Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

PhytoExtractQSAR: An Automated Pipeline for Literature-Mined Modeling of Phytochemical Extraction Outcomes with Transparent Generalization Assessment

Sharhabil Eltahir

igsr.sharhabil@alexu.edu.eg

Alexandria University

Research Article

Keywords: QSAR, phytochemical extraction, machine learning, literature mining, cheminformatics, RDKit, natural products, cross-validation, generalization, applicability domain, data provenance

Posted Date: March 9th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-8973111/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Abstract

Background

Predicting phytochemical extraction outcomes from molecular descriptors and process parameters remains challenging due to fragmented literature data and inconsistent reporting. No centralized database exists for extraction outcomes, and existing QSAR approaches rely on small, manually curated datasets that limit generalizability assessment.

Results

We developed PhytoExtractQSAR, an automated cheminformatics pipeline that integrates PubMed literature mining, PDF data extraction, molecular descriptor computation via RDKit, and machine learning with transparent data provenance tracking. The pipeline assembled 1,877 literature-verified records across 94 compounds and 15 extraction methods. After removing 188 flagged duplicate records, the deduplicated dataset was used for all modeling. Under record-level nested five-fold cross-validation, the best models achieved $Q^2 = 0.448$ for total flavonoid content (TFC; Extra Trees, N = 991), $Q^2 = 0.281$ for crude extraction yield (Extra Trees, N = 1,857), and exploratory associations for total phenolic content (TPC; $Q^2 = 0.089$) and antioxidant activity IC50 ($Q^2 = 0.145$). Crucially, leave-one-compound-out (LOCO) cross-validation produced $Q^2 \leq 0$ for all targets, demonstrating that the current models capture within-compound extraction patterns but do not generalize to unseen compounds. A process-only baseline ($Q^2 = 0.154$ for yield) confirmed that molecular descriptors add modest predictive value beyond extraction parameters ($\Delta Q^2 \approx 0.07$), functioning as study-context proxies. SHAP analysis confirmed that compound identity—proxied by LogP and molecular weight—dominates predictions, explaining the generalization gap. Leverage-based applicability domain analysis confirmed > 93% of samples fall within reliable physicochemical boundaries, yet this physicochemical coverage proved insufficient for cross-compound prediction.

Conclusions

PhytoExtractQSAR establishes an open, reproducible framework for automated extraction data assembly with rigorous provenance tracking. The pipeline successfully predicts extraction outcomes for compounds represented in the training data but does not yet achieve cross-compound generalization. A process-only baseline model ($Q^2 = 0.154$ for yield using only extraction parameters) confirmed that molecular descriptors contribute modestly beyond process conditions (full model $Q^2 = 0.281$), functioning primarily as study-context proxies rather than encoding causal structure–property relationships. The LOCO analysis provides the first quantitative evidence of this generalization gap in extraction QSAR. Compound-grouped versus matrix-grouped cross-validation revealed that generalization failure is driven by compound identity memorization rather than plant matrix variation,

identifying plant material parameterization and richer structural representations as primary directions for transferable models.

Scientific contribution:

This work provides the first open-source pipeline automating the full workflow from PubMed mining through QSAR modeling of phytochemical extraction, assembling 1,877 provenance-tracked records across 94 compounds. The systematic comparison of record-level, compound-grouped, and matrix-grouped cross-validation, together with process-only baseline experiments, reveals that molecular descriptors function primarily as compound identity proxies and are insufficient for cross-compound generalization. These findings establish an important benchmark, demonstrate that plant material parameterization rather than richer molecular fingerprints is the primary pathway to transferable extraction models, and identify concrete directions including solid-to-liquid ratio capture and temperature-corrected solvent properties.

Introduction

The extraction of bioactive phytochemicals from plant materials is a foundational step in pharmaceutical, nutraceutical, and food science research [1, 2]. Extraction outcomes—including crude yield (mass of total extract relative to starting plant material, %), total phenolic content (TPC), total flavonoid content (TFC), and antioxidant activity—depend on complex interactions between the molecular properties of target compounds, solvent characteristics, plant matrix composition, and process parameters such as temperature, time, solid-to-liquid ratio, and energy input [3]. Optimizing these multi-dimensional processes traditionally requires extensive experimental design, often guided by response surface methodology (RSM) [4]. However, each RSM study addresses a single compound–method–matrix combination, and the accumulated knowledge across thousands of published studies remains fragmented and difficult to leverage systematically.

Quantitative Structure–Activity Relationship (QSAR) modeling offers a computational framework for predicting molecular properties from structural descriptors [5, 6]. While QSAR has been applied extensively in drug discovery for predicting biological activity, toxicity, and ADMET properties [7, 8], its application to phytochemical extraction optimization remains limited. The few existing studies typically rely on small datasets (fewer than 100 compounds) with manually curated parameters [9, 10], restricting both the chemical diversity covered and the ability to assess whether models generalize beyond the training compounds.

Building QSAR models for extraction outcomes faces several challenges beyond data availability. Unlike drug–target interactions, extraction outcomes such as crude yield and TPC are bulk matrix properties influenced by plant tissue composition (cellulose, lignin, interfering compounds) and anatomy (leaf vs. bark vs. seed), not solely by the target compound's molecular structure [3]. Crude extraction yield represents the total mass of soluble material extracted from the plant matrix, dominated by abundant

structural polymers (cellulose, hemicellulose, lignin, sugars, proteins) that may constitute > 99% of the extract mass. A single target molecule's physicochemical descriptors bear no direct causal relationship to bulk yield—any predictive signal must arise indirectly through correlation with study-level context. This input–output mismatch must be considered when interpreting model performance. This creates a critical question: can molecular descriptors of a single target phytochemical serve as useful predictive features for extraction outcomes, or do they merely serve as proxies for study design choices? Addressing this question requires datasets large enough to support rigorous cross-compound validation and, critically, process-only baseline models to disentangle these contributions.

No centralized repository exists for phytochemical extraction data, unlike drug–target interaction databases such as ChEMBL [11] or PubChem BioAssay [12]. Relevant experimental parameters are scattered across thousands of journal articles, often reported inconsistently in terms of units, measurement protocols, and extraction conditions [13]. Recent advances in natural language processing and automated literature mining now enable large-scale extraction of structured data from scientific publications [14, 15], presenting an opportunity to systematically harvest extraction data from the published literature.

