Journal of Chromatography A 1784 (2026) 467218

Contents lists available at ScienceDirect

Journal of Chromatography A

journal homepage: www.elsevier.com/locate/chroma

Stereo-sensitive modelling of gas chromatographic retention indices of mono-cycloalkanes in jet fuel range

Hannes Lüdtke *, Florian Pütz, Samuel Grams, Thomas Gröger, Barbara Giocastro, Uwe Bauder, Markus Köhler, Andreas Huber, Patrick Oßwald

Institute of Combustion Technology, German Aerospace Center (DLR), 70569 Stuttgart, Germany

ARTICLE INFO

Keywords:
Gas chromatography
Retention index
Isomers
Stereoisomers
Cycloalkanes
Fuel assessment
QSPR / QSRR modelling
GCxGC Fuel Analysis

ABSTRACT

Accurate assessment of novel aviation fuels requires detailed insight about structural differences relative to conventional jet fuels. To support compositional characterization, retention behavior of cycloalkanes needs to be investigated further to evaluate fuels on an isomeric level. Current available literature models for the prediction of saturated cyclic retention indices (RI) are limited in carbon range or by neglection of stereo isomeric effects. This study addresses the challenge of comprehensive RI prediction of mono-cycloalkanes by utilizing available retention data from the NIST23 database and the combination of descriptors from PaDEL, Mordred and RDKit. Multiple model approaches (extreme gradient boosting, multi-layer perceptron regression, support vector regression and ridge regression) were tested via repeated k-fold cross-validation to evaluate robustness and reliability in prediction. As the available literature data was skewed, model tests were run for data up to carbon number C13 and the whole dataset to C20 individually. Cross-validation results showed the best error metrics for the validation sets with ridge regression (e.g. C13-models mean absolute error: MAE = 11.01 ± 1.32 RI; C20-model MAE = 13.68 ± 1.67 RI). Correct stereo sensitive retention of diastereomeric structures was achieved in 78 %, with *cis/trans* differentiation being the best represented stereo effect. These advancements will enable a more detailed characterization of compositional differences and will subsequently enhance evaluation of novel fuel production paths.

1. Introduction

Sustainable aviation fuels (SAF) can reduce climate impact from aviation, targeting both carbon footprint and non-CO_2 effect reduction [1–3]. It is referred to as the most-ready technology in contrast to alternative propulsion system ideas in aviation based on hydrogen or battery. Currently, it is considered to be the most viable solution for the foreseeable future, particularly for long-haul flights [4–6]. The use of SAF percentages is becoming mandatory in accordance to EU regulations. To meet the necessary capacities of SAF, research on novel production routes of SAF is ongoing. These developments are costly and time-consuming, and require early feedback to allow for fast adjustments to the design of compliant SAF products. For process development and fuel formulation the precise, detailed knowledge of the fuels molecular composition is crucial. Without abundant volumes the evaluation of novel candidates is usually done by the so-called prescreening procedure [7–9]. As a first step, the fuel is analyzed on its composition, usually by GCxGC, which serves as input for fuel property predictions to ensure the allowed specification limits of the SAF product. This is especially crucial since SAF is generally narrowly regulated compared to conventional jet fuels.

Regarding the chemical composition, jet fuels commonly include compounds into the carbon range of up to C20 with *n*-/isoparaffins, mono-cycloalkanes and monoaromatics as major chemical classes. The latest CRC fuel survey even indicates mono-cycloalkanes to have the highest median concentration (29.80 mass%) in jet fuel [10–12]. Established SAF/SBCs (synthetic blending components) production routes like Fischer-Tropsch, hydroprocessed esters and fatty acids (HEFA) or alcohol to jet (ATJ) often consist mainly of isoalkanes. The resulting products are therefore often also referred as synthetic paraffinic kerosene (SPK). Novel procedures are starting to explore cycloalkanes as substitute for aromatics due to their potential for non-CO_2 effect reduction whilst maintaining fuel handling relevant properties like o-ring swelling [13–18]. Due to the complex composition of most

* Corresponding author.
E-mail address: Hannes.Luedtke@dlr.de (H. Lüdtke).

https://doi.org/10.1016/j.chroma.2026.467218
Received 24 April 2026; Received in revised form 23 June 2026; Accepted 24 June 2026
Available online 25 June 2026

fuels, a determination of the exact molecular composition is not possible, even for the rather simple isoalkanic SBCs (e.g. FT, HEFA). The state-of the-art evaluation of fuel compositions by GCxGC provide carbon number vs group-type (GT) quantifications only [19–21]. It has been shown by Lüdtke et al. that the isomeric distribution of isoalkanes can be determined beyond the conventional GT-analysis. Especially for novel fuel candidates the composition contains varying degrees of structural isomeric diversity [22]. If the isomer distribution is noticeably different from conventional fuel compositions, fuel prediction models can struggle to properly portray the effect on physical properties (such as distillation behavior, viscosity or cetane number). Depending on the model implementation and the observed fuel property, property variation from isomeric structures can directly impact fuel predictions. It has been shown before, that the model evaluation can be improved by implementation of the specific isomeric detail [9,23]. To improve general fuel evaluation and support development of novel SAF routes, better understanding of specific fuel components and their influence on fuel properties is needed. The investigation on the saturated mono-cycloalkanes is targeted in this work.

For the determination of fuel compositions, GCxGC is typically coupled with mass spectrometry [19–21]. While mass spectrometry can help to a certain degree with identification of group-types it can obtain nearly identical spectra for structurally similar compounds. Limitation of relative amount of database entries reduces certainty of discrimination. These identification uncertainties can be mitigated by retention index (RI) differences, which usually delivers no unique identification but can provide additional context [24]. As retention behavior within hydrocarbon classes correlates with certain properties its implementation for fuel analyses can help understand fuel property changes due to isomeric differences. Increased certainty on the isomeric distribution can eliminate the inclusion of unlikely structures in fuel property predictions, enhancing their precision [9,22,23].

Available reference data usually includes more RI data than mass spectral entries. Nevertheless, in contrast to the number of possible isomers, the relative availability can still be insufficient for complex separation problems. To address this issue, an abundance of varying regression or computational approaches for predictive retention models is available in chromatographic literature. Thereby a common technique is application of the quantitative structure-property relationships (QSPR), which is also known as QSRR if the property of interest is the retention. There, certain functionalities of molecular structures are mathematically described, ranging from simple molecular weight to more complex definitions of topological connectivity or shape-dependent parameters [25–27]. Using the combination of several molecular descriptors, various computational approaches can build connections with physical properties like the RI. Neural networks have gained popularity in recent literature, but typically require representative and comprehensive datasets to prevent overfitting [26–30]. Multi linear regressions have long been successfully used for QSPR and QSRR models, but linearity might not always provide the optimal fit. To capture complex non-linear relationships, tree based models like random forest and extreme gradient boosting (XGB) or kernel based methods like the support vector regression (SVR) are commonly considered techniques for property regression tasks [26,27,30–32]. The literature provides a multitude of RI models that are considering alkanes, alkenes, aromatics, heteroatom containing hydrocarbons [22,33–40]. Several studies also investigated general hydrocarbon mixtures, including subsets of cycloalkane structures [28,29,41–45]. Comprehensive analyses on cycloalkane retention behavior that include the carbon range (C20) necessary for jet fuel evaluations remain scarce. Farkas et al. have pointed out that the possible degrees of molecular complexity within saturated cyclic hydrocarbons are by factors higher to straight chain saturates, making the RI prediction in combination difficult [46]. Additionally, full stereo isomeric sensitivity was usually only implemented into literature RI models, when relying on availability of other physical properties [43,47]. A detailed description of a selection of available literature models with included mono-cycloalkanes is provided in comparison to the present approach in the discussion section. In the following, the generation and evaluation of a comprehensive and reliable method to predict retention of jet-fuel relevant and stereo-sensitive mono-cyclic components will be presented by using molecular descriptors only.

Due to the complex possibilities in the isomerism of mono-cycloalkanes, a short introduction to structural diversity and corresponding isomeric effects on RI of mono-cycloalkanes are explained subsequently. Examples for C9 mono-cycloalkanes isomers are illustrated by Fig. 1.

The initial mono-cycloalkane definition is given by the general chemical formula C_nH_{2n} and the presence of a given (single) ring structure. The so-called constitutional isomers are distinguishable by order of connection. This affects the size of the ring structure, number, type and position of connected sidechains (see methylcyclooctane or propylcyclohexane, Fig. 1). With higher number of substituted positions on a cycloalkane ring the chance of configurational isomers increases. A carbon atom bonded to four distinct substituents is defined as a chiral center. When a compound exhibits multiple chiral centers, the relative orientations of the side chains to one another can cause spatial interactions, possibly influencing the physical properties of a molecule. In cases of high molecular symmetry, cycloalkanes with multiple side chains (e.g., 1,4-dimethylcyclohexane or 1,3,5-trimethylcyclohexane) can also have stereo isomeric diversity even in the absence of chiral centers. Compound pairs with same constitution but different spatial arrangements of the substituents and are termed diastereomers [49]. The spatial orientation is usually portrayed as dashed bond lines for substituents that project backwards from the plane (β herein), while substituents pointing forward out of the plane are illustrated as wedged bonds (α herein; see 1,2,4-trimethylcyclohexanes in Fig. 1). When those bonded substituents are facing towards the same direction it is defined as the *cis*-isomer, opposite directions are defined as *trans*-isomer. With more than two sidechains, the nomenclature becomes more complex. The effect of diastereomeric differences on the retention heavily depends on the structure. Some examples show differences of over 50 RI towards another (e.g. 1-ethyl-2,3-dimethylcyclopentane diastereomers; ααα: 902 RI and αβα: 851 RI [48]), while other cases have barely any reported distinction. If a molecule has chiral centers at the same carbon positions but the spatial arrangement of substituents differs (R/S configuration), the resulting structures can be non-superimposable mirror images and are not identical. Those isomeric pairs are called enantiomers (see βαα and αββ in Fig. 1). Enantiomers rotate the plane of polarized light in different directions, but exhibit identical physical properties in achiral settings. This is usually the case for common gas chromatography columns and consequently enantiomers are not considered as relevant isomers herein. Only if the molecules interact with other chiral environments or modified chiral columns, significant differences can be identified (e. g. physiologic impact on organisms) [50,51]. The impact on structural diversity of the respective isomerism can be seen in Table 1. By constitutional isomers alone, the jet fuel relevant range (until C20) of possible structures contains several million possibilities. Although the total amount of theoretic isomers doesn't generally represent the total number of potentially relevant structures, it serves as illustration of the increased complexity within samples that contain higher boiling mono-cycloalkane fractions. As an example, almost half of the theoretically calculated structures up to C20 are represented by isomers with ring sizes of three or four atoms. These structures would have to be questioned in relevance for fuel analyses due to their high ring strains.

Dicycloalkanes structures were excluded from consideration herein as they show an even higher complexity regarding isomeric diversity (> 2.000.000 constitutional isomers until C16 calculated with MOLGEN [52]) than mono-cycloalkanes and would include additional property effects from steric configuration only of H-atoms (e.g. at the dicyclic bridge of cis-/trans-decalin [53]). Since poly-cycloalkanes are also less prominent in

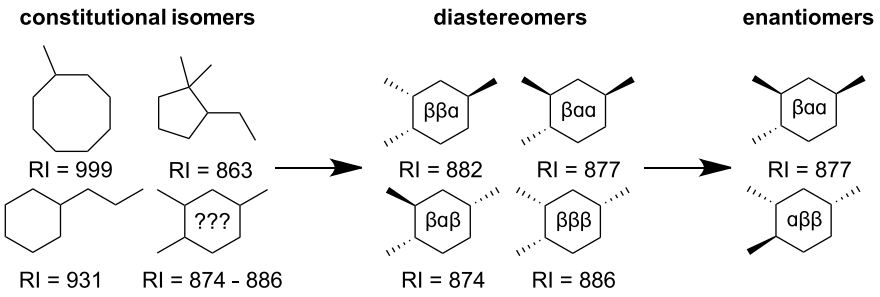

Fig. 1. Illustration of possible isomeric differences and the corresponding retention effect in mono-cycloalkanes. Examples provided for C9, one structure for all four possible diastereomer pairs of 1,2,4-trimethylcyclohexane is shown. One enantiomer pair (αββ & βαα) is exemplary selected. RI taken from NIST23 [48].

