IOCXe 2026 Conference
The 1st International Online Conference on Xenobiotics
MDPI
22–23 June 2026 | Online

QSRR Modeling for GC-MS Retention Prediction of Emerging Xenobiotic Contaminants:

A Robust In Silico Tool for Environmental Screening and Risk Assessment

Nabil Bouarra[1,2*] · Soumaya Kherouf[3] · Loubna Nouri[1] · Adhya-eddine Hamitouche[1] · Khaldoun Bachari[1]

[1] CRAPC, Industrial Zone, Tipaza 42004, Algeria [2] Lab. Environmental Engineering, Badji Mokhtar Univ., Annaba [3] LOMOP, Badji Mokhtar University, Annaba, Algeria

bouarra.nabil@crapc.dz | ** Corresponding author*

INTRODUCTION & AIM

ENVIRONMENTAL CONTAMINATION CONTEXT

EMISSION SOURCES
Pesticides
Pharmaceuticals
Industrial chemicals
Personal care products
~250,000+ known xenobiotics

release

ENV. MATRICES
Surface water
Groundwater & Soil
Sediments
Biota & Food chain
Complex multi-matrix

exposure

HEALTH & ECO RISKS
Endocrine disruption
Carcinogenicity
Reproductive toxicity
Ecosystem collapse
Urgent monitoring needed

X Traditional Monitoring	✓ QSRR In Silico Approach
Costly reference standards required	No reference standards needed
Lengthy laboratory procedures	Rapid computational prediction
Limited to known compound libraries	Applicable to any novel structure
Not scalable for thousands of compounds	Fully scalable — 382 compounds & beyond
New contaminants remain undetected	Supports environmental risk assessment

VS

DATASET — 382 Xenobiotic Compounds (GC-MS Retention Times)

382 Total compounds

269 — Training Set (70.4%)
113 — Test (29.6%)

128	94	73	87
Pesticides	Pharmaceuticals	Industrial Chem.	Other Emerging

WHY QSRR? — KEY ADVANTAGES

- **Structure-Based** — Links molecular features to RT directly
- **Rapid & Scalable** — Thousands of compounds in seconds
- **Universal** — Any novel xenobiotic structure
- **Rigorously Validated** — OECD guidelines int. + ext. validation

RESEARCH OBJECTIVE

Develop a robust MLR-QSRR model using Genetic Algorithm (GA-VSS) descriptor selection to predict GC-MS retention times of 382 structurally diverse xenobiotic compounds supporting rapid environmental screening and risk assessment

METHOD

COMPUTATIONAL METHODOLOGY PIPELINE

1. **ChemDraw** V7.0 — 2D Structure Drawing → Output: SDF / MOL files
2. **MOPAC** 2016 — 3D Geometry Optimization via PM7 → Output: Optimized 3D structures
3. **ALVADESC** V3.0 — Descriptor Calculation → Output: ~5000+ descriptors
4. **GA-VSS** QSARINS V2.2 — Feature Selection → Output: 5 optimal descriptors
5. **MLR + Validation** — Model Build & Validate → Validated QSRR Model

PM7 SEMI-EMPIRICAL METHOD — 3D GEOMETRY OPTIMIZATION (MOPAC 2016)

- **F — Fast**: 100x faster than DFT Ab initio methods. Ideal for large datasets
- **A — Accurate**: Reliable geometries for structurally diverse mols. Validated vs experimental
- **P — Parametrized**: Optimized for organic & environmental mols. Including heteroatoms
- **S — Scalable Batch**: Applied to all 382 compounds systematically. Automated workflow

GENETIC ALGORITHM — VARIABLE SUBSET SELECTION (GA-VSS) — QSARINS V2.2

Initialize Population — Random descriptor subsets from ~5000+ candidates. Population size: 100
→ **Evaluate Fitness** — LOO cross-validation Maximize Q^2LOO score. + penalize collinearity
→ **Crossover + Mutation** — Evolve toward better descriptor subsets. Generations: 200+
→ **Convergence** — 5 optimal descriptors DBI, R8i, TPSA_eff, Hy, SM1_B(p). Q^2LOO = 0.877