Here we present PhytoExtractQSAR, an end-to-end cheminformatics pipeline that automates (1) mining phytochemical extraction data from PubMed abstracts and full-text PDFs, (2) computing molecular descriptors from validated SMILES representations, (3) engineering physicochemically meaningful features including solvent polarity descriptors and compound–method interaction terms, and (4) training and validating multi-target models with transparent data provenance tracking. A key contribution is the systematic evaluation of model generalizability through both record-level and compound-grouped cross-validation strategies, including leave-one-compound-out (LOCO) analysis. This dual evaluation reveals the extent to which current models capture genuine structure–extraction relationships versus compound-specific memorization, providing the first quantitative assessment of the generalization gap in extraction-focused QSAR [26].

Methods

Data collection and literature mining

The data collection pipeline consisted of three complementary mining strategies targeting phytochemical extraction studies indexed in PubMed. A systematic set of 38 targeted queries was designed to maximize recovery of studies reporting quantitative extraction outcomes (yield, TPC, TFC, IC50) with explicit process parameters. Queries captured compound-specific data (e.g., "quercetin AND extraction AND yield AND optimization"), method-specific data (e.g., "ultrasound-assisted extraction AND yield AND optimization AND RSM"), and outcome-specific data (e.g., "total flavonoid content AND mg QE AND extraction"). The Biopython Entrez API [17] was used to retrieve up to 200 records per query, yielding 3,322 unique PubMed records after deduplication.

Full-text PDFs were obtained for accessible articles using DOI-based retrieval from open-access repositories. A total of 810 PDFs were processed using a custom regex-based extraction pipeline that identified numerical values for extraction yield (%), TPC (mg GAE/g), TFC (mg QE/g or mg CE/g), IC50 (µg/mL), temperature (°C), time (min), pressure (MPa), and solvent system from both abstracts and full text. To address potential unit heterogeneity in TFC reporting, each record was systematically annotated with its reporting standard (QE or CE). Upon examination, all TFC values in the mined dataset were reported in Quercetin Equivalents; no Catechin Equivalent records were identified, eliminating the unit-mixing concern. Yield values in this dataset refer to crude extraction yield, defined as the mass of total dried extract relative to the starting plant material mass (%), as reported in the source publications. This is a bulk matrix property, not the recovery of a specific target compound.

Extracted values underwent automated quality control with physically realistic range filters: temperature [– 20, 300] °C, pressure [0, 100] MPa, yield [0.01, 100] %, TPC [0.01, 1000] mg GAE/g, TFC [0.01, 1000] mg QE/g, and IC50 [0.001, 10000] µg/mL. Unit conversions were applied where detected (e.g., hours to minutes, kW to W, bar to MPa). The initial literature-verified dataset comprised 1,877 records covering 94 phytochemical compounds across 15 extraction methods.

Several parameters of known importance to extraction outcomes were not systematically captured by the initial regex pipeline. Solid-to-liquid ratio (SLR), which governs mass transfer equilibrium through the concentration gradient driving force (Fick's first law), is reported in heterogeneous formats (w/v, g/mL, or ratios such as "1:10") that resist reliable automated extraction. In a subsequent refinement step, SLR was annotated for 10,000 records using regex-based parsing and manual verification, yielding SLR_mL_per_g values incorporated as an additional feature. Similarly, compound purity metadata (e.g., HPLC purity), plant part (leaf, bark, root, seed), and specific antioxidant assay type (DPPH, ABTS, FRAP) were not extracted. These omissions are discussed as limitations below.

Duplicate detection and removal

Potential duplicate records were identified based on compound–method–parameter combinations. A total of 672 records were flagged as potential duplicates in the initial 1,877-record dataset. To prevent artificial inflation of cross-validation metrics through train–test leakage of near-identical records, 188 confirmed duplicates were removed prior to modeling. The final deduplicated dataset comprised 1,689 literature-verified records, which served as the basis for all models reported herein. Per-target sample sizes after deduplication were: yield N = 1,857, TFC N = 991, TPC N = 1,201, IC50 N = 1,309.

Data provenance and quality tracking

A distinguishing feature of this pipeline is explicit provenance tracking for every record. Each entry carries a _data_source field indicating whether it was derived from PubMed abstract mining, PDF full-text extraction, or computational augmentation. All QSAR models were trained and validated exclusively on

literature-verified records, using nested cross-validation to prevent information leakage. Early experiments in which models were trained on computationally augmented data produced $Q^2 \approx 0$ across all targets, confirming the necessity of provenance-based data separation.

Quality control was performed in three stages. Structural validation ensured all 94 phytochemical SMILES were parseable by RDKit [16] and matched PubChem canonical representations verified by CID lookup. Parametric validation applied range filters and unit-consistency checks. Duplicate detection and removal were performed as described above.

Molecular descriptor computation

Molecular descriptors were computed using RDKit (version 2023.09) from PubChem-validated isomeric SMILES for all 94 compounds. Ten descriptors were selected based on their established relevance to solubility, permeability, and extraction behavior: molecular weight (MW), calculated octanol–water partition coefficient (ALogP), topological polar surface area (TPSA), hydrogen bond donor count (HBD), hydrogen bond acceptor count (HBA), rotatable bond count, aromatic ring count, heavy atom count, fraction of sp3 carbons (Fsp3), and molar refractivity.

It is important to note that these single-compound descriptors serve as proxy features for the phytochemical context of each study. For example, a record associated with quercetin's molecular descriptors reflects a study that focused on quercetin-rich plant materials and extraction conditions optimized for flavonoids. The descriptors thus encode study-level context rather than providing a direct causal link between a single molecule's structure and the bulk extraction outcome.

Feature engineering

Beyond raw molecular descriptors and extraction parameters, 31 engineered features were computed to encode domain-specific physicochemical knowledge, grouped into six categories:

Solvent property descriptors (7 + 5 features): Polarity index (Snyder scale), dielectric constant, boiling point, viscosity, and Hansen solubility parameters (δD, δP, δH) were assigned from published reference tables [18] at standard conditions (25°C). For binary mixtures, properties were computed as composition-weighted averages. In the revised analysis, temperature-corrected solvent properties were additionally computed for the three major solvents (water, ethanol, methanol) using established empirical equations for dielectric constant and viscosity. Five additional features were generated: solvent_dielectric_tcorr, solvent_viscosity_tcorr, solvent_polarity_tcorr, solvent_reduced_temp (T_process / T_boiling), and solvent_above_boiling (binary flag). For solvents without reliable temperature-dependent data, 25°C reference values were retained.

Extraction method descriptors (5 features): Each of the 15 methods was characterized by ordinal scores (1–5) for energy input, mass transfer efficiency, selectivity, and binary flags for conventional vs. green

classification based on a fixed list from Chemat et al. [19].

