer valuable insight into the influence of tautomerism and suggest potential strategies for refinement. One such example is shown in Fig. 3, where applying a post hoc average of model predictions across tautomeric forms leads to an improvement over the original model output.

It can be reasonably expected that accounting for tautomerism during model development would improve ecotoxicity prediction. This hypothesis was confirmed in our study through the development of a QSAR model for aquatic toxicity against Tetrahymena pyriformis. We created two model variations: a conventional model (CM), which does not consider tautomers, and a weighted model (WM), which uses descriptors averaged

over all tautomeric forms of the molecule. Both models share 20 common descriptors (see Table 2), with the CM including 7 unique descriptors and the WM including 4 unique ones. This preliminary descriptor analysis indicates that tautomerism influences the descriptor selection process as well. Additionally, Table 2 highlights the concept of "cascading modelling," where the Crippen-LogP QSPR model is selected as an input descriptor for the ecotoxicity QSAR model. This further highlights the importance of accurate and tautomer-aware modelling to build reliable predictive systems and mitigate the risk of cascading error propagation.

Fig. 5 demonstrates how incorporating tautomeric information during model construction enhances predictive performance. QSAR models for *Tetrahymena pyriformis* toxicity were evaluated using both 5-fold cross-validation (denoted as CV5 in the figure) and an external validation set (Ext-Val Set). In both cases, the weighted model (WM), which accounts for tautomerism, outperformed the conventional model (CM), which does not.

On the external validation set, the WM model achieved a higher correlation coefficient (0.75 vs. 0.70) and a lower mean absolute error (MAE) (0.44 vs. 0.51) compared to the CM model. Similar improvements were observed in the cross-validation results, with the correlation coefficient increasing from 0.79 to 0.82 and the MAE decreasing from 0.48 to 0.44. These results support the benefit of including tautomeric variability in model development.

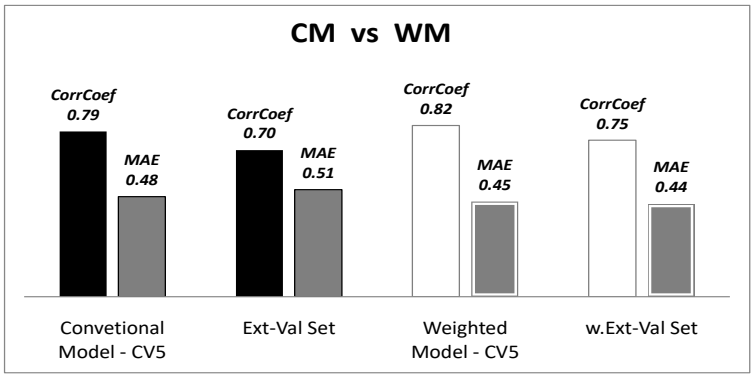


Fig. 5. Statistics comparison for Conventional Model (CM) and Weighted Model (WM) for Aquatic toxicity against *Tetrahymena pyriformis*.

Conclusions

Tautomerism affects molecular descriptors and, in turn, the accuracy of QSAR/QSPR models, as shown in this study. There is currently no systematic approach for integrating tautomers into QSAR/QSPR modelling, and OECD guidelines offer no recommendations for incorporating tautomeric information in ecotoxicity or chemical property assessments.

We propose two practical strategies for addressing tautomerism in QSAR/QSPR modeling: (1) post hoc weighting of model outputs based on tautomeric forms, as shown in Ames mutagenicity and XLogP models; and (2) use of weighted molecular descriptors, demonstrated in our aquatic toxicity model.

Ultimately, this work highlights how rigorous tautomer enumeration and its informed application across data processing, predictive modelling, and regulatory frameworks can enhance chemical safety assessments, particularly for ecotoxicity endpoints. It also supports the advancement of digital chemical innovation in alignment with EU policy objectives and sustainable chemistry initiatives.