Table 1
The possible mono-cycloalkane C_nH_{2n} structures and the available stereo sensitive retention index (RI) data up to $C_{20}H_{40}$.[a.]

nC	Constitutional isomers	Total stereo isomers	Stereo isomers without enantiomer repetition	RI	RI (distinct stereo)
3	1	1	1	1	0
4	2	2	2	2	0
5	5	7	6	6	2
6	12	18	14	9	2
7	29	58	41	25	12
8	73	173	112	30	19
9	185	546	331	92	68
10	475	1,693	977	58	37
11	1,231	5,334	2,975	22	10
12	3,232	16,724	9,097	12	7
13	8,506	52,725	28,158	16	8
14	22,565	166,133	87,668	3	0
15	60,077	524,644	274,203	2	0
16	160,629	1,658,075	861,539	7	0
17	430,724	5,247,658	2,713,112	8	2
18	1,158,502	16,624,945	8,567,302	7	0
19	3,122,949	52,731,978	27,099,092	7	0
20	8,437,289	167,431,350	85,876,656	2	0
Total	13,406,486	244,462,064	125,521,286	309	167

[a] constitutional isomers were calculated with MOLGEN 5.0 and enu. Stereoisomers were calculated with enu. Retention index (RI) data listed was gathered from literature measurements on semi-standard non-polar columns and curated in the NIST 23 database [48,52,54].

jet fuels [10] this study focusses only on the investigation of saturated monocyclic hydrocarbons and there complex isomeric diversity.

2. Methods

2.1. Collection of available RI data

Since commercially available reference standards are limited, a more comprehensive retention data set of mono-cycloalkanes was collected from the curated NIST23 (National Institute of Standards and Technology) retention database [48]. The goal was to utilize only data with correct illustration of stereo variability. The most commonly used column setup for reported retention data was the semi-standard non-polar (357 up to C20). This includes columns like DB-5, BPX-5 and Squalane. RI data on non-polar columns was limited (150 up to C20) and mainly available for the same isomers as semi-standard. Small polarity effects were noticed on mono-cycloalkane RI. Due to observed inconsistencies in retention behavior among isomers, no general inclusion of non-polar data points was applied.

While the isoalkane isomers are usually reported with constitutional isomer detail, the mono-cycloalkanes are often listed by steric differentiation. Structures with one or no sidechains or structures with two sidechains at the same ring position have no RI relevant stereo structure variations. These constitutional structures are included in the herein used dataset (e.g. 1,1-dimethylcyclohexane, cyclooctane or pentylcyclopentane). To ensure accurate representation of stereochemical sensitivity, retention index entries for isomers missing configurational information were removed (e.g. 1,4-dimethylcyclohexane without *cis/trans* information). This reduced the dataset from C3-C20 mono-cycloalkanes to 309 unique datapoints with the applied filter for stereo sensitivity. Depending on the abundance of literature sources per structure, uncertainties were given as standard deviation. Uncertainties were available for 207/309 isomers (mean $\Delta = 3.7$ RI, median $\Delta = 3.0$ RI). Despite looking into stereo effects from ring connected substituents, stereo effects can also be present in complex sidechain substitutions. Some studies have shown this effect to be relevant for long chain isomers [55–57]. This is rarely considered in databases and therefore difficult to assess individually. Ring structure effects like the distinction between the chair- or boat conformation of cyclohexanes are not represented by the data and believed to be within an equilibrium-averaged molecule as conformation changes are rapid and need low activation energy [58]

The final available data included 309 RIs, 142 of those entries only included constitutional differences. Zero or only one sidechain was present in 109/142 cases, while also including more complex substituents like isopropyl or tert-butyl functionalities. Among the 18 out of 142 constitutional structures with two sidechains, the maximum number of chiral centers was one, due to both sidechains being attached to the same ring position not resulting in diastereomeric differences. For the remaining 15/142 even three or four sidechains were connected, but the connections again allowed a maximum of one chiral center so no further diastereomeric diversity was present. Within the stereo sensitive data (167/309 data points, see Tab. 1), the main steric effects were described by a simple *cis/trans* related distinction (123/167). This translated to 58 available diastereomeric data pairs and seven structures without a reported stereo partner. In 53/58 of the *cis-trans* pairs only two sidechains were present. The other five examples had three sidechains (3 pairs) or four sidechains (2 pairs) with only two diastereomeric options due to multiple substitutions on the same ring position. In contrast, stereo isomeric groups with higher diversity were rarely reported in the database and not always included all possible diastereomers. The dataset included nine constitutional isomer groups with three sidechains, creating multiple diastereomers (37/167). One example is 1,2,4-trimethylcyclohexane, with four possible diastereomeric structures (4 enantiomeric pairs, see Fig. 1). The NIST database included only one example with multiple entries for a complex diastereomeric diversity of four sidechains (1,2,3,4-tetramethylcyclopentane). For this constitutional isomer, 8 diastereomer pairs are possible, but only five unique RI values are given plus one pair of enantiomeric data (7/167). Deviating enantiomeric entries were found in seven stereo cases. Since no definite quality evaluation was possible, all 14 datapoints were kept and used with the given uncertainty. The total used dataset up to carbon number C20 is listed in Tab. 1.

The distribution of retention data over the separation space is further illustrated in Fig. 2. The retention index data for mono-cycloalkanes is mainly available until around 1200 RI (C10). After the carbon number

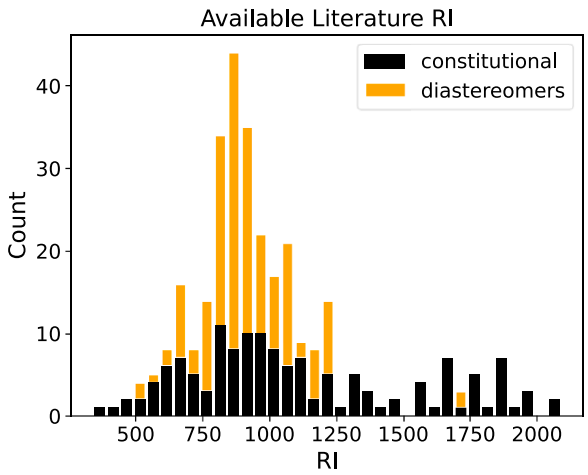

Fig. 2. Illustration of retention index (RI) distribution over the retention space. Constitutional isomers without stereo diversity (142/309) in black and stereo sensitive diastereomers (167/309) in orange.

C13 the amount of datapoints is very scarce and strongly influenced by a few individual structures. The retention data for high carbon range is mainly build from non-stereo containing structures with only one stereo sensitive pair with two substituents, while the rest only consists of large rings without any sidechains or one sidechain structures. The limited availability of structural diversity above C13 is expected to lead to higher uncertainty in prediction attempts for the higher carbon ranges. To achieve optimal fits and minimize bias from the limited number of heavy molecules, modelling is performed on the well-represented dataset up to C13 (273/309) individually. For the prediction of higher carbon numbers, the full dataset (C20) is used, since no reliable C14+ model is feasible from only 36/309 available RI in that range.

2.2. Generation of molecular structures and descriptors

To accomplish stereo-sensitivity in QSPR models, molecular descriptors must capture stereo-specific structural features and spatial differences of isomers accordingly. The creation of 3D coordinates and structure data files (SDF) for molecular structures was explored. RDKit, Open Babel and PaDEL tools as well as the webtool of the CAAD Group of the National Cancer Institute (NCI; running on CACTVS and CORINA) [59–66] were tested based on structure generation from SMILES (simplified molecular-input line-entry system) keys. The SMILES notation is as a natural grammar for chemical structure specifications, which was designed to improve interpretability for both humans and computers [67]. A SMILES code is explicit for one structure, but the same isomer could be described by different grammar. Surprisingly, most tools generated partly varying molecule coordinates with repeated test runs, independent of the allowed iteration count for geometric optimization. It was found that only the NCI tool was able to reproduce structures repeatedly when given the identical SMILES input. For consistency over the whole dataset, the three-dimensional SDFs were generated by the NCI tool and provided as input for the molecular descriptors. Descriptors were calculated with the python modules of PaDELPy, RDKit and Mordred [59,60,68]. Internal coordinate adjustments or optimizations were disabled for PaDELPy and RDKit. Some of the generated descriptors are given by multiple descriptor tools, but overall, a quantity of around 2500 descriptors was generated for the available dataset molecules. Duplicated descriptors were detected by name and value repetitions.

2.3. Selection of molecular descriptors

In order to make a selection of suitable molecular descriptors (i.e. molecular features) for RI prediction, the descriptors were first prepared by robust scaling. Descriptors with zero or low variance were excluded directly. The remaining descriptors were tested on the full RI dataset, as well as on different subsets. This evaluation process of RI subsets involved the individual assessment of carbon numbers, as well as constitutional and stereo-specific isomer sets. Additionally, the data subsets were assessed individually using only 2D, only 3D or only descriptors originating from one specific tool (PaDELPy, RDKit or Mordred). Metrics to evaluate descriptors were feature importance scores from Spearman rank [69] and Pearson [70] correlations with the target variable, as well as random forest and gradient boosting rankings [71, 72]. Different test runs were conducted with and without allowance of co-correlating features above $|R^2| > 0.8$. Through this prefiltering a reduced list of 84 features was generated with the ability to describe both, general retention, as well as constitutional or stereo specific retention effects. The reduced feature matrix was tested again against the full constitutional, and stereo including data subsets. The descriptors with highest scores within Pearson correlation, Spearman correlation, random forest or gradient boosting were kept and extended by some already reported features from literature [47,73,74]. From the remaining 40 descriptors different regression models were tested with varying feature combinations. The descriptor sets used for model comparison (described in chapter 2.4) included 15 descriptors for C13 and C20 test cases (see Tab. S1). Following the initial model comparisons, the descriptor sets of the best-performing model architecture was further evaluated using bootstrap resampling (described in chapter 2.5) for each carbon range (final descriptors listed and described in chapter 3.1.4 in the results section).

2.4. Examination of model types

To evaluate the predictive performance of the chosen feature matrices (Tab. S1) across the C13 and C20 datasets, a selection of modelling approaches was examined. The compared models were extreme gradient boost (XGB), support vector regression (SVR) with two different kernels (linear and radial basis function; rbf), a neural network (NN), as well as the ridge cross validation (RidgeCV). Scikit-learn (1.4.2) [71,72] python module was used for SVR, NN (multi-layer perceptron regressor) and RidgeCV, while XGB tests were conducted with the python module xgboost 3.0.2 [75]. XGB is not affected by scaling factors, but RidgeCV, SVR and NN models need scaled input data. The data preparation was implemented consistently before every model approach test. The descriptor array was scaled using the standard scaler, which transforms the data to have a mean value of zero and a standard deviation of one. A hold-out test selection of 20% was randomly sampled from the respective RI datasets (C13: 55/273; C20: 62/309). Hyperparameters were optimized for SVR, NN and XGB with a GridSearchCV on the remaining 80% of training data (C13: 218/273; C20: 247/309). This way, a five-fold cross validation was performed on all possible parameter combinations within the given grid, targeting the lowest possible mean squared error (MSE) of prediction (resulting parameters see Tab. S2).

The five model architectures were then evaluated with the optimized hyperparameters on the training data. To reduce the dependence of performance metrics on a single random seed, a five-fold cross-validation was performed with 10 random seeds, resulting in 50 unique training (C13: 175/218; C20: 198/247) and validation folds (C13: 43/218; C20: 49/247). This way, the model stability could be assessed on average performance and standard deviations across all 50 folds. The training metrics reflect how well the respective model can fit the given data. Validation metrics help to detect overfitting as they assess the model's generalization ability on data not used during the training phase within each cross-validation fold. As individual error metrics can provide different insights into the predictive capabilities of a model, each model was evaluated on multiple metrics. The evaluated error metrics were the mean squared errors (MSE; penalizes outliers), mean absolute errors (MAE; rates errors linearly), median absolute

Table 2
Performance metrics of the five compared model approaches are listed. Training (train) and validation (val) set metrics are given with standard deviation and derived from repeated k-fold cross validation on 80% of the dataset until C13 (218/273). MSE: mean squared error, MAE: mean absolute error, MedAE: median absolute error, MRE: mean relative error.