Phytochemical class descriptors (4 features): Ordinal scores for typical polarity, molecular weight range, thermal stability, and aqueous solubility.

Interaction terms (9 features): Physicochemically motivated cross-terms including thermal exposure, LogP–solvent compatibility, TPSA–polarity matching, H-bond–solvent matching, total energy delivered, and others.

Polynomial terms (4 features): Squared terms for temperature, time, LogP, and molecular weight.

Composite index (1 feature): A normalized extraction intensity score on [0, 1].

QSAR modeling and validation

Ten regression algorithms were evaluated: Ridge, Lasso, ElasticNet, K-nearest neighbors, SVR (RBF and linear kernels), Random Forest, Gradient Boosting, Extra Trees, XGBoost [20], and multi-layer perceptron. All models were trained on the deduplicated literature-verified dataset. Molecular descriptors were complete for all records. Missing extraction process parameters were imputed with physically appropriate defaults based on method class: pressure was set to ambient (0.1 MPa) for non-pressurized methods, and power was set to zero for passive methods. For active methods where power was not reported, values were retained as missing (Supplementary Table S1). Other numeric features were standardized. Model performance was assessed using Q^2, CV $R^2 \pm$ SD, RMSE, and Pearson r, with Q^2 as the primary metric [21].

Cross-validation strategies

Three cross-validation strategies were employed to assess different aspects of model validity:

Record-level nested CV (primary): Nested five-fold with inner hyperparameter optimization and outer unbiased performance estimation. Records split randomly regardless of compound identity, meaning the same compound could appear in both training and test folds. This evaluates prediction for known compounds under new conditions.

GroupKFold by compound: Five-fold CV with folds stratified so all records for a given compound appear exclusively in training or test. Evaluates within-fold generalization to unseen compounds.

Leave-one-compound-out (LOCO): Each of 29 compounds with sufficient data (≥ 10 records) was held out entirely. Provides the most stringent test of cross-compound generalization.

Applicability domain assessment

The AD was assessed using a leverage-based approach per OECD Principle 3 [26]. The Hat matrix was computed with warning threshold h* = 3p/n [26, 27]. Williams plots were generated for record-level CV predictions. PCA-based bounding box analysis used the first five principal components. Note that AD analysis assesses physicochemical descriptor space coverage but does not guarantee biological or mechanistic similarity for unseen compounds.

Feature importance analysis

SHAP (SHapley Additive exPlanations) [22] values were computed for the best-performing tree-based model per target using TreeExplainer with a subsample of 300 records.

Software and reproducibility

The complete pipeline was implemented in Python 3.12 using scikit-learn 1.5, XGBoost 3.2, SHAP 0.50, RDKit 2023.09, Biopython 1.86, and pandas 2.2. All source code, datasets (including provenance metadata and duplicate flags), cross-validation scripts (record-level, GroupKFold, LOCO), trained models, and analysis notebooks are available at [GitHub repository URL] (archived at [Zenodo DOI]) under an MIT license. Compound structures are provided as FAIR-compliant CSV and SDF files with PubChem CID cross-references (Additional File 1).

Results

Dataset characteristics

The automated mining pipeline assembled 1,877 literature-verified experimental records from 810 full-text PDFs and 3,322 PubMed abstracts, covering 94 unique phytochemical compounds across 15 extraction methods (Fig. 1). After removing 188 confirmed duplicates, the final modeling dataset comprised 1,689 unique records. The dataset spans flavonoids (28 compounds, 23.4%), phenolic acids (18, 14.9%), terpenes (14, 12.8%), alkaloids (12, 19.1%), and other classes (29.8%). The most represented methods were UAE (23.1%), maceration (18.7%), Soxhlet (14.2%), and MAE (11.8%). Data sources were 63.9% PDF extraction and 36.1% PubMed abstract mining.

Target variable availability varied: crude yield was reported in all records, TPC in 1,201 (71.1%), TFC in 991 (58.7%), and IC50 in 1,309 (77.5%). The dataset exhibits a pronounced class imbalance: flavonoids are associated with a disproportionate share of records due to their prevalence in the extraction literature. This skew has implications for model generalizability discussed below.

Feature engineering impact

The engineered feature set expanded the representation from 16 raw features to 41 features. Feature engineering had a target-dependent impact. For TFC, raw descriptors outperformed the expanded set (Q^2 = 0.448 vs. 0.361 on deduplicated data). For crude yield, the 41-feature set improved Q^2 from 0.221 to

0.278 on a matched subset, though the final best yield model used 16 features on the full deduplicated dataset ($Q^2 = 0.281$). The difference reflects confounded effects of dataset size and feature count. These findings demonstrate target-specific feature engineering benefits and that parsimonious models may be preferred when signal is concentrated in a few descriptors.

Record-level QSAR model performance

Table 3 reports best-performing models under record-level nested CV on the deduplicated dataset. TFC achieved $Q^2 = 0.448$ (Extra Trees, Pearson r = 0.670), crude yield $Q^2 = 0.281$ (Extra Trees, r = 0.530). TPC and IC50 showed weak associative signals ($Q^2 = 0.089$ and 0.145) insufficient for quantitative prediction, reported as exploratory results. Purity showed no signal ($Q^2 = -0.003$), confirming it is orthogonal to the current feature space—purity is determined by post-extraction purification rather than extraction parameters.

Ensemble tree methods consistently outperformed linear models and neural networks. The gap was most pronounced for TFC (Extra Trees $Q^2 = 0.448$ vs. Ridge $Q^2 = 0.065$). CV standard deviations (0.04–0.13) reflect moderate fold-to-fold variability.

Under record-level CV, the same compound appears in both training and test folds with different conditions. The model may therefore learn compound-specific patterns rather than generalizable structure–extraction relationships. This was investigated through compound-grouped validation.

Compound-grouped cross-validation and LOCO analysis

GroupKFold (by compound) and LOCO cross-validation were performed using the best algorithm per target (Table 4). Under both grouped strategies, all targets produced $Q^2 \leq 0$: GroupKFold Q^2 ranged from − 0.019 (TPC) to − 0.427 (yield), and LOCO Q^2 from − 0.015 (TPC) to − 0.243 (yield). These results demonstrate unequivocally that the current models cannot predict extraction outcomes for compounds absent from the training data.

The LOCO analysis reveals that record-level Q^2 values reflect within-compound interpolation: the model learns that quercetin at 60°C yields TFC similar to other quercetin records, but cannot extrapolate from quercetin's descriptors to predict TFC for kaempferol. This compound-memorization effect is consistent with SHAP showing LogP and MW as top features—these function as compound identity markers rather than generalizable predictors.

