Predicting Reversed-Phase Liquid-Chromatography Retention Time from Molecular Structure with Gradient-Boosted Trees on the METLIN SMRT Dataset

July 12, 2026

Abstract

Retention time (RT) is an orthogonal, information-rich property that can sharpen small-molecule identification in untargeted liquid-chromatography–mass-spectrometry (LC–MS) metabolomics, yet its predictive modelling has historically been limited by the scarcity of large, uniformly acquired training data. The METLIN small-molecule retention-time (SMRT) dataset—80,038 authentic compounds measured under a single reversed-phase (RP) chromatographic method—removed that barrier and has become the reference benchmark for quantitative structure–retention relationship (QSRR) modelling. Here we develop a structure-based RT predictor on the original 2019 SMRT release. Beginning from the raw 80,038 entries, we additionally applied the 2024 erroneous-entry filter of Khrisanfov et al., removing the 1,299 flagged structures, and combined this with RDKit-based standardisation and deduplication to yield 78,629 curated compounds (1,409 entries removed in total). Each molecule was represented by a 2048-bit radius-2 Morgan fingerprint concatenated with 211 physicochemical descriptors. Compounds were split 80/20 by structure (random seed 42), and a gradient-boosted decision-tree regressor (LightGBM, L_1 objective) was selected over ridge regression and single-representation baselines by three-fold cross-validation. On the 15,726-compound hold-out set the model achieved a mean absolute error (MAE) of 45.99 s, a median absolute error of 25.09 s, a coefficient of determination $R^2 = 0.842$, and 79.2% of compounds predicted within 60 s of the measured RT. Predictions on the poorly-retained subpopulation eluting within the first 300 s were markedly worse (MAE 342.0 s; 5.3% within 60 s), and the model systematically over-predicted their retention—an applicability-domain effect consistent with the near-void behaviour of non-retained analytes. Feature-attribution analysis showed that physicochemical descriptors carried 74.6% of the total split gain, with the Wildman–Crippen octanol–water partition estimate (`MolLogP`) alone accounting for 6.25%, followed by partial-charge- and surface-area-weighted polarity descriptors (`PEOE_VSA`, `SlogP_VSA`, `VSA_EState`). These drivers recapitulate the classical hydrophobic-partitioning mechanism of RP chromatography and demonstrate that a compact, interpretable, tabular model reproduces the retention behaviour of a large, structurally diverse chemical space while transparently exposing where such models should—and should not—be trusted.

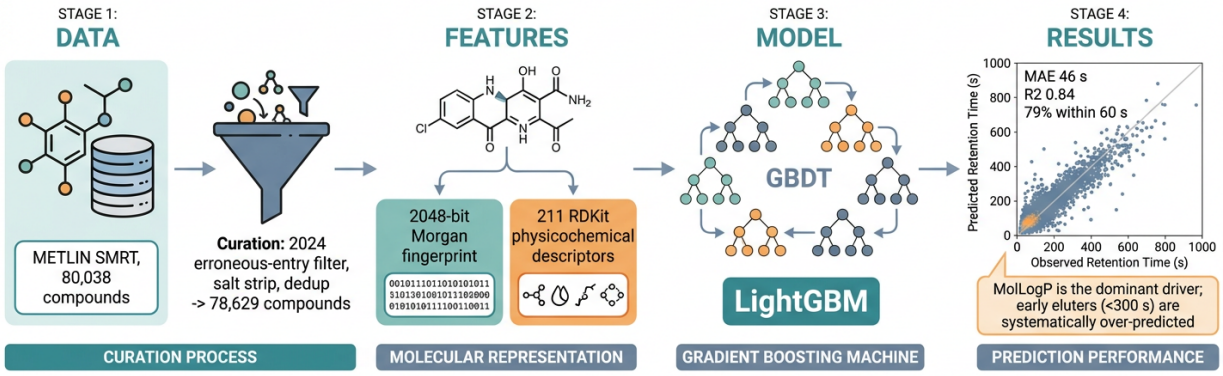

Figure 1: Graphical abstract. From the 80,038-compound METLIN SMRT release, curation (2024 erroneous-entry filter, salt stripping, deduplication) yields 78,629 compounds; each is encoded as a Morgan fingerprint plus RDKit descriptors and modelled with a gradient-boosted tree ensemble (LightGBM). The hold-out predictor attains MAE 46 s, $R^2 = 0.84$ and 79% of compounds within 60 s, with `MolLogP` the dominant driver and early eluters systematically over-predicted.

Keywords: retention time prediction; QSRR; reversed-phase liquid chromatography; METLIN SMRT; gradient boosting; LightGBM; molecular descriptors; metabolite annotation.

1 Introduction

Untargeted metabolomics and small-molecule screening by liquid-chromatography–mass-spectrometry (LC–MS) generate thousands of chromatographic features per sample through automated peak-detection and alignment pipelines [1, 2], but converting those features into confident chemical identities remains the central bottleneck of the field [3, 4]. Accurate mass and tandem-mass spectra constrain the space of plausible structures, yet isobaric and isomeric candidates routinely survive these filters, and reference spectral libraries cover only a small fraction of known metabolomes [5, 6]. Chromatographic retention time (RT) is an orthogonal, physically grounded property that is measured essentially for free in every LC–MS experiment, and it encodes information—overall hydrophobicity, polar surface area, hydrogen-bonding capacity, ionisation state—that mass spectrometry alone does not resolve. A reliable *in silico* predictor that maps molecular structure to RT can therefore act as a powerful additional filter, demoting candidates whose predicted elution is incompatible with the observed peak and thereby reducing false annotations [7–9].

The relationship between molecular structure and chromatographic retention has been studied for decades under the banner of quantitative structure–retention relationships (QSRR), the chromatographic analogue of quantitative structure–activity relationships [10]. Classical QSRR models regress retention factors or retention indices onto a small number of physically interpretable descriptors—logarithmic octanol–water partition coefficients, polar surface area, hydrogen-bond

donor and acceptor counts, and solvatochromic (linear solvation energy) parameters—and have proven invaluable both for rationalising separation mechanisms and for guiding method development [10–12]. In reversed-phase (RP) chromatography, where a non-polar stationary phase is eluted with an increasingly organic mobile phase, retention is governed predominantly by hydrophobic partitioning: more lipophilic analytes are retained longer, so descriptors of hydrophobicity dominate most RP-QSRR models [10–12]. The principal historical limitation of QSRR has not been the modelling formalism but the data: retention is exquisitely sensitive to the column chemistry, mobile-phase composition, gradient programme, temperature and instrument, so models trained on one laboratory's small, idiosyncratic dataset rarely transfer to another [13–15].

The METLIN small-molecule retention-time (SMRT) dataset transformed this landscape by providing experimentally measured RTs for 80,038 authentic molecular standards acquired under a *single*, uniform RP method [16]. By decoupling the structure–retention modelling problem from the confounding variability of heterogeneous instrumentation, SMRT created, for the first time, a benchmark large enough to train and fairly evaluate expressive machine-learning models. The original release trained a deep neural network on molecular descriptors and fingerprints and demonstrated that predicted RT could rank the correct identity among the top three candidates in roughly 70% of test cases, establishing RT prediction as a practical annotation aid [16]. In the five years since, SMRT has underpinned a rapidly growing body of work spanning gradient-boosted trees, graph neural networks and transformer architectures, as well as projection and calibration schemes that transfer SMRT-trained predictions onto other chromatographic systems [8, 9, 17–19].