C13 models	Unit	RidgeCV	SVR lin	SVR rbf	NN	XGB
MSE train	RI^2	167.39 ± 10.84	176.09 ± 11.93	162.50 ± 7.96	228.92 ± 151.51	1501.05 ± 91.46
MAE train	RI	10.05 ± 0.35	9.77 ± 0. 35	10.60 ± 0.32	11.09 ± 3.56	20.77 ± 0.43
MedAE train	RI	7.88 ± 0.39	7.12 ± 0.52	9.66 ± 0.63	8.76 ± 3.23	12.50 ± 0.65
MRE train	%	1.13 ± 0.04	1.09 ± 0.04	1.22 ± 0.04	1.29 ± 0.41	2.65 ± 0.09
MSE val	RI^2	210.13 ± 46.53	219.49 ± 53.35	242.18 ± 66.20	416.68 ± 214.64	2062.58 ± 1244.48
MAE val	RI	11.24 ± 1.35	11.41 ± 1.37	12.00 ± 1.73	15.29 ± 3.40	26.26 ± 6.39
MedAE val	RI	8.99 ± 1.61	9.02 ± 1.48	9.64 ± 1.96	12.36 ± 3.36	16.47 ± 3.07
MRE val	%	1.29 ± 0.18	1.29 ± 0.17	1.39 ± 0.23	1.80 ± 0.38	3.32 ± 1.19

error (MedAE; non-responsive to outliers; indicates symmetry of error distribution in context to MAE) and mean relative error (MRE; scale independent) from the predicted value to the experimental/literature RI. The model comparisons for the C13- and C20-dataset are discussed subsequently below.

2.5. Bootstrapping and final optimization of selected models

Among the models evaluated, the one with the highest performance on the predefined descriptor set was selected for further analysis in each carbon range. Initial descriptor significance and 95% confidence intervals (CI; using the 2.5th and 97.5th percentile) for model coefficients were assessed for the selected models, using bootstrap resampling (n = 5000) [76,77]. Based on these results, descriptors deemed statistically insignificant were removed to optimize the descriptor subsets. The cross-validation procedure described in chapter 2.4 was then applied to these optimized descriptor subsets to reevaluate model stability for each carbon range. The models derived from full training sets were subsequently evaluated on the independent hold-out datasets to assess its external performance and measure the predictive ability on unseen data. Due to data scarcity, all final model fitting and repeated bootstrap resampling (n = 5000) were conducted on the complete respective datasets without further splitting. Final parameter uncertainty was reported with mean values and 95% CIs. Additionally, mean RI and their corresponding 95% CI and were derived for each compound were derived from the out-of-bag (OOB) predictions generated during bootstrap resampling iterations [78].

2.6. investigation of applicability domain (AD)

The theoretical chemical space of possible mono-cycloalkane isomers is extensive, yet the available dataset is limited in comparison. The applicability domain (AD) was defined using the leverage method, identifying potential extrapolation issues where new compounds would exceed the structural range of the training data [79,80]. Williams plots were generated by plotting the standardized prediction residuals of the available data against the leverage (h) values both model ranges and the combined model set. This visualization characterizes response outliers (i.e. $> \pm 3$ standard residuals) and assesses the structural proximity of each isomer relative to the centroid of the training descriptor space. The leverage threshold is defined as $h^* = 3(p + 1)/n$, with p being the number of selected descriptors and n representing the number of datapoints. Since experimental RI values are unavailable for the theoretically enumerated stereo isomers, these compounds cannot be included in the Williams plot. As an estimate of descriptor-space extrapolation, leverage values were computed for of all theoretically enumerated stereoisomers from C3-C12 (24,820 isomers) and compared with the applicability domain threshold (h^*) of the C13 model.

3. Results

3.1. General model performance comparison

The following subsections detail the performance of the different model architectures across the tested data splits, concluding with the final model specifications.

3.1.1. Evaluation of model performances up to carbon number C13

The RidgeCV model slightly outperformed the linear SVR model on the average validation set in the cross-validation procedure of the small dataset (C13 see Tab. 2). Both SVR models had good training metrics but marginally worse performances on the validation sets. When using the linear kernel approach for SVR, better results were achieved consistently than with the non-linear rbf kernel. The non-linear approaches were less

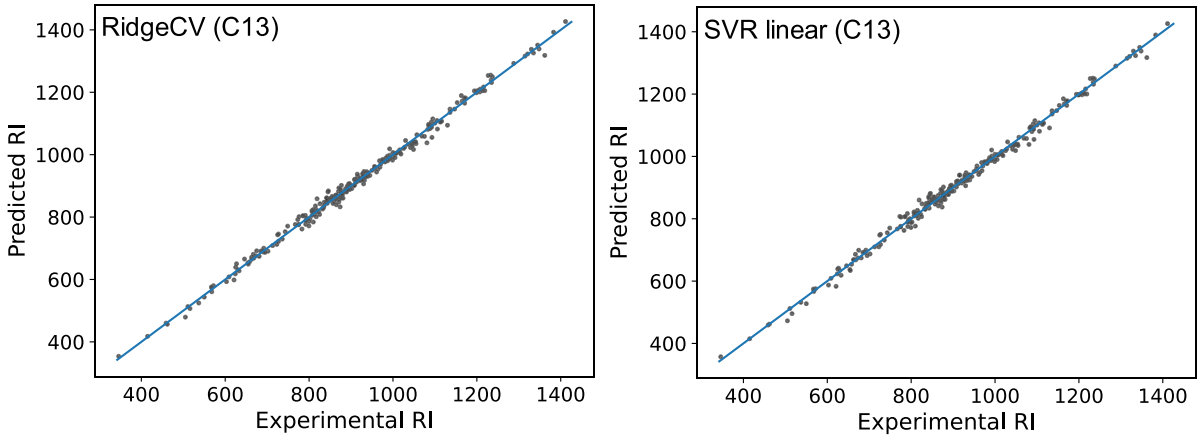

Fig. 3. Illustration of predicted retention indices against the experimental database retention indices for the dataset up to the carbon number C13. Results for RidgeCV and SVR (linear). Unity Line given in blue.

Table 3
Performance metrics of the five compared model approaches are listed. Training (train) and validation (val) set metrics are given with standard deviation and derived from repeated k-fold cross validation on 80% of the full dataset to C20 (247/309). MSE: mean squared error, MAE: mean absolute error, MedAE: median absolute error, MRE: mean relative error.

C20 models	Unit	RidgeCV	SVR lin	SVR rbf	NN	XGB
MSE train	RI^2	280.30 ± 30.00	324.04 ± 32.37	348.73 ± 37.57	346.13 ± 257.74	2818.73 ± 105.13
MAE train	RI	12.45 ± 0.58	12.67 ± 0.43	13.03 ± 0.44	13.41 ± 4.08	29.99 ± 0.56
MedAE train	RI	10.14 ± 0.64	9.57 ± 0.56	9.10 ± 0.65	10.61 ± 3.55	18.21 ± 1.13
MRE train	%	1.31 ± 0.05	1.32 ± 0.04	1.32 ± 0.04	1.43 ± 0.46	3.53 ± 0.09
MSE val	RI^2	444.87 ± 307.48	492.96 ± 281.25	491.03 ± 215.18	717.43 ± 469.20	3973.20 ± 1952.14
MAE val	RI	14.00 ± 1.95	14.78 ± 2.10	15.08 ± 1.96	18.43 ± 4.50	36.73 ± 7.24
MedAE val	RI	10.92 ± 1.32	11.18 ± 1.61	11.07 ± 1.77	14.26 ± 3.62	21.51 ± 3.40
MRE val	%	1.45 ± 0.16	1.52 ± 0.17	1.50 ± 0.15	1.95 ± 0.47	4.11 ± 1.10

accurate overall, as the NN and XGB models could not achieve comparable metrics to RidgeCV and SVR (linear), especially for the validation set. The XGB metrics were noticeably worse than all alternative model architectures. NN was consistently less accurate than SVR (rbf). Standard deviations for NN and XGB further indicated that performances were highly affected by the cross-validation splits. The standard deviations for RidgeCV and SVR models were generally comparable, but dependent on the observed individual metric. Both linear approaches were noticeably more stable than the SVR with rbf kernel as well and showed good functionality over the prediction range (see Fig. 3). In summary, the multi-linear model architectures showed better ability of generalization, while also providing the most stable metrics. The RidgeCV model demonstrated slightly improved performances in terms of accuracy and standard deviation of the MSE and MAE for the validation set, when compared to the linear SVR.

3.1.2. *Evaluation of model performances up to carbon number C20*

In general, the inclusion of the data points above the carbon number of C13 negatively affected all assessed error metrics noticeably (see Tab. 3). Not only the absolute metrics number deteriorated, but also the standard deviation within the cross validation increased. The MSE of the validation was affected the most in relation to C13. The high standard deviation demonstrates difficulties from the inclusion of some heavy structures, when they were not considered in training data of the respective cross-validation fold. The standard deviation of the other metrics in contrast are not affected as strongly, suggesting the MSE change to be mainly affected by individual outliers. The advantages of the RidgeCV over the competing models are more pronounced across almost all validation metrics and are further supported by the standard deviations, indicating greater generalization and robustness independent of cross-validation folds. For the SVR models the choice of the kernel was almost irrelevant, only showing noticeable differences for the MAE metric in preference of the linear SVR. The NN and XGB approaches demonstrated no suitability again, with less precision on both the training and validation data, indicating difficulties for complex architectures with limited datasets. Fig. 4 illustrates that despite acceptable overall performance metrics, both linear models struggled to properly implement the sparse additional data of heavy molecules. For the linear SVR model this was apparent by the loss in linearity and monotony represented by almost horizontal step-like patterns in the prediction above RI's of 1500. This indicated a lack of predictive resolution for heavier molecules within the SVR framework. RidgeCV also showed higher difficulties in precisely modelling the heavy molecules, but recreated the general tendencies found in the experimental RIs more accurately.

Overall, RidgeCV was determined to demonstrate the most consistent modelling outcomes within the established error metrics and their variability for the present dataset. This model was selected as the preferred approach for both datasets (C13 & C20) and will be used for further subsequent discussions.

3.1.3. *RidgeCV model optimization and final performances on hold-out test*

Cross-validation demonstrated that RidgeCV outperformed all other models on both datasets (C13 and C20). Bootstrap-based uncertainty analysis indicated that the descriptors FNSA2, JGT10, MDEC-33 and Vu did not significantly contribute to the C13 model as their coefficient confidence intervals included zero. Consistent with this, individual feature elimination tests confirmed that these descriptors could be excluded without compromising model performance. For the C20 model, the descriptors WTPT-1 and FMF were similarly found to be non-essential within RidgeCV, as their removal did not decrease predictive accuracy. The final optimized RidgeCV model results are presented in Tab. 4.