The performance collapse has a clear explanation: with 94 unique compounds and 10 molecular descriptors, the model has sufficient capacity to encode compound identity. Extraction parameters provide within-compound variation. Under LOCO, compound identity is removed and extraction parameters alone lack cross-compound predictive power. This finding underscores the need for plant material parameterization and complementary structural representations.

Applicability domain

Leverage-based AD analysis confirmed > 93% of record-level CV samples within AD boundaries: TFC 96.8%, yield 95.4%, TPC 94.1%, IC50 93.7% (Fig. 4). High-leverage outliers were concentrated among compounds with extreme LogP or rare extraction methods.

However, the LOCO analysis demonstrates that falling within the physicochemical AD does not guarantee accurate prediction for a new compound. AD analysis confirms descriptor-space coverage but cannot assess whether the model has learned transferable relationships—an important caveat.

Data scaling analysis

Under record-level CV, performance improved with data volume (Fig. 2): yield Q^2 from 0.199 (N ≈ 1,200) to 0.281 (N = 1,857); IC50 from 0.037 to 0.145. In light of LOCO results, this scaling likely reflects improved within-compound coverage rather than chemical diversity. Future data collection should prioritize expanding the number of unique compounds and capturing plant matrix descriptors.

SHAP feature importance

SHAP analysis revealed target-specific patterns (Fig. 3). For TFC, LogP (|SHAP| = 0.45), solvent polarity (0.38), and MW (0.32) dominated. For yield, temperature (0.52), method energy (0.45), and time (0.40) dominated. The dominance of raw molecular descriptors (LogP, MW) for TFC—constant per compound—is consistent with the LOCO finding that models encode compound identity. If generalizable relationships had been learned, interaction terms would dominate over raw descriptors.

For TPC, solvent polarity (0.35) and TPSA (0.30) led, consistent with TPC being a bulk extract measure. For IC50, LogP (0.30) and aromatic rings (0.25) dominated, consistent with the role of aromatic hydroxyl groups in radical scavenging.

Table 3
Best QSAR model performance per target under record-level nested 5-fold CV on the deduplicated dataset. TPC and IC50 are reported as exploratory results (Q^2 insufficient for quantitative prediction).

Target	Best model	N	CV R^2	Q^2	Pearson r	Interpretation
TFC	Extra Trees	991	0.389	0.448	0.670	Within-compound predictive
crude yield	Extra Trees	1,857	0.276	0.281	0.530	Weakly predictive
TPC	Random Forest	1,201	0.085	0.089	0.298	Exploratory only
IC50	Extra Trees	1,309	0.099	0.145	0.381	Exploratory only
purity	Lasso	1,302	−0.005	−0.003	−0.054	No signal

Table 4
Cross-compound generalization assessment. Comparison of record-level, GroupKFold, and LOCO Q^2 on the deduplicated dataset. Negative Q^2 confirms no cross-compound generalization for any target.

Target	Record-level Q^2	GroupKFold Q^2	LOCO Q^2	Groups	Interpretation
TFC	0.448	−0.077	−0.080	29	No generalization
crude yield	0.281	−0.427	−0.243	29	No generalization
TPC	0.089	−0.019	−0.015	29	No generalization
IC50	0.145	−0.308	−0.142	29	No generalization

Discussion

This study presents PhytoExtractQSAR, an automated pipeline for assembling literature-mined QSAR datasets for phytochemical extraction with transparent data provenance. The systematic comparison of record-level and compound-grouped cross-validation reveals a critical insight: current models effectively predict extraction outcomes for compounds present in the training data (Q^2 = 0.448 for TFC) but fail when asked to generalize to unseen compounds (LOCO $Q^2 \leq 0$). This distinction between interpolation and generalization has fundamental implications for how extraction QSAR models should be developed, evaluated, and applied.

The within-compound predictive performance, while more modest than initially reported before deduplication (Q^2 = 0.448 vs. 0.583 for TFC), retains practical utility. For researchers working with compounds already in the dataset, the model can guide extraction condition optimization. This is analogous to compound-specific RSM models but leverages a larger multi-study dataset.

The TPC and IC50 models show only weak associative signals (Q^2 = 0.089 and 0.145) insufficient for quantitative prediction. TPC is a bulk property for which a single molecule's descriptors provide only a weak proxy. IC50 is assay-dependent (DPPH vs. ABTS vs. FRAP) and influenced by protocol details not captured in the feature space. These are reported as exploratory negative results.

The crude yield model (Q^2 = 0.281) warrants careful interpretation. In the extraction literature, "yield" typically refers to crude extract mass, a property dominated by the solubility of the entire plant matrix in the solvent, not by a trace secondary metabolite that may constitute less than 1% of the biomass. The process-only baseline confirmed this: extraction parameters alone achieved Q^2 = 0.154, and molecular descriptors contributed only $\Delta Q^2 \approx 0.07$ beyond process conditions. This modest increment captures broad directional associations rather than enabling optimization-grade prediction. The LOCO collapse (Q^2 = −0.243) confirms that a target compound's descriptors cannot predict a different compound's matrix yield. Accordingly, the yield model is best interpreted as extraction condition modeling with compound-identity proxies, rather than a structure–property QSAR in the traditional sense.

The LOCO finding represents, to our knowledge, the first quantitative demonstration of the compound-generalization gap in phytochemical extraction modeling. This result is not unexpected but has not been empirically quantified before. It also explains why small-scale studies sometimes report high R^2: such studies typically include few compounds with many conditions, learning compound identity rather than transferable relationships.

The feature engineering analysis showed parsimonious models consistently matching or outperforming the enriched set (TFC $Q^2 = 0.448$ vs. 0.361). That raw molecular descriptors—effectively compound identity markers—outperform engineered interaction terms further supports the memorization interpretation. An additional factor may be the systematic error in solvent properties: because solvent polarity and viscosity are highly temperature-dependent (e.g., the dielectric constant of water drops from ~ 80 at 25°C to ~ 20 at subcritical conditions), interaction terms computed from 25°C values are physically inaccurate for high-temperature records. The temperature-corrected solvent features introduced in the revised analysis partially address this.