Two features of large, community-curated benchmarks demand explicit attention. First, datasets of this scale inevitably contain errors—mis-assigned structures, transcription mistakes and mislabelled peaks—that can silently degrade or bias a model. Khrisanfov et al. recently interrogated SMRT with an ensemble of five predictive models and flagged approximately 1,299 entries as potentially erroneous, offering an explicit exclusion list for downstream users [20]. Whether and how such a filter is applied is a material methodological choice that should be reported. Second, a substantial fraction of any RP dataset consists of poorly retained, near-void-volume species whose elution is dominated by column dead time rather than by analyte–stationary-phase interaction; these compounds lie at the edge of, or outside, the applicability domain of a hydrophobicity-driven QSRR model [11, 14]. Characterising model behaviour on this subpopulation is essential for honest deployment.

Among the modelling options, gradient-boosted decision-tree ensembles occupy a pragmatic sweet spot for structure–property prediction on tabular fingerprint-and-descriptor inputs [21–23]. They natively handle the mixed binary/continuous, high-dimensional, sparse feature matrices produced by cheminformatics pipelines; they are robust to uninformative features and monotone transformations; they train in minutes on tens of thousands of molecules; and, crucially for scientific interpretation, they expose transparent feature-attribution statistics that connect predictions back to physicochemical descriptors [24]. These properties make them a strong, interpretable baseline against which more complex deep-learning approaches must justify their added cost.

In this work we build a structure-based RT predictor on the original 2019 SMRT release and report a complete, reproducible account of every methodological decision. We (i) curate the raw 80,038 entries, explicitly applying the 2024 erroneous-entry filter and reporting the exact number of removed structures; (ii) represent each molecule with a Morgan fingerprint concatenated with RDKit physicochemical descriptors; (iii) hold out 20% of compounds by structure using a stated random seed and select a LightGBM regressor by cross-validation; and (iv) report on the hold-out set the mean and median absolute error, the coefficient of determination, and the percentage of

compounds predicted within 60 s of the measured RT, together with a dedicated analysis of the poorly-retained (≤ 300 s) subpopulation and complete per-compound predictions. Finally, through feature-attribution and residual analyses we connect the model's numerical performance to the underlying chromatographic chemistry and delineate precisely where the predictor can and cannot be trusted.

2 Materials and Methods

The entire workflow is deterministic and reproducible; its four stages—curation, feature engineering, model selection, and evaluation—are summarised in Figure 2.

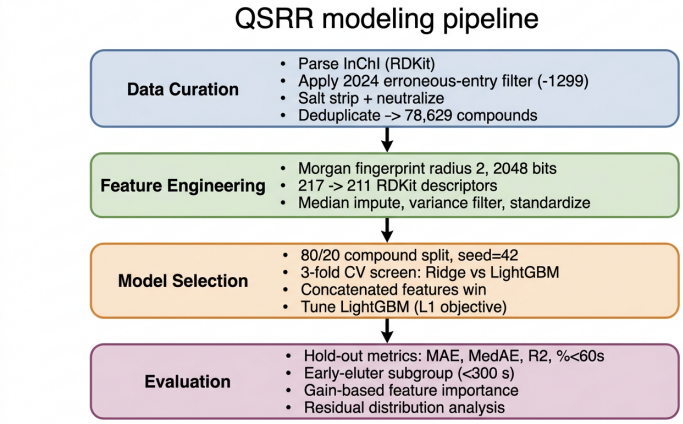

Figure 2: Computational QSRR pipeline. Raw SMRT entries are curated (structure parsing, 2024 erroneous-entry filter, salt stripping/neutralisation and deduplication), encoded as Morgan fingerprints plus RDKit descriptors, split 80/20 by compound, and modelled by a cross-validated, hyper-parameter-tuned LightGBM regressor. Evaluation reports hold-out error metrics, an early-eluter subgroup analysis, gain-based feature importance and the residual distribution.

2.1 Dataset and provenance

The METLIN SMRT dataset was obtained from its public Figshare deposition accompanying the original 2019 release [16]. The distributed file (`SMRT_dataset.csv`, semicolon-delimited) contains three fields per record—the PubChem compound identifier, the experimental retention time in seconds, and the molecular structure as an InChI string—for 80,038 unique compounds. Retention times were acquired by the originating laboratory under a single reversed-phase gradient method, so no chromatographic method descriptors are required or available; the modelling task is thus a pure structure→RT mapping conditioned on one fixed separation. The raw file contained 80,038 records with unique PubChem identifiers and retention times spanning 0.3 s to 1471.7 s.

2.2 Chemical curation

All structure handling used RDKit (version 2026.03.3) [25], and the curation pipeline was designed to reconcile exactly against the raw entry count. Four sequential operations were applied.

(A) Structure parsing. Each InChI string was parsed to an RDKit molecule and converted to a canonical SMILES; 81 entries whose structures could not be parsed were removed, leaving 79,957

records.

(B) 2024 erroneous-entry filter. We additionally applied the published exclusion list of Khrisanfov, Matyushin and Samokhin, who used an ensemble of five retention-time predictors and five-fold cross-validation to flag the ≈2% of SMRT entries whose measured RT deviated most strongly from consensus predictions across all five models [20]. The list of 1,299 canonical structures was retrieved from the authors' public repository. Independent canonicalisation reproduced the flagged set exactly: all 1,299 listed structures matched dataset entries (1,299/1,299), and **1,299 entries were removed** by this filter, leaving 78,658 records. We report this step explicitly because whether the 2024 filter is applied is a material choice affecting the training distribution.

(C) Standardisation. Surviving structures were standardised by largest-fragment selection (stripping salts and solvents), charge neutralisation (RDKit `Uncharger`) and normalisation/cleanup. Consistent with METLIN providing desalted parent structures, no multi-fragment records were encountered; six records had formal charges neutralised, and no standardisation failures occurred.

(D) Deduplication. Duplicate standardised canonical structures were resolved to guarantee a strict compound-level train/test separation downstream. Of 18 duplicated structures (38 entries), the 9 structures whose replicate RTs agreed to within 30 s were collapsed to a single median-RT representative (removing 9 redundant rows), while the 9 structures with discordant replicate RTs were removed entirely (removing 20 rows) to avoid ambiguous ground truth. The accounting reconciles exactly ($80{,}038 - 81 - 1{,}299 - 9 - 20 = 78{,}629$), yielding **78,629 curated compounds**. We deliberately retained low-RT, near-void species at this stage; the more aggressive RT > 400 s pre-filter used by some authors to discard non-retained analytes is a modelling-specific choice and was not imposed on the curated set, so that its impact could be examined directly (Section 3.4).