The corresponding hold-out test sets were used for final evaluation

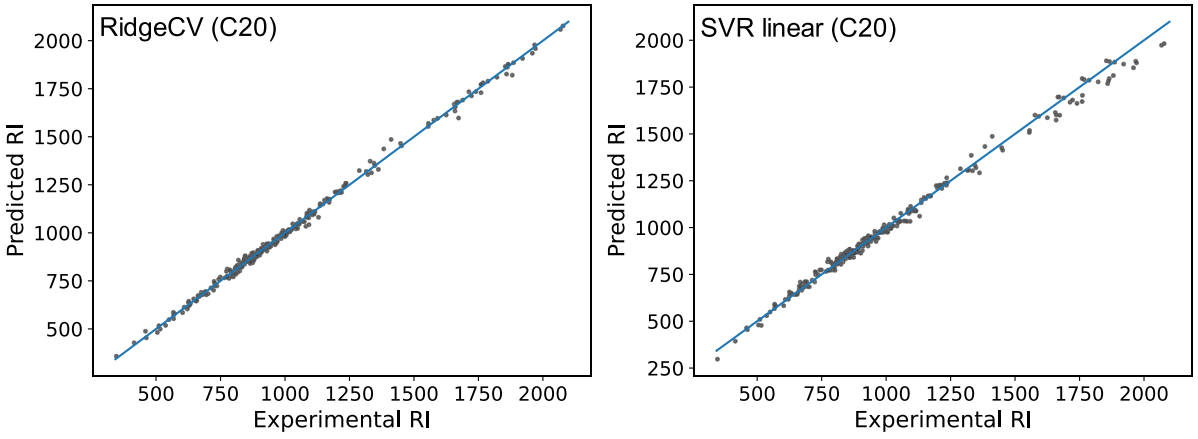

Fig. 4. Illustration of predicted retention indices against the experimental database retention indices for the dataset up to the carbon number C20. Results for RidgeCV and SVR (linear). Unity Line given in blue.

Table 4
Performance metrics of the adjusted RidgeCV models after bootstrapping. Training (train) and validation (val) set metrics are given with standard deviation and derived from repeated k-fold cross validation on 80% of the dataset until C13 (218/273) or the full dataset to C20 (247/309). Hold-out test metrics of the RidgeCV models are listed below (C13: 55/273; C20: 62/309).

Optimized Models	Unit	RidgeCV (C13)	RidgeCV (C20)
MSE train	RI^2	174.06 ± 11.84	278.40 ± 25.42
MAE train	RI	10.15 ± 0.36	12.47 ± 0.41
MedAE train	RI	7.85 ± 0.40	10.23 ± 0.42
MRE train	%	1.14 ± 0.04	1.31 ± 0.04
MSE val	RI^2	204.81 ± 49.28	366.72 ± 145.62
MAE val	RI	11.01 ± 1.32	13.68 ± 1.67
MedAE val	RI	8.87 ± 1.62	11.02 ± 1.55
MRE val	%	1.25 ± 0.17	1.44 ± 0.15
Hold-out test			
MSE	RI^2	154.84	225.57
MAE	RI	9.31	11.52
MedAE	RI	7.39	10.58
MRE	%	1.13	1.26

on external data excluded during optimization (see Tab. 4 and Fig. S1). Results demonstrate strong performance on unseen data, consistent with cross-validation metrics. This indicates sufficient capability of generalization by both models. The C13 model again exhibited lower overall errors. When applying the C20 model function only on the hold-out test molecules C13 and below the test metrics (MSE: 238.58 RI^2, MAE: 11.91 RI, MedAE: 10.63 RI, MRE: 1.33 %) were found to be worse than for the individual C13 model fit as well. It suggests that the C13 model is better at recreation of small steric differences that weren't met as effectively by the full model. Thus, it is proposed to apply the RidgeCV (C20) model only for structures above the carbon number C13 and use the C13 model for more detailed small structure evaluations.

3.1.4. *final model formulas (RidgeCV)*

For the reported model functions, the full respective datasets were used as training set for the RidgeCV model approaches. Model functions were generated with scaled datasets. Coefficients are presented in scaled form for importance evaluation as well as in rescaled form to allow direct application without the need for descriptor scaling. Descriptor set optimization reduced the C13 model to 11 descriptors, whereas 13 descriptors were retained in the C20 model. Features within both sets overlapped substantially, giving a total of 16 relevant descriptors across both models. The descriptors are listed with coefficients in Tab. 5. The lists partially include co-correlated features. Yet, it was found that especially the 3D descriptors could provide varying individual differentiation of stereoisomeric separation. Thus, they were used as combined input data to help fitting stereo isomeric effects. The bootstrap-based 95% confidence interval results for the coefficients are presented within Tab. S3 in the supplement.

Stereospecific distinction was mainly provided by the 3-dimensional descriptor class of differently weighted (W) or fractionated (F) positively (P) or negatively (N) 'charged partial (S) surface areas (A)' (WNSA, FPSA, FNSA), developed by Stanton and Jurs who already used them for modeling of different physical properties [73]. Molecular distance edge vector descriptors (MDEC) have been previously reported for the estimation of boiling points by Liu et al. [74]. They describe the proximity of carbons with specific substitution levels to one another. Global descriptions of molecule structure are further provided by the total walk counts (TWC) and additional assignment of weightings for specific bond types (WTPT-1) [25,81]. Retention is naturally calibrated by molecular weights (MW) but impact of ring size and number of branches ([CX4H3]) on physical properties was also already reported elsewhere [46,82]. Platts et al. proposed a structural group combination approach to create molecular linear free energy relation descriptors (MLFER_L) [83]. It was found to be highly correlated to the overall retention behavior but only roughly differentiates between constitutional isomers. The 'AATSC3i' descriptor is an average Broto-Moreau autocorrelation index that considers the topological distances of the third order in the molecular graph, weighted by the ionization potential [25,60]. The FMF descriptor was developed by Yang et al. and describes the molecules complexity based on the size of the molecules framework weighted by the total molecular size [84,85]. Further details about individual descriptors can be found in the respective publications and descriptor calculation software packages [59,60,68]. The list of included stereo isomers is available in the supplementary information along with C13- and C20 model predictions as well as descriptor values and bootstrapping results.

3.2. *Evaluation of model sensitivity on stereo isomerism*

As the declared objective was to correctly map stereoisomeric effect on retention, stereo containing retention data was analyzed independently from the total model performance. In general, stereo specific retention data in the NIST database was mainly reported for the C13 model range (165/167). The stereo-specific compounds (34/55) within the hold-out test of the RidgeCV (C13; 165/273) model exhibited the error metrics MSE: 180.10 RI^2; MAE: 9.92 RI; MedAE: 7.58 RI; MRE: 1.23 %. For the model up to C20 data only two new stereo specific structures (1-decyl-2-methyl-cyclohexane; *cis/trans*) were available, so the relative amount of stereo sensitive data was decreased (167/309). The stereo-specific hold-out test set compounds (33/62) of the C20 model exhibited the error metrics MSE: 233.66 RI^2, MAE: 12.05 RI, MedAE: 11.55 RI, MRE: 1.38 %.

For both model ranges the respective error metrics slightly worsened in contrast to the total hold-out set. Yet, the errors are still in good comparison to the validation metrics and show that the overall

Table 5
Descriptor coefficients for the RidgeCV models in C13 and C20 range. Respective software packages for calculation were PadelPy$^{(p)}$, Mordred$^{(m)}$ and RDKit$^{(r)}$. All descriptors exhibiting 3-dimensional distinction are marked with an asterisk (*) [59,60,68].

C13 model	RidgeCV Scaled	RidgeCV Rescaled	C20 model	RidgeCV Scaled	RidgeCV Rescaled
*WNSA3$^{(m)}$:	5.66949E+01	1.77773e+02	*FPSA2$^{(m)}$:	−1.64728E+03	−1.08327E+04
MLFER_L$^{(p)}$:	−1.24105E+03	−1.37281e+03	MW$^{(r)}$:	5.08889E+03	1.16188E+02
*FPSA2$^{(m)}$:	−2.48010E+02	−2.92925e+03	MDEC-22$^{(p)}$:	−1.26152E+02	−1.71539E+01
MW$^{(r)}$:	1.48264E+03	5.85249e+01	MLFER_L$^{(p)}$:	−2.81984E+03	−1.76198E+03
WTPT-1$^{(p)}$:	2.56670E+02	7.00941e+01	*FPSA4$^{(m)}$:	−9.90124E+01	−1.47615E+05
*FPSA4$^{(m)}$:	−8.72197E+01	−1.47523e+05	*FNSA2$^{(m)}$:	1.94343E+02	1.04581E+04
*FNSA3$^{(m)}$:	7.02486E+01	5.08634e+04	*FNSA3$^{(m)}$:	6.51186E+01	4.05156E+04
ring size$^{(r)}$:	−3.72677E+01	−2.10471e+01	*WNSA3$^{(m)}$:	5.35718E+01	1.04462E+02
[CX4H3]$^{(r)}$:	−3.34953E+02	−3.62193e+02	TWC$^{(p)}$:	−5.24569E+01	−6.29301E+00
FMF$^{(p)}$:	2.59291E+01	4.30180e+02	[CX4H3]$^{(r)}$:	−5.12552E+02	−5.54674E+02
AATSC3i$^{(p)}$:	6.95432E+00	4.31795e+01	ring size$^{(r)}$:	2.39079E+01	1.28278E+01
Intercept:	9.11158E+02	2.73340E+03	MDEC-12$^{(p)}$:	−1.12778E+01	−8.71893E+00
			MDEC-23$^{(p)}$:	−1.22640E+01	−4.89675E+00
			Intercept:	1.01002E+03	2.95132E+03

Table 6
Selection of available literature models considering cycloalkanes within their prediction. Utilized data, descriptors and model techniques are described. Reported original errors were copied from references in column 5. To enable better comparison, uniform metrics (MAE, MedAE, MRE) were herein recalculated from provided data of the references in column 6.

Reference	Dataset	Descriptors	Model approach	Error metrics reported by Reference	New error metrics recalculated from original data from the references
Dimov et al. 1978 [47]	30/30 C5-C9 cycloalkanes	-PCI (physico-chemical index): vapor pressure, molecular volume (solute and reference) -structural number: ring size, number of substituents, side chain length, number of sidechain branches, number of free ring positions, number of specific *cis/trans* structures *stereo sensitive due to distinct **physical property** differences*	Multilinear regression	variance: $s^2 = 0.97$	*no hold-out test* MAE: 0.88 RI MedAE: 0.6 RI
Bermejo et al. 1984 [43]	15/55 C6-C8 cycloalkanes Linear, branched and cyclic alkanes	-physical properties: boiling point, refractive index, density *stereo sensitive due to distinct **physical property** differences*	Multilinear regression	standard deviation $s = 4.1$ *no report on cyclic only errors*	*no hold-out test* all data: MAE: 3.3 RI MRE: 0.49 % MedAE: 2.8 RI cyclo *only data* MAE: 3.87 RI MRE: 0.52 % MedAE: 3.3 RI
Duvenbeck et al. 1993 [42]	C5-C9 cycloalkanes 22/217 datapoints Acyclic- and, cyclic alkanes, alkenes, alcohols, esters, ketones, ethers	-number of identical vertices, the atomic level indices, electrotopological state, bond type indices, ring parameter *no stereo differentiation; use **2D molecular descriptors** only*	Multilinear regression	MAE 7.0 – 16.7 RI Standard errors: 10.2 – 24.8 *no report on cyclic only errors*	*no hold-out test* *best model* cyclo *only data* MAE:11.05 MedAE: 7.35 MRE: 1.54 %
Yan et al. 1998 [29]	Cyclic and acyclic alkanes, Alkenes, alcohols, ethers, ketones, esters Cyclo: 20/184 From C5-C9	-physical properties: boiling point, molecular weight, density *stereo sensitive due to distinct **physical property** differences*	Neural network	"the average relative error is no more than 2%." report **only on test set** *no report on cyclic only errors*	*from hold-out test* MAE: 11.16 RI MedAE: 8.35 RI MRE: 1.52 % *4/34* cyclo*only hold-out test* MAE: 23.73 RI MedAE: 24.1 RI MRE: 3.32 %
Yin et al. 2001 [44]	"55 cyclanes" from C5-C9 *Exported from total dataset of 150 hydrocarbons*	15-bit numerical code (partially stereo sensitive) *limited stereo description (only includes bit for if cis/trans is present), use **molecular descriptors** only*	Multilinear regression	standard deviation (best model) $s = 20.24$ *no report on cyclic only errors*	*no hold-out test* *all data best model* MAE: 15.59 RI MedAE: 12.3 RI MRE: 1.95 % *"Cyclanes" only best model* MAE: 16.32 RI MedAE: 15.38 RI MRE: 2.04 %
Geer et al. 2024 [41]	Total NIST EI-MS spectral libraries of standard semi polar columns	Atomic number, formal charge, number of atoms, bond type, bond conjugation, shortest topological distance along the path, ring descriptors (size, aromaticity) *no stereo differentiation, use **2D molecular descriptors** only*	Neural network	Reported test set metrics: MAE: 15.1 RI MedAE: 8.0 RI *no report on cyclic only errors*	Lit. database C13 (273): MAE: 12.06 RI MedAE: 9.30 RI MRE: 1.34% Lit. database C20 (309): MAE:11.51 RI MedAE: 8.60 RI MRE: 1.23% *no train/test split known*

performance is not substantially affected by separation problems of higher complexity. Still, it indicates that there may be marginally higher uncertainties with stereo effect predictions and better fits to constitutional isomerism.