5 SELECTED DESCRIPTORS — FULL NAME & PHYSICAL INTERPRETATION

DBI	R8i	TPSA_eff	Hy	SM1_B(p)
Dragon Branching Index	R Autocorrelation lag 8 / Ionization Potential	Polar Surface Area / Molecular Size Ratio	Hydrophilic Factor (Todeschini)	Spectral Moment order 1 Burden matrix / polarisab.
Mol. branching ↓ contact → ↓ RRT	Distant e- topology ↓ polarisab. → ↓ RRT	Surface polarity density ↓ volatility → ↑ RRT	H-bond self-association ↓ volatility → ↑ RRT	Atomic polarisability ↑ dispersion → ↑ RRT
coeff = -0.149	coeff = -0.073	coeff = +0.720 ★	coeff = +0.127	coeff = +1.979 ★★

RIGOROUS VALIDATION STRATEGY — OECD QSAR GUIDELINES

Internal Validation	External Validation	Y-Scrambling	Applicability Domain
R^2 = 0.883	Q^2F1=0.812 Q^2F2=0.810	R^2Yscr ≈ 0 (avg ~0.003)	Williams plot analysis
Q^2LOO = 0.877	Q^2F3=0.837	Q^2Yscr ≈ 0 (avg -0.025)	Leverage threshold h* = 0.067
RMSEtr = 0.117, s = 0.118	CCCext=0.893 RMSEext=0.138	No chance correlation	Identifies structural outliers
Training set (269 cpds)	Test set (113 compounds)	100 scrambling iterations	Majority within AD

QSRR MODEL EQUATION | ChemDraw · MOPAC 2016 (PM7) · ALVADESC V3 · QSARINS V2.2

$$Tr = -4.19 - 0.149 \cdot DBI - 0.073 \cdot R8i + 0.127 \cdot Hy + 0.720 \cdot TPSA\_eff + 1.979 \cdot SM1\_B(p)$$

RESULTS & DISCUSSION

Internal Validation				External Validation				
0.883	0.877	0.117	0.118	0.812	0.810	0.837	0.893	0.138
R^2	Q^2LOO	RMSEtr	s	Q^2F1	Q^2F2	Q^2F3	CCCext	RMSEext