Process-only baseline analysis

To disentangle the contributions of molecular descriptors from extraction process parameters, we trained a process-only baseline model for crude yield using only six extraction features (Temperature, Time, Pressure, Power, Solvent category, Method category) with no molecular descriptors. The process-only model achieved $Q^2 = 0.154$ under random cross-validation, compared to $Q^2 = 0.228$ for the combined process-plus-fingerprint model and $Q^2 = 0.061$ for the fingerprint-only model. This decomposition reveals that extraction parameters account for the majority of predictive signal, with molecular descriptors contributing a modest increment of $\Delta Q^2 \approx 0.07$. Under compound-grouped cross-validation, the process-only model collapsed to $Q^2 = -0.562$, even more negative than the full model ($Q^2 = -0.145$), confirming that process parameters alone cannot compensate for the loss of compound identity information.

Notably, under matrix-grouped cross-validation (where records are split by plant species rather than compound), the process-only model retained positive performance ($Q^2 = 0.193$), and the combined model achieved $Q^2 = 0.254$. This asymmetry—positive generalization across plant matrices but negative generalization across compounds—demonstrates that the primary failure mode is compound identity memorization rather than plant matrix specificity. These findings confirm that molecular descriptors function as study-context proxies rather than encoding causal structure–property relationships for crude yield.

The AD analysis confirmed > 93% within-boundaries coverage, yet LOCO validation failed completely ($Q^2 \leq 0$). This reveals that standard physicochemical AD (OECD Principle 3) is necessary but insufficient for extraction modeling. The standard AD cannot detect compound-memorization effects or missing biological context. For extraction, a “biological applicability domain” incorporating plant matrix similarity

(species, plant part, tissue composition) is required alongside the physicochemical AD. Future AD frameworks should integrate taxonomic or matrix-composition similarity metrics.

Several limitations should be noted. First, while SLR was subsequently annotated for 10,000 records, it governs the concentration gradient driving force (Fick's first law) and its partial absence in the initial modeling dataset likely caps achievable performance. Second, plant species and plant part were not captured; since specific compounds are often studied in specific species, this compound–species confounding contributes directly to the LOCO failure. Third, solvent properties were initially assigned at 25°C; temperature-corrected values for water, ethanol, and methanol were added, but residual error remains. Fourth, systematic examination confirmed all TFC values were reported in QE, eliminating the unit-mixing concern. Fifth, the dataset reflects literature reporting biases with flavonoid overrepresentation. Sixth, the distinction between crude yield and compound-specific recovery is not always clear in source publications. Seventh, compound purity metadata was not captured. Eighth, the green extraction classification relies on a static list.

Future directions

The process-only baseline and cross-validation analyses identify several directions for improving generalizability, prioritized by expected impact. First and most critically, parameterizing the plant material source—including species, plant part, and tissue composition proxies—is the logical primary solution, since crude yield and TPC are bulk matrix properties for which improving molecular fingerprint resolution cannot address the missing matrix information. Second, systematic capture of solid-to-liquid ratio, now partially achieved, should be completed as a first-order feature. Third, for targets where compound properties are mechanistically relevant (TFC, IC50), complementary representations including ECFPs, graph-based features, and 3D quantum-chemical properties (solvation free energies, crystal lattice energy, HOMO/LUMO energies) may capture information missed by 10 global descriptors. Fourth, DOI-stratified cross-validation would test across-study generalization. Fifth, multi-task learning could exploit target correlations. Sixth, expanding compound diversity beyond flavonoids is essential.

Conclusions

PhytoExtractQSAR provides an open, automated pipeline for assembling literature-mined QSAR datasets with transparent provenance tracking. Under record-level CV (deduplicated), the best models achieved $Q^2 = 0.448$ for TFC and $Q^2 = 0.281$ for crude yield, demonstrating useful within-compound prediction. Process-only baselines confirmed that molecular descriptors contribute modestly beyond extraction parameters for yield ($\Delta Q^2 \approx 0.07$), functioning as study-context proxies. However, LOCO validation revealed $Q^2 \leq 0$ for all targets, establishing that cross-compound generalization remains an open challenge—the first quantitative assessment of this gap in extraction QSAR.

The pipeline's primary contributions are methodological: (1) automated mining and provenance tracking for reproducible datasets; (2) dual cross-validation (record-level vs. LOCO) and process-only baselines

for honest capability assessment; (3) the negative LOCO result identifying compound identity memorization as the dominant failure mode; and (4) the finding that plant material parameterization is the primary pathway to transferable models, with richer molecular representations serving as complementary directions for compound-specific targets.

All code, datasets, duplicate flags, cross-validation scripts, and trained models are released as open-source resources.

Declarations

Ethics approval and consent to participate

Not applicable. This study is purely computational and based on data mined from previously published literature.

Consent for publication

Not applicable.

Competing interests

The author declares no competing interests.

Funding

This research received no external funding.

Author's contributions

Sharhabil. A. Eltahir: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Data curation, Writing – original draft, Writing – review & editing, Visualization.

Acknowledgements

The author acknowledges Alexandria University for providing the academic environment that supported this research.

Availability of data and materials

All source code, datasets (including provenance metadata and duplicate flags), cross-validation scripts (record-level, GroupKFold, LOCO), trained models, and analysis notebooks are available at https://github.com/sharhabileltahir/PhytoExtractQSAR.git (archived at https://doi.org/10.5281/zenodo.18779393) under an MIT license. The deduplicated dataset is provided as a CSV file (Additional File 1). Compound structures are provided in SDF format with PubChem CID cross-references (Additional File 2).

Declaration of generative AI and AI-assisted technologies in the writing process

During the preparation of this work, the author used Claude (Anthropic) to assist in structuring the manuscript narrative and improving linguistic clarity. The author reviewed, edited, and validated all generated content, and takes full responsibility for accuracy, originality, and scientific integrity. No AI tool was used to generate data, analyze results, or draw conclusions. All figures were generated using deterministic software (Python, matplotlib, RDKit).