4. Discussion

4.1. Comparison of model performance to literature

The evaluation of model performances varies strongly depending on the literature reference. The reported performance metrics are generally not the same over all references and therefore difficult to compare against each other. To enable a better comparison to the obtained performance of the presented model herein, MAE, MedAE and MRE were calculated from the original full data and cycloalkane subsets (see last column in Tab. 6), when sufficient experimental data was provided by the literature sources. As not all literature studies have the distinction of assessing modelling performance on hold-out test sets or by cross-validation, reported results in the literature may potentially be biased by the fit to the known training data. Additionally, given performance metrics often describe the total used dataset which is generally not only consisting of mono-cycloalkanes. A selection of reported models from literature can be found in Tab. 6. Mono-cycloalkane specific retention models are not commonly reported. Only Dimov et al. was found to do an individual analysis [47]. In other studies, it is typically only a part of the total dataset with 15–50 considered compounds. As already discussed in the introduction, cycloalkane modelling has been reported to be a more complex problem in contrast to other hydrocarbon families like paraffins, olefins or aromatics [46]. In the Tab. 6 it can be seen that in contrast to the overall model performance the cycloalkane specific results tend to show higher error metrics in the respective references. Throughout the references, the incorporated mono-cycloalkane structure range is mostly limited to small molecular weights, usually only going as high as C9 with the majority of data in the smaller carbon numbers [29,42–44]. Only Geer et al. have incorporated a comprehensive dataset for a wider range of mono-cycloalkanes, using the full available dataset of NIST libraries [41].

Some of the reported models incorporated physical properties to model retention behavior. Common connections for the correlation are the boiling point and vapor pressure, which are in itself closely related to retention behavior [29,43,47]. For those model cases, stereo sensitive results are achievable due to differences already being represented by the physical properties. Retention predictions from properties are being reported with high precision and errors as low as MAE < 1 RI [47], but mainly include small structure ranges (below C9). Yan et al. however shows that even with property data incorporation the prediction with complex model approaches like a neural network might be difficult due to limited data availability (MAE > 20 RI) of cyclic components [29]. Without the availability of property data, models utilized descriptors mainly referring to two-dimensional structure quantification. Therefore, the configurational effect of steric sidechain interactions cannot be represented in the fit, neglecting partially significant differences of stereo sensitive behavior. Without 3D descriptors some constitutional isomers might even be equally resolved, when vicinity of structural differences is not properly quantified. From the three listed models without property inclusion, only the Yin et al. model incorporated stereo-specific description [44]. The developed numerical code however only differentiates if *cis* or *trans* is present and disregards more complex stereo diversity, which sometimes led to equal RI predictions of stereo isomers. From the models that strictly relied on structure quantification, the best absolute results were achieved by Duvenbeck et al., where comparable absolute errors are reported to the herein described performance [42]. Their study however was only trained on a small selection of light structures and relative RI differences were therefore noticeably higher. The exclusion of a hold-out test might additionally present the metrics in a more positive light, as only trained structures are predicted. For the present work the model of Geer et al. is possibly the primary benchmark to compare to, as it incorporated mono-cycloalkane datasets within their comprehensive database over all chemical functionalities of the NIST-RI database [41]. Reported absolute metrics results from their hold-out test are comparable to the high carbon models results, although the higher difference between MAE and MedAE indicate a good general fit but stronger influence of outliers. The C13 model outperforms the results of Geer et al. on the given metrics. Assessing the quality of the model from Geer et al. on cycloalkanes structures only is complicated due to the unknown diversity of chemical functionalities present in the used test data. By reproducing the procedure published by Geer at al., predictions for the herein used cycloalkane datasets were generated and used for the calculated metrics column in Tab. 6. The indicated performance may be influenced by the neural network's use of an extensive training dataset and the unknown nature of the excluded compounds in the hold-out test set. This raises concerns about the generalization, as their model may perform well on known cycloalkanes but its ability to predict unseen structures remain uncertain. The results in Tab. 2 and Tab. 3 (Chapter 3) already demonstrated that, across all modeling approaches, training set metrics typically outperformed those obtained from the validation set. This was especially noticeable for more complex non-linear architectures. While this may partially explain the improved predictions for higher carbon ranges with limited data availability, the overall assessment of the model by Geer et al. remains challenging due to the possible biases. Despite the questions regarding the neural network's generalizability, the C13 model surpassed the performance of Geer et al. on the assessed metrics. The C20 model's results from validation imply less precision on heavy molecules, but the hold-out test performance from C20 lies well within the neural-networks range. Due to Geer et al. not incorporating 3-dimensional descriptors, stereoisomeric effects were completely excluded in their evaluation.

In conclusion, while the herein proposed models do not match the precision of those using experimental properties as input, they offer a practical solution in the absence of comprehensive physical data. By enabling evaluation across a greater carbon range and incorporation of stereochemical sensitivity, improvements were achieved in comparison to literature models that rely on molecular descriptors. Although predictive accuracy decreases for higher carbon numbers, the models still provide the ability to assess heavier structures within carbon ranges relevant to jet fuel analytics.

Additional work will be conducted to evaluate the impact of the herein presented model on fuel analysis through its application to GCxGC-based subgrouping approaches, analogous to previous studies on isoalkanes [22]. Further improvements may be achieved through the inclusion of synthetically generated data points based on homologous series, thereby extending coverage of the chemical space (e. g. use cyclooctane, methylcyclooctane and ethylcyclooctane to extrapolate towards propylcyclooctane). In addition, previous studies have explored the development of QSRR models for comprehensive two-dimensional GC [86,87], which may provide improved capabilities for solving complex separation problems.

4.2. Remaining uncertainties regarding model performance and applicability domain

As illustrated in Fig. 5, the majority of datapoints from the NIST database are well represented by the combination of the reported models (only molecules above C13 were predicted by the C20 model). The highlighted region with an absolute RI difference ($|\Delta RI|$) < 20 between prediction and literature value contains 270/309 data points. The Williams plots for each models individual applicability domain are presented in the supplement (see Fig. S2). By the combination of the applicability domain 6/309 molecules were detected as response outliers (i.e. $> \pm 3$ standard residuals), while 11/309 isomers were found above the respective leverage thresholds (h^*). The highest uncertainties of prediction were found for the C14 ($\Delta RI = 74.82$) and C16 ($\Delta RI =$

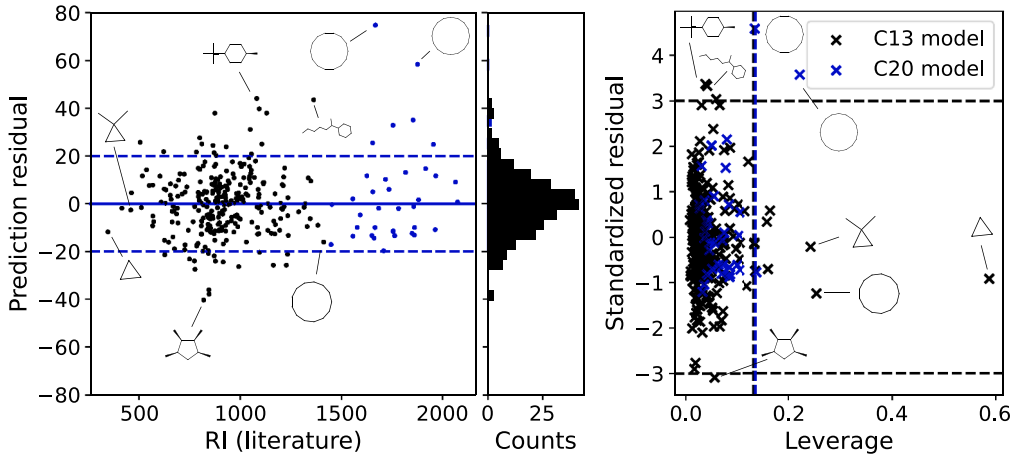

Fig. 5. Illustration of prediction residuals against retention index of the herein created models (left). The Williams plot for the applicability domain (right). Leverage thresholds individually plotted for C13 (h* = 0.132) and C20 (h* = 0.136). RI predictions are combined, taken first from the C13 model (black) and extended towards the high carbon range by the C20 model (blue). Individual structures are highlighted.

58.38) rings without sidechains. Since all other "full rings" from C3-C13 have a mean $|\Delta RI| = 9.23$ without the two outliers (mean $|\Delta RI| = 18.06$ with outliers) these two deviations might have a higher data uncertainty within the database. Due to the C20 model showing general higher errors this effect might also be partially related to the C20 fit to such individual structures with no sidechains (total MAE = 28.33 RI; without outliers MAE C3-C12: 15.93 RI; 21.38 RI with C13). As it can be seen from the Williams plot (Fig. 5), several ring structures without any sidechains were found outside of the leverage threshold our residual boundaries. Larger ring structures were generally underrepresented as 34/36 compounds above C13 were 5- or 6-membered rings, making them difficult to calibrate sufficiently. In total, only 36/309 of the represented isomers possessed ring sizes greater then six members, including the 13 unsubstituted cycloalkanes without side chains. A similar problem can be discussed in connection to complex sidechains like tert-butyl. Only 24/309 molecules included internally branched sidechains, while 48/309 had the ring not connected to the end of a sidechain (like 1-propylbutyl or 1-ethylmethyl which is a isopropyl rest). The majority of data points was given with only methyl or other straight chain alkyl (224/309) ring substituents, possibly explaining better model generalization on lower complexity structures. Despite the fact that the *cis/trans* 1-methyl-4-tertbutylcyclohexanes were both predicted around 40 RI off, the relative elution effect of their stereo difference was well represented (Literature: *cis* 1093 RI, *trans* 1081 RI; Prediction: *cis* 1053.3 RI, *trans* 1036.9 RI). Due to both structures beeing only reported once in literature, the literature RI uncertainty is unknown [88]. For other compounds literature RI were reported with high given errors from NIST summarization. The given RI in the database indicated that 1,2,3,4-tetramethylcyclopentane (ααααα; all *cis*) elutes as earliest stereo isomer (RI ± 23) [48]. Other sources hint that this isomer actually has a noticably higher boiling point or consequently RI than the *cis-trans* mixed isomers [89,90]. The other found 1,2,3,4-tetramethylcyclopentane isomers also have a mean $|\Delta RI| = 6.43$ between literature and prediction values. The here shown 1,2,3,4-tetramethylcyclopentane (ααααα) outlier therefore might have a strong dependency on the reported literature source. Further consideration regarding the quality of literature values can depend of chromatographic conditions and individual sources. RI deviations have previously been linked to variations in temperature programming [91,92], which may not be consistent across the dataset compiled by NIST. Additionally, the curation of literature data by NIST can contain some general reliability questions as reported by Matyushin et al. [93]. Such RI uncertainties are therefore automatically included and can influence potential model predictions. Due to general limitation in availability of data or reference standards it still provided the most comprehensive base to gather representative data for modeling approaches.