DIAGNOSTIC PLOTS

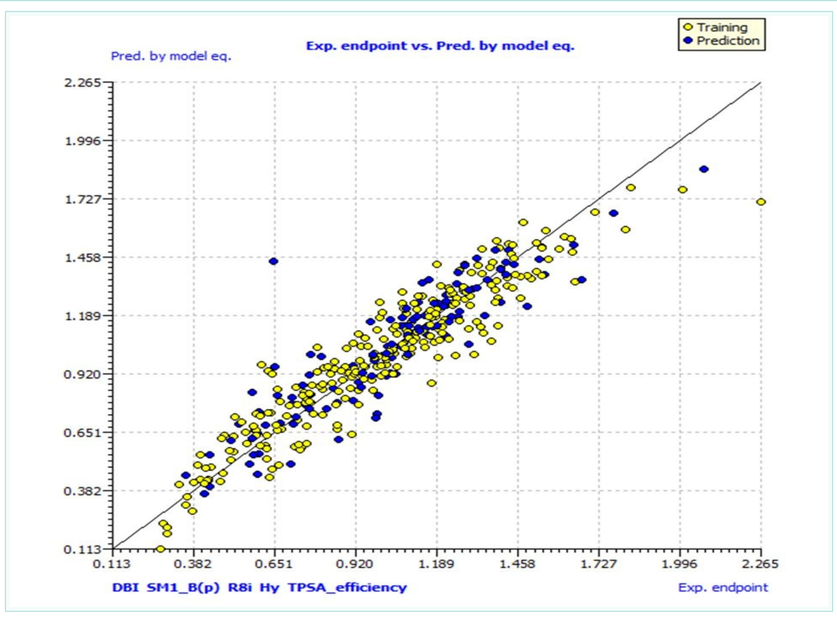

Fig. 1 — Experimental vs. Predicted Endpoint

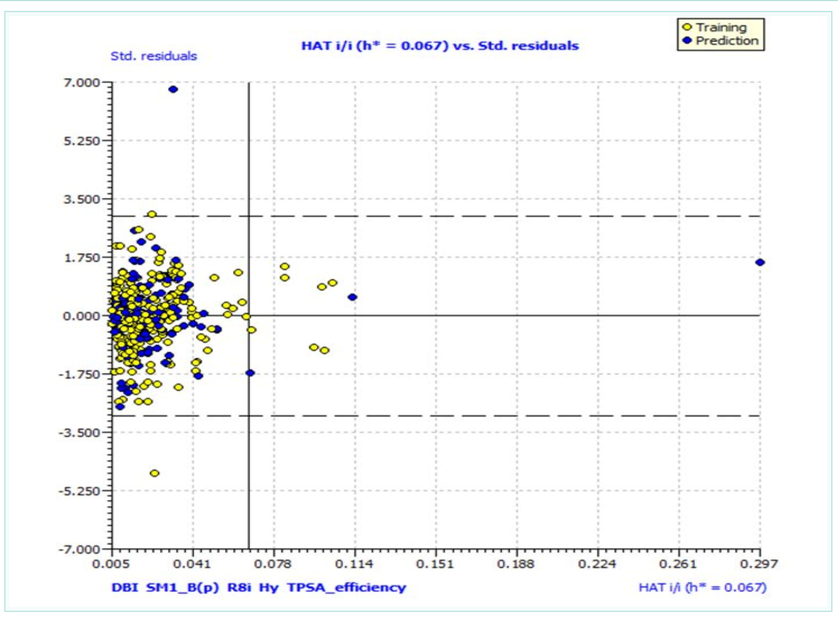

Fig. 2 — Williams Plot: Applicability Domain (h=0.067)*

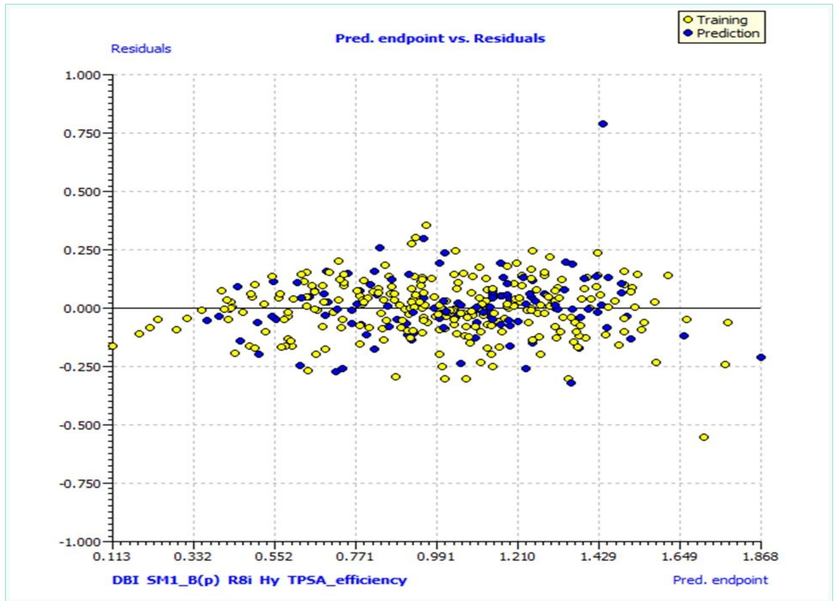

Fig. 3 — Predicted Endpoint vs. Residuals

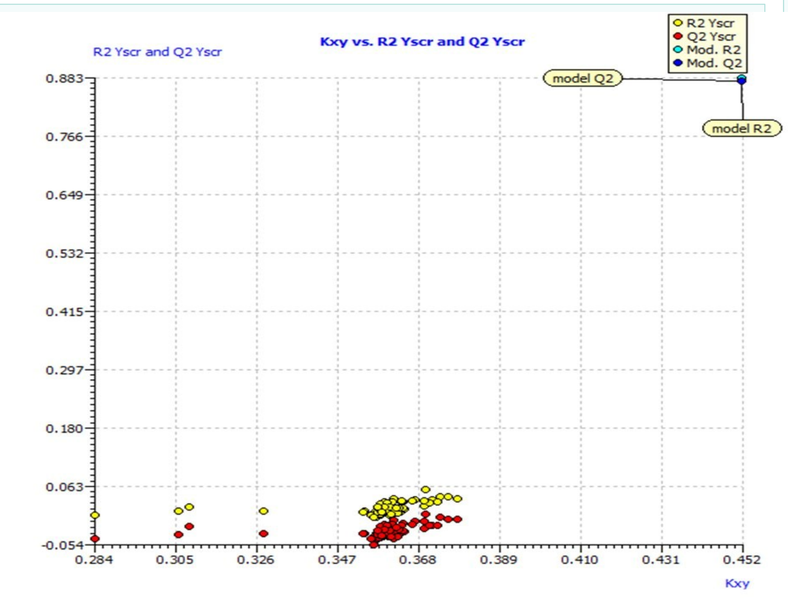

Fig. 4 — Y-Scrambling (Kxy vs R^2/Q^2 Yscr)

DESCRIPTOR STATISTICS

Descriptor	Coefficient	SE Coef	T-Value	P-Value	VIF	Physical Interpretation
Constant	−4.188	0.129	−32.41	<0.001	—	Intercept
DBI	−0.149	0.012	−12.00	<0.001	1.97	Mol. branching → ↓ contact → ↓ RRT
R8i	−0.073	0.018	−3.98	<0.001	1.44	Distant e⁻ topology → ↓ polarisab. → ↓ RRT
Hy	+0.127	0.022	5.80	<0.001	1.64	H-bond self-assoc. → ↓ volatility → ↑ RRT
TPSA_eff	+0.720	0.088	8.23	<0.001	1.78	Surface polarity density → ↓ volatility → ↑ RRT
SM1_B(p)	+1.979	0.050	39.58	<0.001	2.43	Global atomic polarisab. → ↑ dispersion → ↑ RRT

CONCLUSION

- ✓ A robust 5-descriptor MLR-QSRR model was successfully developed for GC-MS retention time prediction of 382 structurally diverse xenobiotic compounds, including pesticides, pharmaceuticals and industrial chemicals.
- ✓ GA-VSS identified the optimal descriptor set: DBI, R8i, TPSA_efficiency, Hy, SM1_B(p) — encoding molecular branching, electronic topology, surface polarity density, hydrophilicity, and atomic polarisability.