2.3 Molecular feature engineering

Features were computed from the standardised SMILES for all 78,629 compounds. Two complementary representations were generated. First, a 2048-bit circular Morgan fingerprint of radius 2 (the ECFP4 analogue) was computed with RDKit's current `MorganGenerator` API and stored as a dense binary vector; circular fingerprints encode the local atomic environments that determine substructural identity and are the community-standard structural representation for ligand-based machine learning [26, 27]. Second, the full RDKit physicochemical descriptor set (217 descriptors) was computed, capturing global molecular properties including molecular weight, the Wildman–Crippen octanol–water partition estimate (`MolLogP`) and molar refractivity [28], topological polar surface area [29], Gasteiger partial-charge-weighted surface areas (the `PEOE_VSA` family) [30], the van der Waals surface-area descriptors of Labute (the `SlogP_VSA`, `SMR_VSA` and `VSA_EState` families) [31], hydrogen-bond donor/acceptor counts, rotatable-bond counts, aromatic-ring counts and fragment indicators. Non-finite values were unified to `NaN` and imputed column-wise with the median; six zero-variance descriptors were removed by a variance threshold, and the remaining 211 descriptors were standardised to zero mean and unit variance. Fingerprint bits were retained as raw binary features. The final per-molecule representation was the concatenation of the 2048 fingerprint bits and the 211 descriptors (2,259 features).

2.4 Train/test split

The 78,629 compounds were partitioned into an 80% training set (62,903 compounds) and a 20% hold-out test set (15,726 compounds) with a fixed random seed of **42**. Because duplicate structures

had already been resolved during curation, a row-wise random split is a strict compound-level split; this was verified programmatically by confirming that the training and test index sets, and the corresponding PubChem identifier sets, were disjoint. All model selection and hyper-parameter tuning used the training set only; the hold-out set was touched exactly once, for final evaluation.

2.5 Model development and selection

We first screened two learning algorithms against three feature representations by three-fold cross-validation on the training set, scoring each configuration by mean absolute error (MAE). The baseline was ridge regression—a regularised linear model [32]—and the candidate non-linear model was LightGBM, a histogram-based gradient-boosted decision-tree implementation optimised for large, high-dimensional data [21, 23]. The three representations were the Morgan fingerprint alone, the descriptors alone, and their concatenation. LightGBM on the concatenated features gave the lowest cross-validated error and was carried forward. We then tuned LightGBM over a grid of learning rate $\in \{0.03, 0.05, 0.1\}$ and maximum leaves $\in \{127, 255\}$ at 800 boosting iterations, again by three-fold cross-validation. Throughout, LightGBM was trained with the L_1 (least-absolute-deviation) objective so that optimisation targeted the reported MAE directly, with subsampling (0.8), column subsampling (0.8) and L_2 leaf regularisation to control over-fitting. The winning configuration (learning rate 0.1, 127 leaves, 800 trees) was refit on the entire training set and evaluated once on the hold-out set. For comparison against related work we also relate our tabular baseline to published deep-learning and boosting predictors trained on SMRT [8, 9, 16, 17].

2.6 Evaluation metrics

On the hold-out set we report the mean absolute error (MAE), the median absolute error (MedAE), the coefficient of determination R^2, and the percentage of compounds whose predicted RT falls within a tolerance window of ± 60 s of the measured value—an annotation-relevant criterion reflecting the RT precision needed to discriminate candidates. Absolute error for compound i is $|y_i - \hat{y}_i|$ and the signed residual is defined throughout as $r_i = y_i - \hat{y}_i$ (observed minus predicted), so that a negative residual denotes over-prediction. As a targeted stress test, the same metrics were recomputed on the subpopulation of poorly retained analytes with measured RT ≤ 300 s.

2.7 Feature-importance and residual analyses

To interpret the fitted ensemble we extracted, for every one of the 2,259 features, both the total split *gain* (the summed reduction in the loss function attributable to splits on that feature) and the split *count*. Gain was used as the primary ranking statistic because it reflects the actual contribution of a feature to loss reduction. Importances were aggregated by feature type (fingerprint bit versus descriptor) to quantify the relative contribution of substructural versus physicochemical information. Separately, the hold-out residuals were summarised by their mean, standard deviation, median, skewness and excess kurtosis using SciPy, both for the full test set and for the early-eluting subpopulation.

2.8 Software and reproducibility

The pipeline was implemented in Python 3.12 with RDKit 2026.03.3 [25], LightGBM ($\geq$4.6) [23], scikit-learn ($\geq$1.5) [32], NumPy, pandas, SciPy and Matplotlib. All random operations were seeded

(seed 42). Curated data, per-compound predictions, the full feature-importance table and the analysis scripts are provided (Data and Code Availability).

3 Results

3.1 Curation outcome

Curation removed 1,409 of the 80,038 raw entries (1.76%): 81 unparsable structures, 1,299 structures flagged by the 2024 erroneous-entry filter, and 29 duplicate-resolution removals (9 collapsed consistent duplicates and 20 discordant duplicates). The resulting 78,629-compound set spans retention times from 0.3 s to 1471.7 s with a mean of 789 s and a median of 773 s. The distribution is strongly right-shifted toward well-retained species: only 2.6% of curated compounds (2,054 molecules) elute within the first 300 s, so poorly retained analytes are a small but non-negligible minority of the chemical space.

3.2 Model and feature-representation selection

The cross-validation screen (Table 1) cleanly separated the candidates. LightGBM outperformed ridge regression for every feature representation, and for both learners the concatenation of fingerprints and descriptors outperformed either representation alone—evidence that substructural and physicochemical features carry complementary retention information. The best screened configuration, LightGBM on concatenated features, reached a cross-validated MAE of 50.4 s, comfortably ahead of the best ridge configuration (63.7 s). Hyper-parameter tuning (Table 2) produced only a modest further improvement, with the optimum at learning rate 0.1, 127 leaves and 800 trees (cross-validated MAE 49.5 s); the flatness of the tuning surface indicates a robust, well-regularised model rather than one balanced on a knife-edge of hyper-parameters.

Table 1: Three-fold cross-validated mean absolute error (MAE, seconds) on the training set for each learner × feature representation. Lower is better; the selected configuration is in **bold**.

Model	Feature representation	CV MAE (s)	SD (s)
LightGBM	**concatenated**	**50.40**	**0.66**
LightGBM	descriptors only	56.66	0.50
LightGBM	Morgan fingerprint	58.52	0.89
Ridge	concatenated	63.65	0.57
Ridge	Morgan fingerprint	74.05	0.31
Ridge	descriptors only	85.70	0.64

Table 2: LightGBM hyper-parameter tuning (three-fold cross-validated MAE, seconds) on the concatenated features at 800 boosting iterations. The selected setting is in **bold**.