Besides the given uncertainties from combined literature sources, additional uncertainties can be introduced by the descriptor generation. CORINA based 3D coordinates from the NCI-Webtool were reproducible with the same input SMILES keys. The reported structure however is not certainly the most optimized, depending on process to find the energetic minima. When the enantiomeric smiles key was given, the optimized coordinates changed. Sometimes they were just inverted, resulting in identical descriptors. Other times, the descriptors were affected by the differences in structure optimization. This could lead to enantiomeric retention differences, which are not physically expected in an achiral column setup. For the respective stereo sensitive structures herein, that could have enantiomeric partners (90/309), a mean absolute difference of 1.09 RI was calculated. Despite in theory beeing physically identical the seven included enantiomeric literature pairs showed noticable RI differences. With reported $|\Delta RI|$ between 7 – 39, the accuracy of prediction is difficult to assess without knowing the "correct value". With predictions of enantiomers having an neglactable deviation, one of the literature values was usually well characterized, while the second showed higher deviations. Due to the given uncertainty, they were not included in subsequent qualitative discussions.

At last, the applicability on unfamiliar data needs to be addressed. The available dataset represents only a limited subset of the theoretical possible isomeric diversity of mono-cycloalkanes (see Tab. 1). Since the total number of possible stereo isomers within jet range is too extensive, a general estimation of applicability was tested on the 24,820 isomers within the range of C3-C12. The AD included 62.5 % of all considered theoretical stereo isomers. 37.5 % of the theoretic structures until C12 exceeded the leverage threshold ($h^* = 0.132$) of the C13 model (see Fig. S3). The majority of data points that were found outside of the AD could be assigned to certain structural elements. Many of those isomers were found to have ring sizes of three or four carbons (7757/9316; 83.3 %) or be highly branched with five or more alkyl branches (7162/9316; 76.9 %). On the contrary, the AD still included 52.7% (8657/16,414) of small ring isomers and 33.7 % (3633/10,795) of high branched structures (5+ methyl groups). Leverage is not always an indicator for prediction errors, as illustrated by the Williams plot in Fig. 5. However, high leverage reveals less representation by the training set of the model and consequently higher prediction uncertainty due to extrapolation. Consequently, the model should be used with caution when applied on isomers that are structurally distant from the validated descriptor space.

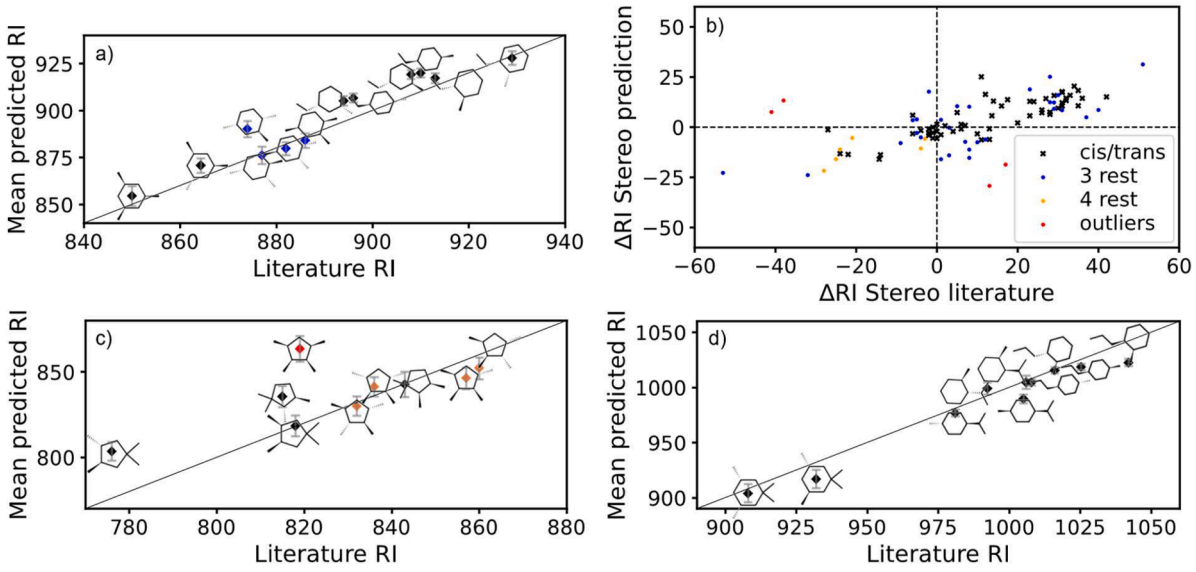

Fig. 6. Illustration of mean model performance from bootstrapping plus corresponding 95% confidence intervals for stereo isomeric separation examples. Only structures shown where multiple stereo isomers were available. a) C9-cyclohexanes; 1,2,4-trimethylcyclohexanes highlighted (blue). b) diastereomer pair RI differences in literature and prediction; different stereo cases and 1,2,3,4-tetramethylcyclopentane (αααα) outlier highlighted (red). c) C9-cyclopentanes with four methyl groups; 1,2,3,4-tetramethylcyclopentane (outlier, red) highlighted (orange). d) C10-cyclohexanes with *cis*/trans isomerism. Enantiomer data excluded due to uncertainties in literature.

4.3. Assessment of the model's stereo sensitivity

The majority of available data for relative retention differences of stereo-isomers was present for C9 and C10 mono-cycloalkanes (105/167, see also distribution in Fig. 2). Selected examples are illustrated in Fig. 6.

For each diastereomer the differences for literature and prediction RIs were calculated with every possible diastereomeric partner. This usually resulted only in data pairs for *cis/trans* cases but created a matrix of stereo pairs for more complex structures (3+ rests). The relative retention order was generally well recreated by the models. The absolute retention differences between included stereo data pairs had higher uncertainties. Predicted RI differences usually were smaller than the reported stereo effect in the literature (see axes Fig. 6b). Still, in 73/97 of the available isomeric data pairs the correct relative retention order was predicted. If only data pairs with RI differences above the average literature uncertainty ($|\Delta RI| = 3.7$) are considered the correct relative order fits in 62/80 cases. This reduced set will be further discussed in the following paragraph. The bootstrap-derived prediction intervals (see Fig. 6 and supplementary information) revealed that the majority of predicted RI differences of stereoisomer pairs statistically exceeded the half-width of the model's 95 % confidence intervals (CI_{hw}) for the respective isomers (59/80). This indicates the model's capability of assessing and reflecting the impact from stereochemistry on retention indices. Especially examples with only two diastereomeric configurations showed well representation by the model and usually reflected the literature derived elution order. From the final model 40/47 assignments were correct, with 34/40 RI differences above the CI_{hw}. Two diastereomeric isomers are usually possible, when two sidechains are not connected to the same ring atom. Additionally, only two structures might be possible with high molecular symmetry, like for 1,3,5-trimethylcyclohexane. There the only valid options are ααα- & ααβ- (*cis/trans*). All isomeric pairs shown in Fig. 6d), like 1,1,3,5-tetramethylcyclohexane or 1-methyl-4-propylcyclohexane only have two viable options (*cis/trans*). In contrast, stereo isomeric groups with higher diversity showed more uncertainties but still had good success rates (22/33 total; 18/22 of those above CI_{hw}). One example illustrated in Fig. 6a) is 1,2,4-trimethylcyclohexane. It can be observed that one structure deviated from the trend (βαβ−1,2,4-trimethylcyclohexane), whereas the predicted mean RI of 3/4 of the stereo isomers still follow the trend reported in the literature. This already resulted in 3/18 incorrect data pairs in the overall evaluation of mean predicted RI due to a single outlier. For 1,2,3,4-tetramethylcyclopentane (Fig. 6c), only five of eight unique RI values could be evaluated. With exception of the αααα-isomer the relative retention order was again well reproduced by the model. The uncertainty of this specific literature value was already discussed above. This however created four added wrong diastereomeric data pairs due to one outlier.

In general, the model is capable of predicting stereo sensitive retention behavior. The replication of relative retention order for diastereomeric pairs is accurate for around 78% (62/80) of the evaluable examples. Due to some diastereomeric pairs eluting closer to one another the confidence intervals of the individual isomers could overlapped occasionally. For 83% (52/62) of the correctly assigned stereo isomeric pair predictions both the elution order was correct and ΔRI exceeded the CI_{hw}, making the result also statistically reliable. In the remaining 10 examples, the general trend of mean prediction indicated the correct trend of elution order but size of confidence intervals included the mean of the isomeric partner (e. g. 1,2,4-trimethylcyclohexanes in Fig. 6a). The overall diversity of stereo sensitive separation cases though is limited and mainly includes *cis/trans* compounds. This makes the overall evaluation and training of complex stereoisomeric problems difficult and highly influenced by individual data quality. Facing the isomeric complexity and limited refence data, the models still reflected well, how significant steric differences can affect the retention range and in which order stereo sensitive structures are eluted. To date, it represents the most accurate tool available for determining RI values for complex cycloalkane isomerism in the carbon range relevant to jet fuels,

5. Conclusion

The impact of diastereomeric structure differences is known to be significant on gas chromatographic retention of mono-cycloalkanes. Available models in literature have commonly only considered cycloalkanes of C9 and below while neglecting steric effects, using physical properties as features or describing steric diversity incomplete. Herein retention index modelling was shown for mono-cycloalkanes in the

carbon range of jet fuels (i.e. C20) with inclusion of stereo isomeric sensitivity. Due to skewed data availability in the literature, one model was built from data until C13 with higher precision and one until C20 with slightly lower performance. Together the generation and evaluation of a comprehensive and reliable method to predict retention of jet-fuel relevant and stereo-sensitive mono-cyclic components was presented. The models are relying on a combination of multiple 2- and 3-dimensional descriptors. Although uncertainties arising from structure optimization procedures and data quality have to be considered, the given performance metrics showed improved value against varying literature reports. Especially compounds within the carbon range until C13 were shown to be reliably predicted. Validation errors from internal cross validation showed mean absolute errors of MAE (C13): 11.01 ± 1.32 RI and MAE (C20): 13.68 ± 1.67 RI over 50 iterations, while hold-out tests were MAE (C13): 9.31 RI and MAE (C20): 11.52 RI. Bootstrapped-derived 95% confidence intervals reflected the combined model's stereoisomeric sensitivity, with 65 % of the diastereomeric pairs predicted in correct elution order and statistically reliable (mean RI 78% correct). The classification of *cis/trans* cases was matched best, with 72 % of correct assignments (mean RI 85 % correct). Despite remaining limitations, the results indicate that the proposed model achieves the highest accuracy to date predicting RI values of complex cycloalkane isomers in the carbon range relevant to jet fuels.

Overall, both models produced promising results for the comprehension of cycloalkane retention effects. They will be beneficial for further understanding of fuel differences, fuel property predictions and the comparison of jet fuel and sustainable aviation fuel compositions.

CRediT authorship contribution statement

Hannes Lüdtke: Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Florian Pütz:** Writing – review & editing, Software, Methodology, Formal analysis, Data curation. **Samuel Grams:** Writing – review & editing, Formal analysis. **Thomas Gröger:** Writing – review & editing, Supervision, Project administration. **Barbara Giocastro:** Writing – review & editing. **Uwe Bauder:** Writing – review & editing, Supervision, Funding acquisition, Data curation. **Markus Köhler:** Writing – review & editing, Supervision, Resources, Project administration, Funding acquisition. **Andreas Huber:** Writing – review & editing, Supervision, Resources, Funding acquisition. **Patrick Oßwald:** Writing – review & editing, Supervision, Resources, Funding acquisition, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

The authors gratefully acknowledge funding by the German Federal Ministry for Digital and Transport (BMDV) within the SAFari project (FKZ: 16RK14009E). Furthermore, the support by DLR's Scientific Program Administrations Energy, Aeronautics, Space and Transport in the project framework of DLR NF Ignite is gratefully acknowledged.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.chroma.2026.467218.

Data availability

The used isomer list is provided in the supplement. Reference data is licensed and therefore not permitted to be shared publicly.

References

[1] R. Teoh, U. Schumann, C. Voigt, T. Schripp, M. Shapiro, Z. Engberg, J. Molloy, G. Koudis, M.E.J. Stettler, Targeted use of sustainable aviation fuel to maximize climate benefits, Environ. Sci. Technol. 56 (2022) 17246–17255, https://doi.org/10.1021/acs.est.2c05781.