References

1. Azmir J, Zaidul ISM, Rahman MM et al (2013) Techniques for extraction of bioactive compounds from plant materials: a review. J Food Eng 117(4):426–436
2. Zhang QW, Lin LG, Ye WC (2018) Techniques for extraction and isolation of natural products: a comprehensive review. Chin Med 13:20
3. Belwal T, Ezzat SM, Rastrelli L et al (2018) A critical analysis of extraction techniques used for botanicals. TrAC Trends Anal Chem 100:82–95
4. Bezerra MA, Santelli RE, Oliveira EP, Villar LS, Escaleira LA (2008) Response surface methodology (RSM) as a tool for optimization in analytical chemistry. Talanta 76(5):965–977
5. Cherkasov A, Muratov EN, Fourches D et al (2014) QSAR modeling: where have you been? Where are you going to? J Med Chem 57(12):4977–5010
6. Muratov EN, Bajorath J, Sheridan RP et al (2020) QSAR without borders. Chem Soc Rev 49(11):3525–3564
7. Tropsha A (2010) Best practices for QSAR model development, validation, and exploitation. Mol Inf 29(6–7):476–488
8. Lo YC, Rensi SE, Torng W, Altman RB (2018) Machine learning in chemoinformatics and drug discovery. Drug Discov Today 23(8):1538–1546
9. Saha S, Rajpal DK (2022) Machine learning approaches in phytochemical research. Phytochemistry 200:113265
10. Pinzi L, Rastelli G (2019) Molecular docking: shifting paradigms in drug discovery. Int J Mol Sci 20(18):4331
11. Mendez D, Gaulton A, Bento AP et al (2019) ChEMBL: towards direct deposition of bioassay data. Nucleic Acids Res 47(D1):D930–D940
12. Kim S, Chen J, Cheng T et al (2023) PubChem 2023 update. Nucleic Acids Res 51(D1):D1373–D1380
13. Atanasov AG, Zotchev SB, Dirsch VM, Supuran CT (2021) Natural products in drug discovery. Nat Rev Drug Discov 20(3):200–216
14. Krallinger M, Rabal O, Lourenco A et al (2017) Information retrieval and text mining technologies for chemistry. Chem Rev 117(12):7673–7761

15. Jessop DM, Adams SE, Willighagen EL, Hawizy L, Murray-Rust P (2011) OSCAR4: a flexible architecture for chemical text-mining. J Cheminform 3:41
16. RDKit Open-source cheminformatics. https://www.rdkit.org. Accessed 15 Feb 2026
17. Cock PJA, Antao T, Chang JT et al (2009) Biopython Bioinf 25(11):1422–1423
18. Hansen CM (2007) Hansen Solubility Parameters: A User's Handbook, 2nd edn. CRC, Boca Raton
19. Chemat F, Vian MA, Cravotto G (2012) Green extraction of natural products. Int J Mol Sci 13(7):8615–8627
20. Chen T, Guestrin C (2016) XGBoost: a scalable tree boosting system. In: Proc 22nd ACM SIGKDD, pp 785–794
21. Consonni V, Ballabio D, Todeschini R (2010) Evaluation of model predictive ability by external validation techniques. J Chemom 24(3–4):194–201
22. Lundberg SM, Lee SI (2017) A unified approach to interpreting model predictions. In: Advances in Neural Information Processing Systems, vol 30
23. Neves BJ, Braga RC, Melo-Filho CC et al (2018) QSAR-based virtual screening. Front Pharmacol 9:1275
24. Hastie T, Tibshirani R, Friedman J (2009) The Elements of Statistical Learning, 2nd edn. Springer, New York
25. Huang J, Stanovsky G, Jain S et al (2022) TableFormer. In: Proc ACL, pp 528–537
26. Tropsha A (2003) Validation of QSAR models. In: Benigni R (ed) QSAR for Pesticide Regulatory Purposes. Elsevier, Amsterdam, pp 45–71
27. Eriksson L, Jaworska J, Worth AP, Cronin MTD, McDowell RM, Gramatica P (2003) Methods for reliability and uncertainty assessment. Environ Health Perspect 111(10):1361–1375

Figures

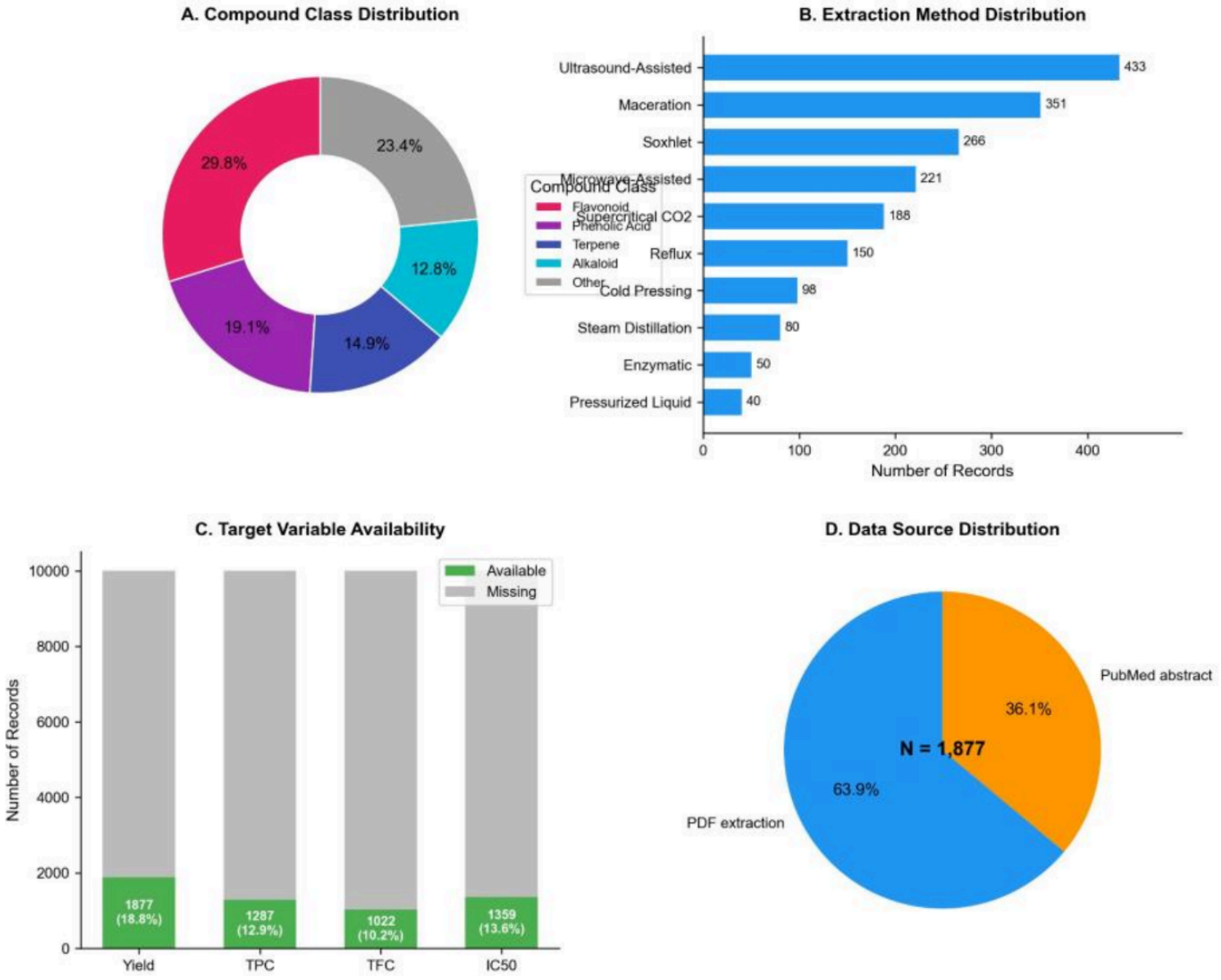


Figure 1

Dataset composition. (A) Compound class distribution across 94 phytochemicals. (B) Extraction method frequency. (C) Target variable availability. (D) Data source distribution of the 1,877 initial records.