Learning rate	Num. leaves	Trees	CV MAE (s)	SD (s)
0.10	**127**	**800**	**49.54**	**0.60**
0.05	255	800	49.61	0.65
0.05	127	800	49.62	0.62
0.10	255	800	50.01	0.70
0.03	255	800	50.22	0.72
0.03	127	800	50.76	0.71

3.3 Hold-out performance

Refit on all 62,903 training compounds and evaluated once on the 15,726 hold-out compounds, the model achieved a mean absolute error of 45.99 s, a median absolute error of 25.09 s, and a coefficient of determination of $R^2 = 0.842$ (Table 3). The substantial gap between the mean and median absolute errors—the median is little more than half the mean—signals a heavy-tailed error distribution in which most predictions are excellent and a minority are badly wrong. Consistent with this, 79.2% of compounds were predicted within ±60 s of the measured RT; tightening or relaxing the tolerance gives 56.6% within 30 s, 88.1% within 90 s and 92.4% within 120 s. The predicted-versus-observed density plot (Figure 3) shows the great majority of compounds packed tightly along the 1:1 identity line across the entire well-retained range from roughly 500 to 1450 s. The single conspicuous departure is a vertical streak of points near an observed RT of ≈80–100 s that the model pushes up toward 400–700 s: the poorly retained analytes, examined next.

Table 3: Hold-out performance on the full test set and on the poorly-retained early-eluter subpopulation (measured RT ≤ 300 s). Residual = observed − predicted.

Subset	N	MAE (s)	MedAE (s)	R^2	Within 60 s (%)
Full test set	15 726	45.99	25.09	0.842	79.2
Early eluters (≤ 300 s)	395	342.02	381.49	−1972.180	5.3
Well retained (> 400 s)	15 331	38.42	–	–	81.1

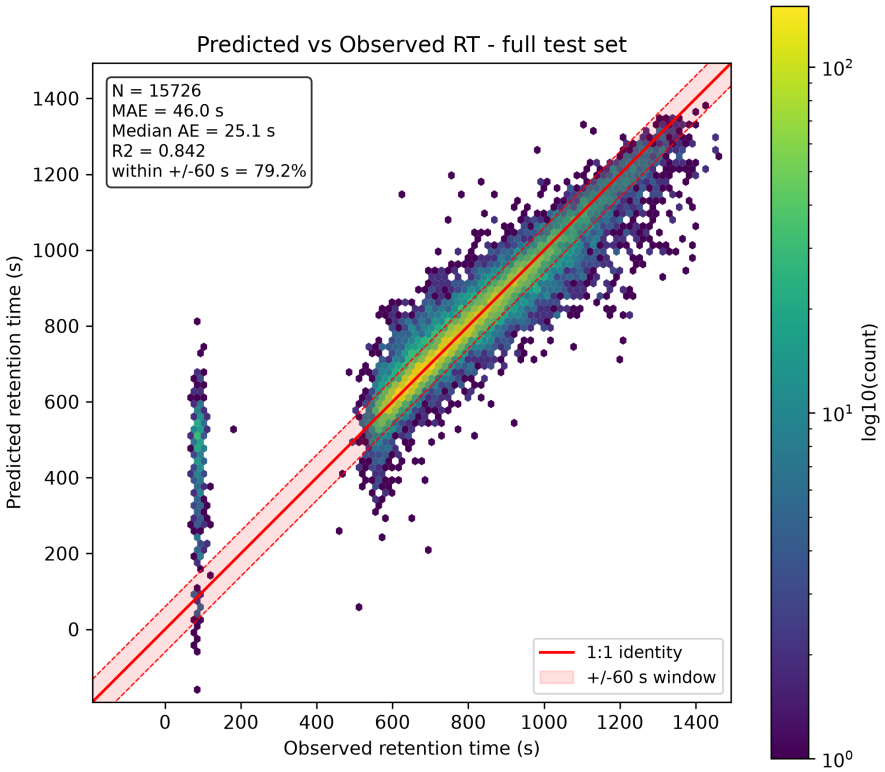

Figure 3: Predicted versus observed retention time on the full hold-out set ($N = 15{,}726$). Colour encodes point density (log scale); the solid line is the 1:1 identity and the shaded band the ± 60 s tolerance. The bulk of compounds tracks the identity line closely; the vertical cluster near an observed RT of $\approx$80–100 s comprises poorly retained analytes that the model over-predicts.

3.4 Poorly-retained (early-eluter) subpopulation

The 395 hold-out compounds with measured RT ≤ 300 s (2.5% of the test set) were predicted far less accurately than the set as a whole, with a subgroup MAE of 342.0 s, a median absolute error of 381.5 s, and only 5.3% falling within the 60 s window (Table 3). The coefficient of determination for this subgroup is strongly negative ($R^2 = -1972$), meaning the model is far worse than simply predicting the subgroup mean; this is expected because the subgroup spans a very narrow observed-RT range while the predictions scatter widely, so any R^2 computed within so compressed a target window is uninformative and should not be over-interpreted. The direction of the error is systematic and one-sided: the mean residual for this subgroup is -335.7 s, i.e. these compounds are almost uniformly *over*-predicted, as the dedicated scatter plot makes plain (Figure 4), where essentially all points sit far above the identity line. The three worst-predicted compounds in the entire test set are all early eluters (observed RTs of 88.5, 96.0 and 88.8 s predicted at 804, 752 and 733 s, respectively). Crucially, this failure is confined to the poorly retained edge of the domain: on the complementary well-retained subset (RT > 400 s, $N = 15{,}331$) the MAE falls to 38.4 s and 81.1% of compounds land within 60 s, appreciably better than the full-set figures that the early-eluter tail inflates.

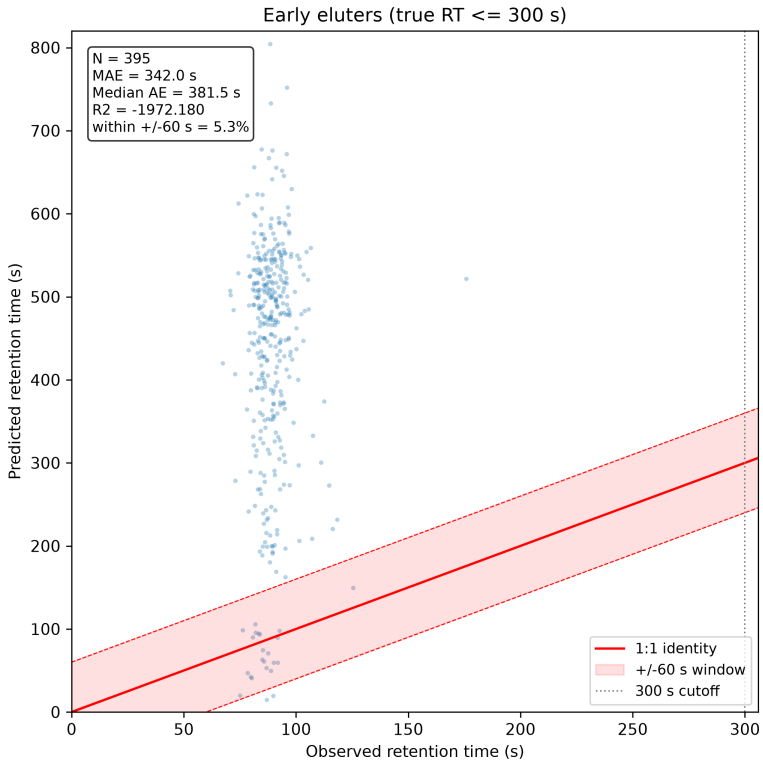

Figure 4: Predicted versus observed retention time for early eluters (measured RT $\leq$ 300 s, $N = 395$). The near-vertical cloud sitting well above the 1:1 line demonstrates a systematic, one-sided over-prediction of poorly retained analytes; only 5.3% fall within the $\pm$60 s band.