Supplementary materials

All datasets and modelling data are available at the following Zenodo repositories: https://doi.org/10.5281/zenodo.16421375; https://doi.org/10.5281/zenodo.16421491; https://doi.org/10.5281/zenodo.16420004 and https://doi.org/10.5281/zenodo.16441958.

Acknowledgments

This study is financed by the European Union-NextGenerationEU, through the National Recovery and Resilience Plan of the Republic of Bulgaria, project № BG-RRP-2.004-0001-C01

References

Caldeira, C., Garmendia, A.I., Tosches, D., Mancini, L., Abbate, E., Farcal, R., Lipsa, D., Rasmussen, K., Rauscher, H., Riego, S.J., & Sala, S. (2023). *Safe and Sustainable by Design chemicals and materials - Application of the SSbD framework to case studies*. European Commission, JRC Technical Report, 201. doi: 10.2760/487955

European Parliament and Council. (2006). Regulation (EC) No. 1907/2006 of the European Parliament and of the Council of 18 December 2006 concerning the Registration, Evaluation, Authorisation and Restriction of Chemicals (REACH) (L 396, pp. 1–849). Official Journal of the European Union. Retrieved from https://eur-lex.europa.eu/eli/reg/2006/1907/oj/eng

Frank, E., Hall, M.A., & Witten, I. H. (2016). *The WEKA Workbench. Online Appendix for "Data Mining: Practical Machine Learning Tools and Techniques", 4th Edition*. Morgan Kaufmann, Burlington.

Knox, C., Law, V., Jewison, T., Liu, P., Frolkis, A., Pon, A., Banco, K., Mak, C., Neveu, V., Djoumbou, Y., Eisner, R., Guo, A., & Wishart, D. (2011). DrugBank 3.0: a comprehensive resource for "omics" research on drugs. *Nucleic Acids Research*, 39, 1035-1041. doi: 10.1093/nar/gkq1126

Kochev, N.T., Paskaleva, V.H., & Jeliazkova, N. (2013). Ambit-tautomer: An open source tool for tautomer generation. *Molecular Informatics*, 32(5–6), 481–504. doi: 10.1002/minf.201200133

NCI database. (n.d.). Retrieved October 29, 2012, from http://cactus.nci.nih.gov/download/nci/

Organisation for Economic Co-operation and Development. (2015). *Harmonised templates for reporting chemical test summaries (OHTs) (OECD Environment, Health and Safety Publications, Series on Testing and Assessment No. 25)*. OECD Publishing. Retrieved from https://www.oecd.org/en/topics/sub-issues/assessment-of-chemicals/harmonised-templates.html

Steinbeck, C., Hoppe, C., Kuhn, S., Floris, M., Guha, R., & Willighagen, E.L. (2006). Recent Developments of the Chemistry Development Kit (CDK) - An Open-Source Java Library for Chemo- and Bioinformatics. *Current Pharmaceutical Design*, 12(17), 2111–2120. doi: 10.2174/138161206777585274

Wang, R., Fu, Y., & Lai, L. (1997). A new atom-additive method for calculating partition coefficients. *Journal of Chemical Information and Computer Sciences*, 37(3), 615–621. doi: 10.1021/ci960169p

Warr, W.A. (2010). Tautomerism in chemical information management systems. *Journal of Computer-Aided Molecular Design*, 24(6–7), 497–520. doi: 10.1007/s10822-010-9338-4

Weininger, D. (1988). SMILES, a Chemical Language and Information System: 1. Introduction to Methodology and Encoding Rules. *Journal of Chemical Information and Computer Sciences*, 28(1), 31–36. doi: 10.1021/ci00057a005

Yap, C.W. (2011). PaDEL-descriptor: An open source software to calculate molecular descriptors and fingerprints. *Journal of Computational Chemistry*, 32(7), 1466–1474. doi: 10.1002/jcc.21707

Received: 28.05.2025
Accepted: 29.06.2025