[2] R.S. Märkl, C. Voigt, D. Sauer, R.K. Dischl, S. Kaufmann, T. Harlaß, V. Hahn, A. Roiger, C. Weiß-Rehm, U. Burkhardt, U. Schumann, A. Marsing, M. Scheibe, A. Dörnbrack, C. Renard, M. Gauthier, P. Swann, P. Madden, D. Luff, R. Sallinen, T. Schripp, P.L. Clercq, Powering aircraft with 100 % sustainable aviation fuel reduces ice crystals in contrails, Atmos. Chem. Phys. 24 (2024) 3813–3837, https://doi.org/10.5194/acp-24-3813-2024.

[3] C. Voigt, J. Kleine, D. Sauer, R.H. Moore, T. Bräuer, P. Le Clercq, S. Kaufmann, M. Scheibe, T. Jurkat-Witschas, M. Aigner, U. Bauder, Y. Boose, S. Borrmann, E. Crosbie, G.S. Diskin, J. DiGangi, V. Hahn, C. Heckl, F. Huber, J.B. Nowak, M. Rapp, B. Rauch, C. Robinson, T. Schripp, M. Shook, E. Winstead, L. Ziemba, H. Schlager, B.E. Anderson, Cleaner burning aviation fuels can reduce contrail cloudiness, Commun. Earth. Environ. 2 (2021) 114, https://doi.org/10.1038/s43247-021-00174-y.

[4] European Parliament, Verordnung (EU) 2023/2405 des europäischen Parlaments und des Rates vom 18. Oktober 2023 zur gewährleistung gleicher wettbewerbsbedingungen für einen nachhaltigen luftverkehr (Initiative ReFuelEU Aviation). http://data.europa.eu/eli/reg/2023/2405/oj, 2023.

[5] ICAO through the ICAO-UNDP Project, financed by GEF, Sustainable Aviation Fuels Guide: version 2. https://www2023.icao.int/environmental-protection/GFAAF/pages/Other-Resources.aspx, 2018.

[6] European Comission: directorate-General for Mobility and Transport and Directorate-General for Maritime Affairs and Fisheries, State of the art on Alternative fuels Transport Systems in the European Union –2020 update. https://data.europa.eu/doi/10.2771/29117, 2020.

[7] J. Heyne, B. Rauch, P. Le Clercq, M. Colket, Sustainable aviation fuel prescreening tools and procedures, Fuel 290 (2021) 120004, https://doi.org/10.1016/j.fuel.2020.120004.

[8] Z. Yang, S. Kosir, R. Stachler, L. Shafer, C. Anderson, J.S. Heyne, A GC × GC tier α combustor operability prescreening method for sustainable aviation fuel candidates, Fuel 292 (2021) 120345, https://doi.org/10.1016/j.fuel.2021.120345.

[9] F. Pütz, H. Lüdtke, A.R. Hernandez, P. Oßwald, P. Le Clercq, U. Bauder, G. Eckel, A. Huber, Influence of branching degree-derived iso -alkane GC×GC subgroups on fuel property prediction, Energy Fuels. 39 (2025) 18001–18012, https://doi.org/10.1021/acs.energyfuels.5c02762.

[10] S. Kosir, A.M. Arts, T. Nicely, C.J. Nesbit, S. Mueller, A. Landsaw, M. Baker, L. Brown, J. Thompson, C. Anderson, L. Shafer, W.H. Steinecker, S. Zabarnick, Z. J. West, World Jet fuel Survey part 1: 2023-2024: CRC report No. AV-33-22. https://crcao.org/crc-project-no-av-33-22/, 2025.

[11] O.J. Hadaller, J.M. Johnson, World fuel Sampling Program: CRC Report No. 647. https://crcao.org/crc-report-no-647/, 2006.

[12] P. Oßwald, J. Zinsmeister, T. Kathrotia, M. Alves-Fortunato, V. Burger, R. van der Westhuizen, C. Viljoen, K. Lehto, R. Sallinen, K. Sandberg, M. Aigner, P. Le Clercq, M. Köhler, Combustion kinetics of alternative jet fuels, part-I: experimental flow reactor study, Fuel 302 (2021) 120735, https://doi.org/10.1016/j.fuel.2021.120735.

[13] C.J. Faulhaber, Elastomer O-ring seal swell measurements for sustainable aviation fuel material compatibility, Honors Theses 384 (2022). https://ecommons.udayton.edu/uhp_theses/384/.

[14] S. Kosir, J. Heyne, J. Graham, A machine learning framework for drop-in volume swell characteristics of sustainable aviation fuel, Fuel 274 (2020) 117832, https://doi.org/10.1016/j.fuel.2020.117832.

[15] Y. Liu, C.W. Wilson, Investigation into the impact of n-decane, Decalin, and Isoparaffinic solvent on elastomeric sealing materials, Adv. Mech. Eng. 4 (2012) 127430, https://doi.org/10.1155/2012/127430.

[16] Z. Yang, Z. Xu, M. Feng, J.R. Cort, R. Gieleciak, J. Heyne, B. Yang, Lignin-based jet fuel and its blending effect with conventional jet fuel, Fuel 321 (2022) 124040, https://doi.org/10.1016/j.fuel.2022.124040.

[17] X. Chen, K.A. Orton, C. Mukarakate, K. Gaston, G.M. Fioroni, R.L. McCormick, M. B. Griffin, K. Iisa, Cycloalkane-rich sustainable aviation fuel production via hydrotreating lignocellulosic biomass-derived catalytic fast pyrolysis oils, Sustain. Energy Fuels 8 (2024) 5504–5513, https://doi.org/10.1039/D4SE01151A.

[18] A. Landera, R.P. Bambha, N. Hao, S.P. Desai, C.M. Moore, A.D. Sutton, A. George, Building structure-property relationships of cycloalkanes in support of their use in sustainable aviation fuels, Front. Energy Res. 9 (2022) 771697, https://doi.org/10.3389/fenrg.2021.771697.

[19] R.C. Striebich, L.M. Shafer, R.K. Adams, Z.J. West, M.J. DeWitt, S. Zabarnick, Hydrocarbon group-type analysis of petroleum-derived and synthetic fuels using two-dimensional gas chromatography, Energy Fuels. 28 (2014) 5696–5706, https://doi.org/10.1021/ef500813x.

[20] J. Melder, J. Zinsmeister, T. Grein, S. Jürgens, M. Köhler, P. Oßwald, Comprehensive two-dimensional gas chromatography: a universal method for composition-based prediction of emission characteristics of complex fuels, Energy Fuels. 37 (2023) 4580–4595, https://doi.org/10.1021/acs.energyfuels.2c04270.

[21] M.K. Jennerwein, M. Eschner, T. Gröger, T. Wilharm, R. Zimmermann, Complete group-type quantification of petroleum middle distillates based on comprehensive two-dimensional gas chromatography time-of-flight mass spectrometry (GC×GC-