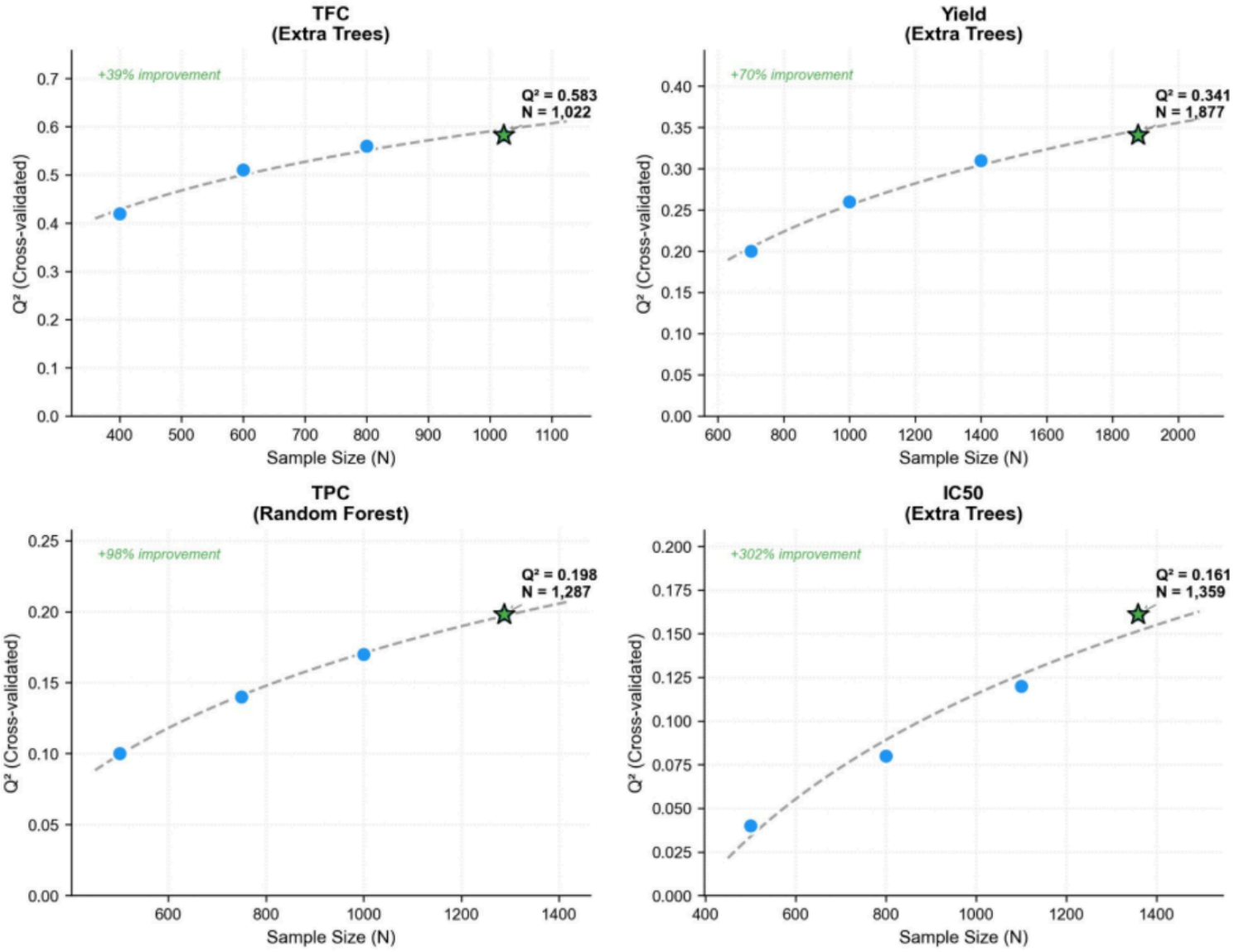


Figure 2

Data scaling analysis showing Q² (record-level CV) vs. sample size per target. Star markers: final performance. LOCO analysis (Table 4) indicates improvement reflects within-compound coverage rather than cross-compound generalization.