3.5 Chemical drivers of retention

Feature attribution over the fitted ensemble revealed a clear hierarchy (Figure 5). Physicochemical descriptors, though only 211 of the 2,259 features, accounted for 74.6% of the total split gain, versus 25.4% shared across all 2,048 fingerprint bits combined; nineteen of the twenty most important features are descriptors, with a single fingerprint bit (`FP_1791`) the lone exception at rank two. The dominant single feature by a wide margin is `MolLogP`, the Wildman–Crippen atom-contribution estimate of the octanol–water partition coefficient [28], which alone carries 6.25% of the total gain—more than three times the next feature. The remaining leaders are overwhelmingly measures of polarity, partial charge and surface-area-weighted hydrophobicity: several members of the Gasteiger partial-charge surface-area family (`PEOE_VSA6`, `PEOE_VSA8`, `PEOE_VSA1`, `PEOE_VSA2`, `PEOE_VSA7`) [30], the LogP-weighted van der Waals surface-area descriptor `SlogP_VSA2` and the electrotopological-state surface-area descriptors (`VSA_EState3`, `VSA_EState8`, `VSA_EState7`, `VSA_EState4`) [31], alongside extreme-partial-charge terms (`MaxPartialCharge`, `MinPartialCharge`), a molar-refractivity surface term (`SMR_VSA3`) and BCUT eigenvalue descriptors. The composition of this list is chemically coherent: it is essentially a ranked inventory of the molecular properties that classical theory predicts should govern reversed-phase retention.

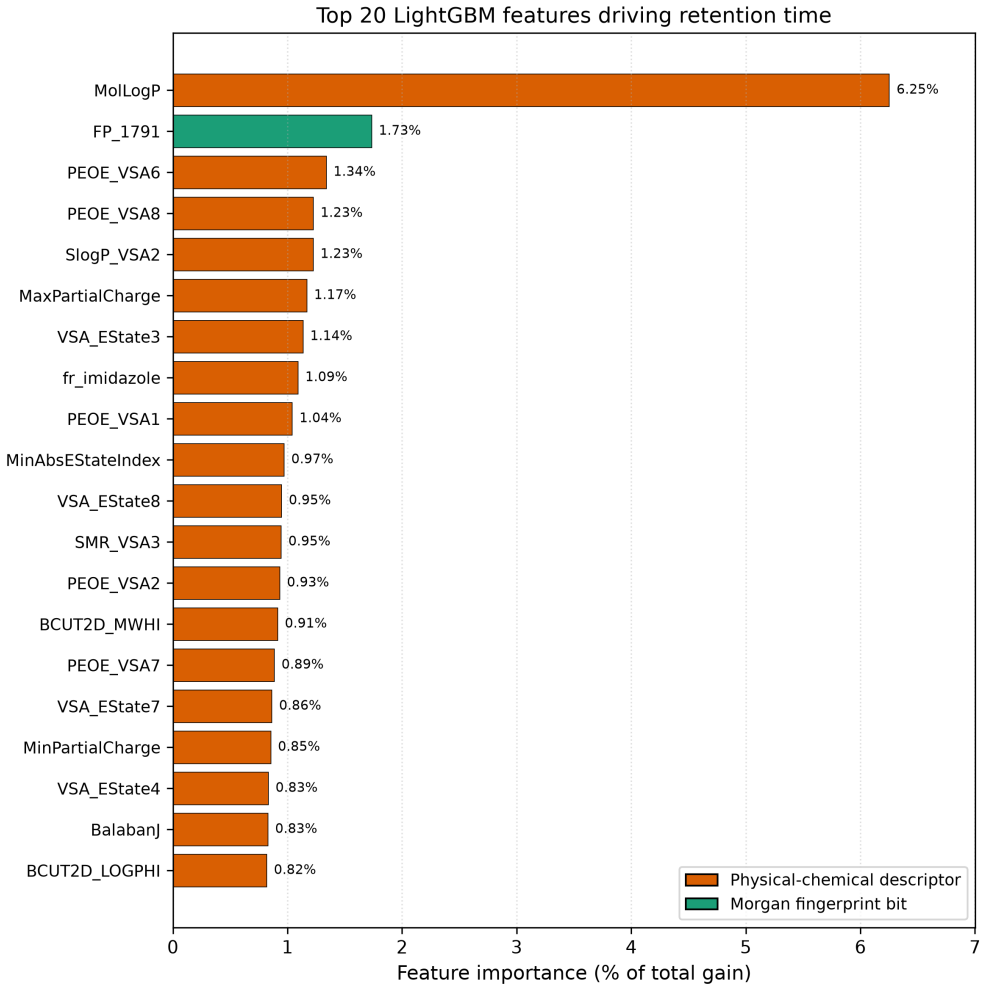

Figure 5: Top-20 features by split gain in the LightGBM model. Bars are coloured by feature type (physicochemical descriptor vs. Morgan fingerprint bit) and annotated with each feature's share of total gain. `MolLogP` dominates (6.25%); the remaining leaders are partial-charge-, surface-area- and electrotopological-state descriptors that encode hydrophobicity and polarity.

3.6 Residual distribution

The hold-out residual distribution (Figure 6) is centred essentially at zero (mean -5.1 s, median 0.07 s), confirming that the model is close to unbiased over the dataset as a whole, but it is sharply peaked and heavy-tailed, with a standard deviation of 82.0 s, a negative skewness of -2.14 and a large excess kurtosis of 13.3. The leptokurtic shape—a tall central spike flanked by long tails—explains the mean/median gap: a dense core of near-perfect predictions coexists with a thin population of large errors. The negative skew reflects the early-eluter over-predictions, which contribute a tail of large negative residuals; within that subgroup alone the mean residual is -335.7 s. In practical terms, for the overwhelming majority of well-retained compounds the residual is small and symmetric, while the distribution's asymmetry is almost entirely a signature of the poorly-retained edge of the applicability domain.