- ✓ Exceptional statistical performance: R^2=0.883, Q^2LOO=0.877 (internal); Q^2F1–$F3$ >0.81, CCCext=0.893, RMSEext=0.138 (external) — well above OECD acceptance thresholds.
- ✓ Y-scrambling confirms no chance correlation. Williams plot defines the applicability domain. Model is fully compliant with OECD QSAR validation principles.
- → This in silico tool enables rapid, cost-effective screening of emerging xenobiotics in complex environmental matrices, directly supporting exposure assessment and risk evaluation workflows.

FUTURE WORK / REFERENCES

Future Work:

- Extend to broader chemical spaces: pharmaceuticals, pesticides, industrial contaminants
- Apply non-linear ML (RF, XGBoost, SVM) for improved accuracy
- Integrate predictions with environmental fate models for holistic risk frameworks
- Explore multi-endpoint QSRR for simultaneous GC parameter prediction

References:

[1] Stan,H.J. J Chromatogr. A, 892, 347–377 (2000).
[2] Gramatica, P et al,. J Comput Chem.,34, 2121–2132 (2013)
[3] MAURI, A. alvaDesc, New York, NY : Springer US, 2020. p. 801-820.

KEY PERFORMANCE INDICATORS

382	R^2=0.883	CCCext=0.893	5 Desc.	MOPAC
Xenobiotic Compounds	Goodness of Fit	Lin's CCC	GA-Optimal Set	Semi-emp. Method
Diverse dataset: pesticides, pharmaceuticals, industrials	Excellent fit on training data (internal validation)	Strong predicted/measured agreement on test set	DBI · R8i · TPSA_eff Hy · SM1_B(p)	PM7 via MOPAC 2016 for all 382 compounds