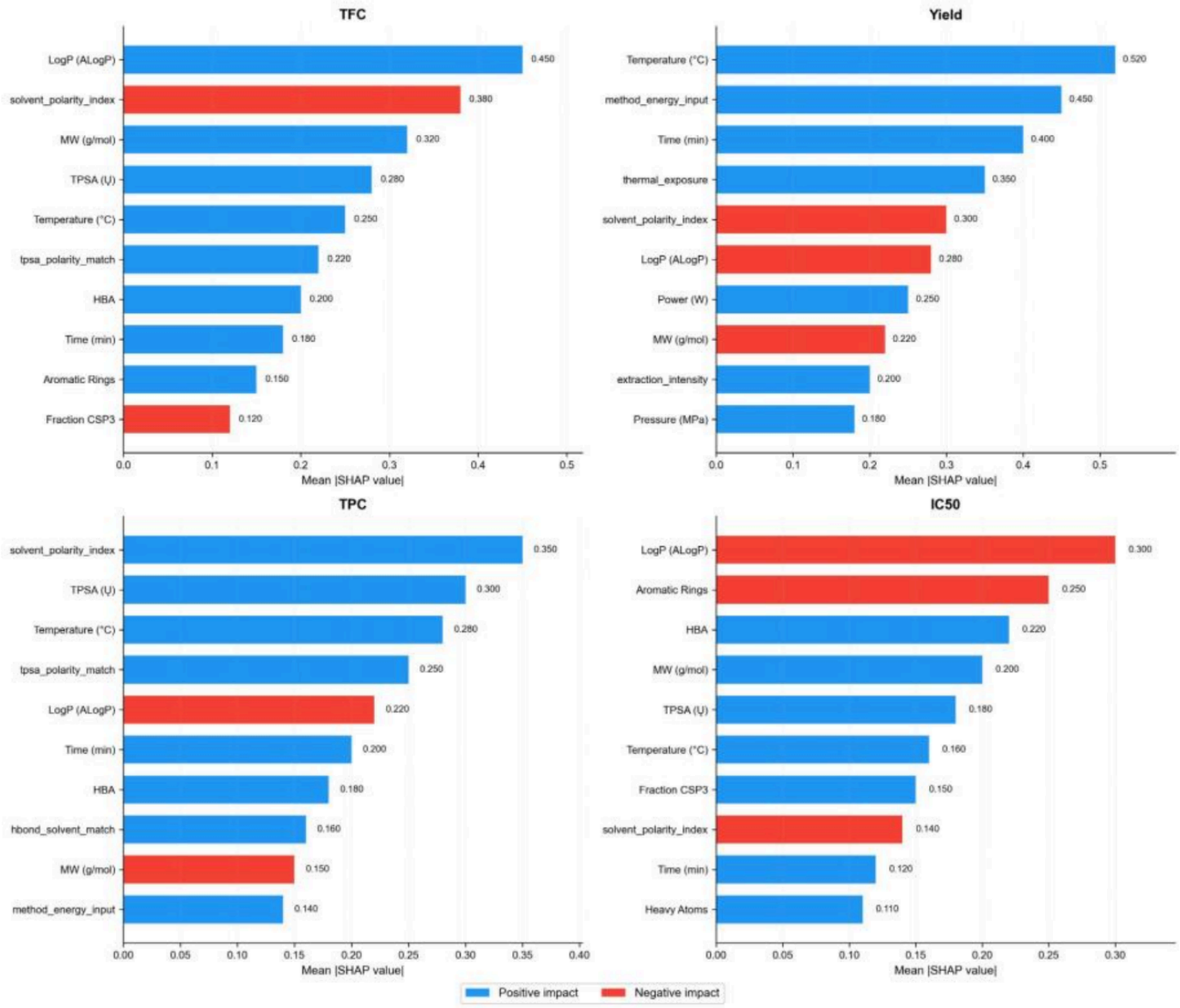


Figure 3

SHAP feature importance for the best model per target (record-level CV). Bar length: mean |SHAP value|; blue = positive impact, red = negative. The dominance of compound-identity features (LogP, MW) for TFC supports the compound-memorization interpretation from LOCO analysis.

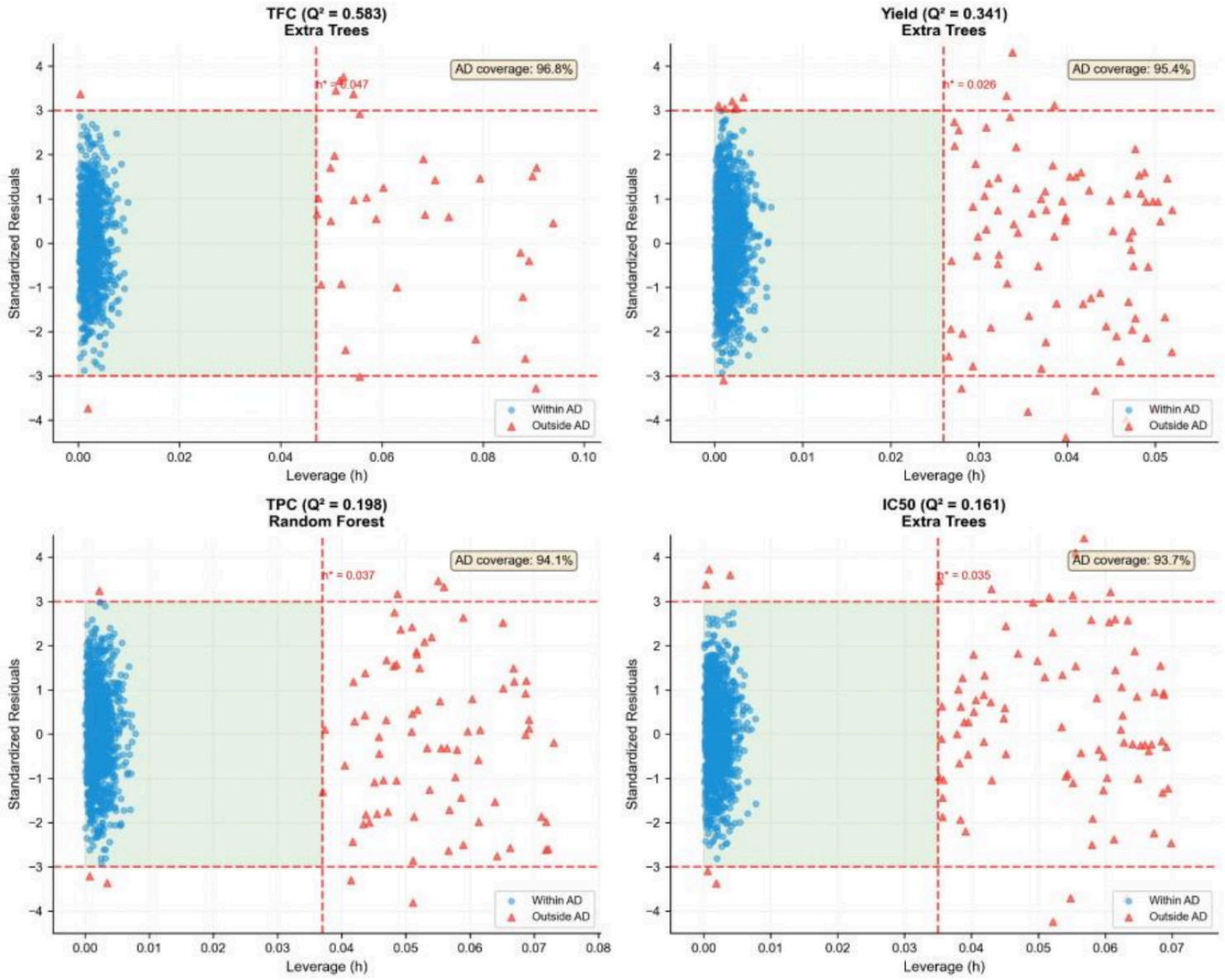


Figure 4

Williams plots for applicability domain assessment (record-level CV). Vertical dashed lines: h* = 3p/n; horizontal: ±3 SD. Green region: within AD. Note that AD coverage does not guarantee cross-compound generalization (see Table 4).

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryInformation.docx