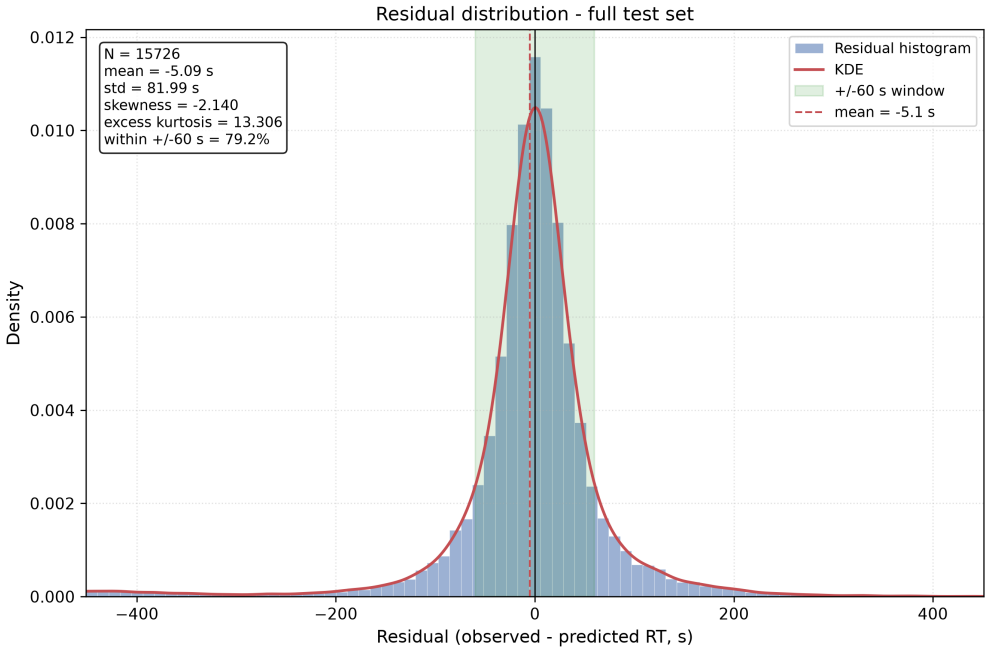

Figure 6: Residual distribution on the full hold-out set (residual = observed − predicted). The histogram and kernel-density estimate are tall and narrow (median ≈ 0 s) but heavy-tailed (excess kurtosis 13.3) and negatively skewed (−2.14); the shaded band marks the ±60 s tolerance containing 79.2% of predictions.

3.7 Per-compound predictions

Predicted, observed and residual retention times for all 15,726 hold-out compounds, keyed by PubChem identifier, are provided in full as `test_predictions.csv` (Data and Code Availability). Table 4 illustrates the range of behaviour with representative best-, median- and worst-predicted compounds. The best predictions are essentially exact; the median-error cases are off by only ≈ 25 s (about 3% of a typical RT); and the worst cases are, without exception, early-eluting analytes over-predicted by more than 600 s.

Table 4: Representative per-compound hold-out predictions (all times in seconds). "Median" rows are drawn from compounds whose absolute error equals the test-set median. All three worst-case compounds are poorly-retained early eluters.

Category	PubChem CID	Observed (s)	Predicted (s)	Residual (s)
Best	83 284 900	771.3	771.3	0.0
Best	53 131 199	1106.2	1106.2	0.0
Best	53 173 082	1038.8	1038.8	0.0
Median	53 164 188	828.9	803.8	25.1
Median	50 831 184	721.1	696.0	25.1
Median	53 170 049	1036.0	1010.9	25.1
Worst	35 397 625	88.5	804.2	−715.7
Worst	50 743 347	96.0	751.7	−655.7
Worst	2 912 138	88.8	732.8	−644.0

4 Discussion

4.1 The model recapitulates the physics of reversed-phase retention

The most reassuring result of this study is not a headline error metric but the composition of the feature-importance ranking, which reads like a textbook account of reversed-phase separation. That a purely data-driven ensemble, given free choice among 2,259 features, elevates `MolLogP` to first place by a factor of three over any competitor is a direct, model-agnostic confirmation that hydrophobic partitioning is the master variable of RP retention [10, 11]. In RP chromatography a non-polar C18 stationary phase is eluted with an increasingly organic mobile phase, and an analyte's elution is governed by the free energy of transfer between the mobile and stationary environments—a quantity that octanol–water partitioning approximates well. The prominence of the Gasteiger partial-charge surface-area (`PEOE_VSA`) and LogP-weighted surface-area (`SlogP_VSA`) descriptors refines this picture: retention responds not only to bulk lipophilicity but to how hydrophobic and polar character is *distributed* over the molecular surface, since it is the solvent-accessible surface that contacts the stationary phase [30, 31]. The electrotopological-state surface descriptors (`VSA_EState`) and extreme partial-charge terms capture the polar and hydrogen-bonding groups that resist partitioning into the non-polar phase and thereby shorten retention. The near-total dominance of interpretable descriptors over fingerprint bits (74.6% versus 25.4% of gain) further indicates that, for this endpoint, global physicochemical character—rather than the presence or absence of specific substructural fragments—carries most of the predictive signal, a finding consistent with the descriptor-only model already outperforming the fingerprint-only model in cross-validation (Table 1). This mechanistic transparency is a concrete advantage of gradient-boosted tabular models over end-to-end deep networks, whose learned representations are far harder to map back onto separation chemistry [9, 24].

4.2 Early eluters and the applicability domain

The systematic over-prediction of poorly retained analytes is the model's most important limitation, and it admits a clean chemical explanation. Compounds eluting within the first few hundred seconds of an RP gradient are, to a large extent, *non-retained*: their elution is dominated by the column void time rather than by analyte-specific partitioning, so their observed RTs are compressed into a narrow band near the dead volume and are only weakly related to structure. A hydrophobicity-driven model trained overwhelmingly on well-retained compounds (97.4% of the curated set elutes after 300 s) has neither the training density nor the mechanistic leverage to place these species correctly; lacking evidence that a compound is non-retained, it defaults toward the bulk of the distribution and over-predicts [11, 14]. The one-sided, large-magnitude residuals in this subgroup (mean -336 s; Figure 4) and their outsized contribution to the negative skew and heavy tails of the overall residual distribution (Figure 6) are two views of the same phenomenon. The strongly negative subgroup R^2 is a statistical artefact of computing a variance-explained metric over a target range narrower than the error scale and should not be read as a coherent goodness-of-fit; the MAE and the direction of the residuals are the interpretable quantities. The practical remedy is to treat non-retention as a distinct regime: many workflows, including the pipeline of the authors who produced the 2024 error filter, impose an RT > 400 s applicability-domain cut-off before modelling [20]. Our results quantify exactly what such a cut-off buys—restricting to RT > 400 s lowers the MAE from 46.0 to 38.4 s and lifts the within-60 s fraction to 81.1%—while also cautioning that any downstream annotation use of predicted RT for early-eluting features should be down-weighted or

accompanied by an explicit out-of-domain flag.

4.3 Relation to prior work

Our hold-out MAE of 46 s sits within the range reported for descriptor- and boosting-based predictors on SMRT and its relatives, and is achieved with a deliberately simple, fast, interpretable tabular model rather than a bespoke deep architecture [8, 14, 16]. The original SMRT study reported that deep-neural-network predictions ranked the correct candidate in the top three in about 70% of cases [16]; gradient boosting has since been shown to be a strong competitor for RT and other molecular-property endpoints [21–24], and dedicated tools such as Retip demonstrated the practical value of tree-based RT predictors for compound annotation [8]. More recent graph-neural-network and transformer models can reach lower errors, particularly when trained on very large proprietary datasets or narrower method spaces [9, 17], but they do so at the cost of interpretability and engineering complexity, and the comprehensive benchmark of Bouwmeester et al. found no single algorithm universally best across datasets, with performance governed more by chemical-space coverage than by model class [14]. Indeed, several influential RT and QSRR predictors are built on random-forest ensembles rather than boosting [7, 15, 33], underscoring that the tree-ensemble family as a whole—not any one implementation—is well matched to tabular structure–retention data. A distinct and complementary line of work addresses the transfer of SMRT-trained predictions onto other chromatographic systems through projection and calibration—PredRet's direct inter-system mapping, MultiConditionRT's explicit method descriptors, and post-projection calibration—which remains essential because a single-method model such as ours predicts RT only for the specific separation on which SMRT was acquired [13, 15, 18, 19].

4.4 Limitations

Three limitations bound the interpretation of our results. First, the model is method-specific: it predicts retention for the single RP gradient underlying SMRT and cannot be applied to another column or gradient without projection or calibration [13, 18]. Second, the representation is two-dimensional; Morgan fingerprints and topological descriptors do not encode three-dimensional conformation or stereochemistry, so retention differences between stereoisomers or conformers that share a 2D structure are, in principle, invisible to the model. Third, the applicability domain is bounded on the low-RT side, as the early-eluter analysis shows, and—because SMRT is dominated by drug-like and metabolite-like small molecules—predictions for chemotypes sparsely represented in the training space should be treated with appropriate caution. We also note that our reliance on the published 2024 exclusion list means our training distribution inherits whatever assumptions underlie that filter; because we report the filter explicitly and reproduced the flagged set exactly, downstream users can reproduce or vary this choice [20].

4.5 Implications for metabolite annotation

For untargeted LC–MS annotation the operational message is twofold. Within the well-retained domain, an interpretable structure-based predictor delivers RT estimates precise enough to serve as a genuine orthogonal filter: with a median absolute error of 25 s and roughly four in five compounds predicted within 60 s, predicted RT can meaningfully rank or eliminate isobaric candidates that mass spectrometry alone cannot separate, complementing spectral and accurate-mass evidence within community identification-confidence frameworks [4, 7, 8, 34]. At the same time, the early-

eluter analysis is a reminder that such filters must be applied with an explicit sense of domain: for features eluting near the void volume, predicted RT is unreliable and should not be used to reject candidates. Reporting per-compound predictions, an honest error profile, and the boundaries of the applicability domain—as we have done here—is therefore not merely good practice but a prerequisite for using RT prediction safely in identification pipelines.

5 Conclusion

Starting from the original 2019 METLIN SMRT release and explicitly applying the 2024 erroneous-entry filter (removing exactly 1,299 flagged structures en route to 78,629 curated compounds), we built a compact, reproducible, interpretable gradient-boosted-tree model that predicts reversed-phase retention time from molecular structure. On a 20% compound-level hold-out (seed 42) the model attained an MAE of 45.99 s, a median absolute error of 25.09 s, $R^2 = 0.842$, and 79.2% of compounds predicted within 60 s of the measured value, with complete per-compound predictions provided. The model's decisions are chemically legible—dominated by `MolLogP` and by partial-charge- and surface-area-weighted polarity descriptors that encode the hydrophobic-partitioning basis of RP separation—and its failures are equally legible: poorly retained analytes eluting within the first 300 s are systematically over-predicted (subgroup MAE 342 s), a bounded and predictable applicability-domain effect that a simple RT $>$ 400 s cut-off largely resolves. Taken together, the work shows that a well-curated benchmark and a transparent tabular model can reproduce the retention behaviour of a large, diverse chemical space while making explicit exactly where predicted retention time can, and cannot, be trusted as an aid to small-molecule identification.

Data and Code Availability

The METLIN SMRT dataset is publicly available from its 2019 Figshare deposition [16]; the 2024 erroneous-entry exclusion list is from Khrisanfov et al. [20]. The curated dataset (78,629 compounds), the complete per-compound hold-out predictions (`test_predictions.csv`, 15,726 rows), the full ranked feature-importance table (all 2,259 features), the cross-validation and tuning logs, the structured metric/residual results, and the analysis scripts accompany this manuscript.

Conflicts of Interest

The authors declare no competing interests.

References

[1] Colin A. Smith, Elizabeth J. Want, Grace O'Maille, Ruben Abagyan, and Gary Siuzdak. XCMS: Processing mass spectrometry data for metabolite profiling using nonlinear peak alignment, matching, and identification. *Analytical Chemistry*, 78(3):779–787, 2006. doi: 10.1021/ac051437y.

[2] Warwick B. Dunn, David Broadhurst, Paul Begley, Eva Zelena, Sue Francis-McIntyre, Nadine Anderson, Marie Brown, Joshua D. Knowles, Antony Halsall, John N. Haselden, et al. Procedures for large-scale metabolic profiling of serum and plasma using gas chromatography and

liquid chromatography coupled to mass spectrometry. *Nature Protocols*, 6(7):1060–1083, 2011. doi: 10.1038/nprot.2011.335.

[3] Gary J. Patti, Oscar Yanes, and Gary Siuzdak. Metabolomics: the apogee of the omics trilogy. *Nature Reviews Molecular Cell Biology*, 13(4):263–269, 2012. doi: 10.1038/nrm3314.

[4] Emma L. Schymanski, Junho Jeon, Rebekka Gulde, Kathrin Fenner, Matthias Ruff, Heinz P. Singer, and Juliane Hollender. Identifying small molecules via high resolution mass spectrometry: Communicating confidence. *Environmental Science & Technology*, 48(4):2097–2098, 2014. doi: 10.1021/es5002105.

[5] Carlos Guijas, J. Rafael Montenegro-Burke, Xavier Domingo-Almenara, Amelia Palermo, Benedikt Warth, Gunda Hermann, Gunda Koellensperger, Tao Huan, Winnie Uritboonthai, Aries E. Aisporna, Dennis W. Wolan, Mary E. Spilker, H. Paul Benton, and Gary Siuzdak. METLIN: A technology platform for identifying knowns and unknowns. *Analytical Chemistry*, 90(5):3156–3164, 2018. doi: 10.1021/acs.analchem.7b04424.

[6] David S. Wishart, AnChi Guo, Eponine Oler, Fei Wang, Afia Anjum, Harrison Peters, Raynard Dizon, Zinat Sayeeda, Siyang Tian, Brian L. Lee, et al. HMDB 5.0: the Human Metabolome Database for 2022. *Nucleic Acids Research*, 50(D1):D622–D631, 2022. doi: 10.1093/nar/gkab1062.

[7] Anna M. Wolfer, Sylvain Lozano, Thomas Umbdenstock, Vincent Croixmarie, Alban Arrault, and Philippe Vayer. UPLC-MS retention time prediction: a machine learning approach to metabolite identification in untargeted profiling. *Metabolomics*, 12(1):8, 2016. doi: 10.1007/s11306-015-0888-2.

[8] Paolo Bonini, Tobias Kind, Hiroshi Tsugawa, Dinesh Kumar Barupal, and Oliver Fiehn. Retip: Retention time prediction for compound annotation in untargeted metabolomics. *Analytical Chemistry*, 92(11):7515–7522, 2020. doi: 10.1021/acs.analchem.9b05765.

[9] Yuting Liu, Akiyasu C. Yoshizawa, Yiwei Ling, and Shujiro Okuda. Insights into predicting small molecule retention times in liquid chromatography using deep learning. *Journal of Cheminformatics*, 16(1):113, 2024. doi: 10.1186/s13321-024-00905-1.

[10] Roman Kaliszan. QSRR: Quantitative structure-(chromatographic) retention relationships. *Chemical Reviews*, 107(7):3212–3246, 2007. doi: 10.1021/cr068412z.

[11] Paul R. Haddad, Maryam Taraji, and Roman Szűcs. Prediction of analyte retention time in liquid chromatography. *Analytical Chemistry*, 93(1):228–256, 2021. doi: 10.1021/acs.analchem.0c04190.

[12] Yabin Wen, Mohammad Talebi, Ruth I. J. Amos, Roman Szucs, John W. Dolan, Christopher A. Pohl, and Paul R. Haddad. Retention prediction in reversed phase high performance liquid chromatography using quantitative structure-retention relationships applied to the Hydrophobic Subtraction Model. *Journal of Chromatography A*, 1541:1–11, 2018. doi: 10.1016/j.chroma.2018.01.053.

[13] Jan Stanstrup, Steffen Neumann, and Urska Vrhovšek. PredRet: Prediction of retention time by direct mapping between multiple chromatographic systems. *Analytical Chemistry*, 87(18):9421–9428, 2015. doi: 10.1021/acs.analchem.5b02287.

[14] Robbin Bouwmeester, Lennart Martens, and Sven Degroeve. Comprehensive and empirical evaluation of machine learning algorithms for small molecule LC retention time prediction. *Analytical Chemistry*, 91(5):3694–3703, 2019. doi: 10.1021/acs.analchem.8b05820.

[15] Amina Souihi, Modestas P. Mohai, Emma Palm, Louise Malm, and Anneli Kruve. MultiConditionRT: Predicting liquid chromatography retention time for emerging contaminants for a wide range of eluent compositions and stationary phases. *Journal of Chromatography A*, 1666: 462867, 2022. doi: 10.1016/j.chroma.2022.462867.

[16] Xavier Domingo-Almenara, Carlos Guijas, Elizabeth Billings, J. Rafael Montenegro-Burke, Winnie Uritboonthai, Aries E. Aisporna, Emily Chen, H. Paul Benton, and Gary Siuzdak. The METLIN small molecule dataset for machine learning-based retention time prediction. *Nature Communications*, 10(1):5811, 2019. doi: 10.1038/s41467-019-13680-7.

[17] Serhii A. Dymura, Oleksandr O. Viniichuk, Kostiantyn P. Melnykov, Dmytro S. Radchenko, and Oleksandr O. Grygorenko. Machine learning-based retention time prediction tool for routine LC-MS data analysis. *Journal of Chemical Information and Modeling*, 65(14):7415–7425, 2025. doi: 10.1021/acs.jcim.5c00514.

[18] Yue Zhang, Fang Liu, Xiaoqiang Li, Yuan Gao, et al. Generic and accurate prediction of retention times in liquid chromatography by post-projection calibration. *Communications Chemistry*, 7(1):54, 2024. doi: 10.1038/s42004-024-01135-0.

[19] Fleming Kretschmer, Eva-Maria Harrieder, Martin A. Hoffmann, Sebastian Böcker, and Michael Witting. RepoRT: a comprehensive repository for small molecule retention times. *Nature Methods*, 21(2):153–155, 2024. doi: 10.1038/s41592-023-02143-z.

[20] Mikhail Khrisanfov, Dmitriy Matyushin, and Andrey Samokhin. Finding potentially erroneous entries in METLIN SMRT. *Journal of Chromatography A*, 1745:465761, 2025. doi: 10.1016/j.chroma.2025.465761.

[21] Jerome H. Friedman. Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, 29(5):1189–1232, 2001. doi: 10.1214/aos/1013203451.

[22] Tianqi Chen and Carlos Guestrin. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16)*, pages 785–794. ACM, 2016. doi: 10.1145/2939672.2939785.

[23] Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. LightGBM: A highly efficient gradient boosting decision tree. In *Advances in Neural Information Processing Systems 30 (NIPS 2017)*, pages 3146–3154, 2017.

[24] Eugene N. Muratov, Jürgen Bajorath, Robert P. Sheridan, Igor V. Tetko, Dmitry Filimonov, Vladimir Poroikov, Tudor I. Oprea, Igor I. Baskin, Alexandre Varnek, Adrian Roitberg, Olexandr Isayev, Stefano Curtalolo, Denis Fourches, Yoram Cohen, Alan Aspuru-Guzik, David A. Winkler, Dimitris Agrafiotis, Artem Cherkasov, and Alexander Tropsha. QSAR without borders. *Chemical Society Reviews*, 49(11):3525–3564, 2020. doi: 10.1039/D0CS00098A.

[25] Greg Landrum and The RDKit Contributors. RDKit: Open-source cheminformatics. `https://www.rdkit.org`, 2024. Version 2026.03.3.

[26] H. L. Morgan. The generation of a unique machine description for chemical structures–a technique developed at chemical abstracts service. *Journal of Chemical Documentation*, 5(2): 107–113, 1965. doi: 10.1021/c160017a018.

[27] David Rogers and Mathew Hahn. Extended-connectivity fingerprints. *Journal of Chemical Information and Modeling*, 50(5):742–754, 2010. doi: 10.1021/ci100050t.

[28] Scott A. Wildman and Gordon M. Crippen. Prediction of physicochemical parameters by atomic contributions. *Journal of Chemical Information and Computer Sciences*, 39(5):868–873, 1999. doi: 10.1021/ci9903071.

[29] Peter Ertl, Bernhard Rohde, and Paul Selzer. Fast calculation of molecular polar surface area as a sum of fragment-based contributions and its application to the prediction of drug transport properties. *Journal of Medicinal Chemistry*, 43(20):3714–3717, 2000. doi: 10.1021/jm000942e.

[30] Johann Gasteiger and Mario Marsili. Iterative partial equalization of orbital electronegativity–a rapid access to atomic charges. *Tetrahedron*, 36(22):3219–3228, 1980. doi: 10.1016/0040-4020(80)80168-2.

[31] Paul Labute. A widely applicable set of descriptors. *Journal of Molecular Graphics and Modelling*, 18(4-5):464–477, 2000. doi: 10.1016/S1093-3263(00)00068-1.

[32] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[33] Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001. doi: 10.1023/A:1010933404324.

[34] Lloyd W. Sumner, Alexander Amberg, Dave Barrett, Michael H. Beale, Richard Beger, Clare A. Daykin, Teresa W.-M. Fan, Oliver Fiehn, Royston Goodacre, Julian L. Griffin, et al. Proposed minimum reporting standards for chemical analysis. *Metabolomics*, 3(3):211–221, 2007. doi: 10.1007/s11306-007-0082-2.