TOFMS) and visual basic scripting, Energy Fuels. 28 (2014) 5670–5681, https://doi.org/10.1021/ef501247h.
[22] H. Lüdtke, F. Pütz, S. Grams, U. Bauder, M. Köhler, A. Huber, P. Oßwald, GCxGC-based iso -alkane subgrouping for enhanced compositional analysis of sustainable aviation fuels, Energy Fuels. 39 (2025) 4434–4449, https://doi.org/10.1021/acs.energyfuels.4c05573.
[23] C. Hall, D.C. Bell, J. Feldhausen, B. Rauch, J. Heyne, Quantifying isomeric effects: a key factor in aviation fuel assessment and design, Fuel 357 (2024) 129912, https://doi.org/10.1016/j.fuel.2023.129912.
[24] H.R. Bizzo, N.S. Brilhante, Y. Nolvachai, P.J. Marriott, Use and abuse of retention indices in gas chromatography, J. Chromatogr. A 1708 (2023) 464376, https://doi.org/10.1016/j.chroma.2023.464376.
[25] R. Todeschini, V. Consonni (Eds.), Molecular Descriptors for Chemoinformatics, Wiley-VCH, Weinheim, 2009.
[26] R. Kaliszan, QSRR: quantitative structure-(chromatographic) retention relationships, Chem. Rev. 107 (2007) 3212–3246, https://doi.org/10.1021/cr068412z.
[27] A.R. Katritzky, M. Kuanar, S. Slavov, C.D. Hall, M. Karelson, I. Kahn, D.A. Dobchev, Quantitative correlation of physical and chemical properties with chemical structure: utility for prediction, Chem. Rev. 110 (2010) 5714–5789, https://doi.org/10.1021/cr900238d.
[28] C. Qu, B.I. Schneider, A.J. Kearsley, W. Keyrouz, T.C. Allison, Predicting Kováts retention indices using graph neural networks, J. Chromatogr. A 1646 (2021) 462100, https://doi.org/10.1016/j.chroma.2021.462100.
[29] A. Yan, R. Zhang, M. Liu, Z. Hu, M.A. Hooper, Z. Zhao, Large artificial neural networks applied to the prediction of retention indices of acyclic and cyclic alkanes, alkenes, alcohols, esters, ketones and ethers, Comput. Chem. 22 (1998) 405–412, https://doi.org/10.1016/S0097-8485(98)00001-1.
[30] C.E. Üstün, R.S.M. de Da Freitas, E.C. Okafor, M. Shahbakhti, X. Jiang, A. Paykani, Machine learning applications for predicting fuel ignition and flame properties: current status and future perspectives, Energy Fuels. (2025), https://doi.org/10.1021/acs.energyfuels.5c02343.
[31] X. Yu, Y. Wang, L. Wu, G. Chen, L. Wang, H. Qin, Comparison of support vector regression and extreme gradient boosting for decomposition-based data-driven 10-day streamflow forecasting, J. Hydrol. 582 (2020) 124293, https://doi.org/10.1016/j.jhydrol.2019.124293.
[32] T.R. Noviandy, A. Maulana, N. Sasmita, R. Suhendra, I. Irvanizam, M. Muslem, G. Idroes, M. Yusuf, H. Sofyan, T. Abidin, R. Idroes, The prediction of Kovats retention indices of essential oils at gas chromatography using genetic algorithm-mulitple linear regression and support vector regression, J. Eng. Sci. Technol. Vol. 17 (2022) 306–326.
[33] S. Ahmadi, S. Lotfi, H. Hamzehali, P. Kumar, A simple and reliable QSPR model for prediction of chromatography retention indices of volatile organic compounds in peppers, RSC. Adv. 14 (2024) 3186–3201, https://doi.org/10.1039/d3ra07960k.
[34] Z. Garkani-Nejad, M. Karlovits, W. Demuth, T. Stimpfl, W. Vycudilik, M. Jalali-Heravi, K. Varmuza, Prediction of gas chromatographic retention indices of a diverse set of toxicologically relevant compounds, J. Chromatogr. A 1028 (2004) 287–295, https://doi.org/10.1016/j.chroma.2003.12.003.
[35] A.R. Katritzky, K. Chen, U. Maran, D.A. Carlson, QSPR correlation and predictions of GC retention indexes for methyl-branched hydrocarbons produced by insects, Anal. Chem. 72 (2000) 101–109, https://doi.org/10.1021/ac990800w.
[36] T.M. McGinitie, J.J. Harynuk, Prediction of retention times in comprehensive two-dimensional gas chromatography using thermodynamic models, J. Chromatogr. A 1255 (2012) 184–189, https://doi.org/10.1016/j.chroma.2012.02.023.
[37] H. Noorizadeh, A. Farmany, QSRR models to predict retention indices of cyclic compounds of essential oils, Chromatographia 72 (2010) 563–569, https://doi.org/10.1365/s10337-010-1660-4.
[38] N. Dimov, General model for precalculation of the retention indices of isoalkanes separated by gas or liquid chromatography, Anal. Chim. Acta 201 (1987) 217–224, https://doi.org/10.1016/S0003-2670(00)85339-8.
[39] B. Skrbić, A. Onjia, Prediction of the Lee retention indices of polycyclic aromatic hydrocarbons by artificial neural network, J. Chromatogr. A 1108 (2006) 279–284, https://doi.org/10.1016/j.chroma.2006.01.080.
[40] J.M. Sutter, T.A. Peterson, P.C. Jurs, Prediction of gas chromatographic retention indices of alkylbenzenes, Anal. Chim. Acta 342 (1997) 113–122, https://doi.org/10.1016/S0003-2670(96)00578-8.
[41] L.Y. Geer, S.E. Stein, W.G. Mallard, D.J. Slotta, AIRI: predicting retention indices and their uncertainties using artificial intelligence, J. Chem. Inf. Model. 64 (2024) 690–696, https://doi.org/10.1021/acs.jcim.3c01758.
[42] C. Duvenbeck, P. Zinn, List operations on chemical graphs. 3. Development of vertex and edge models for fitting retention index data, J. Chem. Inf. Comput. Sci. 33 (1993) 211–219, https://doi.org/10.1021/ci00012a005.
[43] J. Bermejo, M.D. Guillén, Prediction of gas chromatographic retention indices of linear, branched, and cyclic alkanes from their physicochemical properties, J. High Resol. Chromatogr. 7 (1984) 191–195, https://doi.org/10.1002/jhrc.1240070404.
[44] C. Yin, W. Liu, Z. Li, Z. Pan, T. Lin, M. Zhang, Chemometrics to chemical modeling: structural coding in hydrocarbons and retention indices of gas chromatography, J. Sep. Sci. 24 (2001) 213–220, https://doi.org/10.1002/1615-9314(20010301)24:3<213:AID-SSC213>3.0.CO;2-4.
[45] S.E. Stein, V.I. Babushok, R.L. Brown, P.J. Linstrom, Estimation of Kováts retention indices using group contributions, J. Chem. Inf. Model. 47 (2007) 975–980, https://doi.org/10.1021/ci600548y.
[46] O. Farkas, K. Héberger, I.G. Zenkevich, Quantitative structure–retention relationships XIV, Chemom. Intell. Lab. Syst. 72 (2004) 173–184, https://doi.org/10.1016/j.chemolab.2004.01.012.
[47] N. Dimov, D. Papazova, Calculation of the retention indices of C5-C9 cycloalkanes on squalane, J. Chromatogr. A 148 (1978) 11–15, https://doi.org/10.1016/S0021-9673(00)99316-6.
[48] S.E. Stein, NIST/EPA/NIH Mass Spectral Library-PC Version, 2023 Release, National Institute of Standards and Technology NIST, 2023.
[49] V.K. Ahluwalia, Stereochemistry of Organic Compounds, Springer International Publishing, Cham, Switzerland, 2022.
[50] C. Wolf, Dynamic Stereochemistry of Chiral Compounds: Principles and Applications, The Royal Society of Chemistry, Cambridge, 2007.
[51] A.A. Elbashir, H.Y. Aboul-Enein, Multidimensional gas chromatography for chiral analysis, Crit. Rev. Anal. Chem. 48 (2018) 416–427, https://doi.org/10.1080/10408347.2018.1444465.
[52] R. Gugisch, A. Kerber, A. Kohnert, R. Laue, M. Meringer, C. Rücker, A. Wassermann, MOLGEN 5.0, A molecular structure generator, in: G. Restrepo, J. L. Villaveces, S.C. Basak (Eds.), Advances in Mathematical Chemistry and Applications, Advances in Mathematical Chemistry and Applications, 1, BENTHAM SCIENCE PUBLISHERS, Oak Park, 2014, pp. 113–138.
[53] Y. Miyake, C. Boned, A. Baylaucq, D. Bessières, C.K. Zéberg-Mikkelsen, G. Galliéro, H. Ushiki, Stereoisomeric effects on dynamic viscosity versus pressure and temperature for the system cis-+trans-decalin, Chem. Phys. 336 (2007) 171–182, https://doi.org/10.1016/j.chemphys.2007.06.015.
[54] S.R. Rieder, M.P. Oliveira, S. Riniker, P.H. Hünenberger, Development of an open-source software for isomer enumeration, J. Cheminform. 15 (2023) 10, https://doi.org/10.1186/s13321-022-00677-6.
[55] G. Schomburg, Struktur und retentionsverhalten von offenkettigen und cyclischen kohlenwasserstoffen und deren einfacher substitutionsprodukte, Anal. Chim. Acta 38 (1967) 45–64, https://doi.org/10.1016/S0003-2670(01)80560-2.
[56] N.T. Sultanov, L.G. Arustamova, Determination of the boiling points of C10 isoalkanes in an n-decane isomerizate from gas chromatographic retention indices, J. Chromatogr. A 115 (1975) 553–558, https://doi.org/10.1016/S0021-9673(01)98959-9.
[57] M. de Laet, B. Tilquin, Identification D'isomères du Dodécane mesures des indices de rétention a plusieurs Températures, Bull. Soc. Chim. Belg. 100 (1991) 345–355, https://doi.org/10.1002/bscb.19911000408.
[58] F.R. Jensen, D.S. Noyce, C.H. Sederholm, A.J. Berlin, The rate of the chair-chair interconversion of cyclohexane, J. Am. Chem. Soc. 84 (1962) 386–389, https://doi.org/10.1021/ja00862a015.
[59] RDKit: open-source cheminformatics, Release 2023.9.2. https://www.rdkit.org, 2023.
[60] C.W. Yap, PaDEL-descriptor: an open source software to calculate molecular descriptors and fingerprints, J. Comput. Chem. 32 (2011) 1466–1474, https://doi.org/10.1002/jcc.21707.
[61] N.M. O'Boyle, C. Morley, G.R. Hutchison, Pybel: a Python wrapper for the OpenBabel cheminformatics toolkit, Chem. Cent. J. 2 (2008) 5, https://doi.org/10.1186/1752-153X-2-5.
[62] J. Sadowski, C.H. Schwab, J. Gasteiger, CORINA: 3D Structure Generator. Version 3.0 Program Description, Erlangen, Germany, 2003.
[63] W.-D. Ihlenfeldt, J.H. Voigt, B. Bienfait, F. Oellien, M.C. Nicklaus, Enhanced CACTVS browser of the Open NCI Database, J. Chem. Inf. Comput. Sci. 42 (2002) 46–57, https://doi.org/10.1021/ci010056s.
[64] NCI CAAD Group, Online SMILES translator and structure file generator. https://cactus.nci.nih.gov/translate/(accessed 5 June 2025).
[65] J. Gasteiger, C. Rudolph, J. Sadowski, Automatic generation of 3D-atomic coordinates for organic molecules, Tetrahedron Comput. Methodol. 3 (1990) 537–547, https://doi.org/10.1016/0898-5529(90)90156-3.
[66] W.D. Ihlenfeldt, Y. Takahashi, H. Abe, S. Sasaki, Computation and management of chemical properties in CACTVS: an extensible networked approach toward modularity and compatibility, J. Chem. Inf. Comput. Sci. 34 (1994) 109–116, https://doi.org/10.1021/ci00017a013.
[67] D. Weininger, SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules, J. Chem. Inf. Comput. Sci. 28 (1988) 31–36, https://doi.org/10.1021/ci00057a005.
[68] H. Moriwaki, Y.-S. Tian, N. Kawashita, T. Takagi, Mordred: a molecular descriptor calculator, J. Cheminform. 10 (2018) 4, https://doi.org/10.1186/s13321-018-0258-y.
[69] C. Spearman, The proof and measurement of association between two things, Am. J. Psychol. 15 (1904) 72, https://doi.org/10.2307/1412159.
[70] K. Pearson VII, Mathematical contributions to the theory of evolution.—III. Regression, heredity, and panmixia, Phil. Trans. R. Soc. Lond. A 187 (1896) 253–318, https://doi.org/10.1098/rsta.1896.0007.
[71] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, A. Müller, J. Nothman, G. Louppe, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, É. Duchesnay, Scikit-learn: machine learning in Python, J. Mach. Learn. Res. 12 (2012) 2825–2830, https://doi.org/10.48550/arXiv.1201.0490.
[72] Scikit-learn 1.4.2: machine learning in Python. https://scikit-learn.org/1.4/index.html, 2024.
[73] D.T. Stanton, P.C. Jurs, Development and use of charged partial surface area structural descriptors in computer-assisted quantitative structure-property relationship studies, Anal. Chem. 62 (1990) 2323–2329, https://doi.org/10.1021/ac00220a013.
[74] S. Liu, C. Cao, Z. Li, Approach to estimation and prediction for normal boiling point (NBP) of Alkanes based on a novel molecular distance-edge (MDE) vector, λ, J. Chem. Inf. Comput. Sci. 38 (1998) 387–394, https://doi.org/10.1021/ci970109z.

[75] T. Chen, C. Guestrin, XGBoost: a scalable tree boosting system, in: B. Krishnapuram (Ed.), Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, New York, NY, 2016, pp. 785–794. doi:10.1145/2939672.2939785.
[76] C.H. Yu, Resampling methods: concepts, applications, and justification (2003). https://doi.org/10.7275/9cms-my97.
[77] B. Efron, R. Tibshirani, Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy, Stat. Sci. 1 (1986) 54–75, https://doi.org/10.1214/ss/1177013815.
[78] L. Breiman, out-of-bag estimation, Technical report, University of California, Statistics Department, Berkely, CA, 1996.
[79] S. Kar, K. Roy, J. Leszczynski, Applicability domain: a step toward confident predictions and decidability for QSAR modeling, Methods Mol. Biol. 1800 (2018) 141–169, https://doi.org/10.1007/978-1-4939-7899-1_6.
[80] K. Roy, H. Kabir, QSPR with extended topochemical atom (ETA) indices: exploring effects of hydrophobicity, branching and electronic parameters on logCMC values of anionic surfactants, Chem. Eng. Sci. 87 (2013) 141–151, https://doi.org/10.1016/j.ces.2012.10.002.
[81] M. Randic, On molecular identification numbers, J. Chem. Inf. Comput. Sci. 24 (1984) 164–175, https://doi.org/10.1021/ci00043a009.
[82] E. Kováts, Gas-chromatographische charakterisierung organischer verbindungen. Teil 1: retentionsindices aliphatischer Halogenide, alkohole, aldehyde und ketone, Helv. Chim. Acta 41 (1958) 1915–1932, https://doi.org/10.1002/hlca.19580410703.
[83] J.A. Platts, D. Butina, M.H. Abraham, A. Hersey, Estimation of molecular linear free energy relation descriptors using a group contribution approach, J. Chem. Inf. Comput. Sci. 39 (1999) 835–845, https://doi.org/10.1021/ci980339t.
[84] Y. Yang, H. Chen, I. Nilsson, S. Muresan, O. Engkvist, Investigation of the relationship between topology and selectivity for druglike molecules, J. Med. Chem. 53 (2010) 7709–7714, https://doi.org/10.1021/jm1008456.
[85] G.W. Bemis, M.A. Murcko, The properties of known drugs. 1. Molecular frameworks, J. Med. Chem. 39 (1996) 2887–2893, https://doi.org/10.1021/jm9602928.
[86] R. Gieleciak, D. Hager, N.E. Heshka, Application of a quantitative structure retention relationship approach for the prediction of the two-dimensional gas chromatography retention times of polycyclic aromatic sulfur heterocycle compounds, J. Chromatogr. A 1437 (2016) 191–202, https://doi.org/10.1016/j.chroma.2016.02.006.
[87] Y. Ren, H. Liu, X. Yao, M. Liu, An accurate QSRR model for the prediction of the GCxGC-TOFMS retention time of polychlorinated biphenyl (PCB) congeners, Anal. Bioanal. Chem. 388 (2007) 165–172, https://doi.org/10.1007/s00216-007-1188-0.
[88] C. Paris, P. Alexandre, Stereochemical investigation of cyclohexane and terpene compounds by gas chromatography, J. Chromatogr. Sci. 10 (1972) 402–411, https://doi.org/10.1093/chromsci/10.6.402.
[89] V. Zakharenko, I. Delone, O. Morozova, A. Petrov, Thermodynamic stability of C9 cyclopentanes at 600°K, Pet. Chem. 9 (1969) 104–108, https://doi.org/10.1016/0031-6458(69)90003-3.
[90] H.N. Miller, K.W. Greenlee, Synthesis of 1,2,3,4-tetramethylcyclopentanes. I 1,2, J. Org. Chem. 26 (1961) 3734–3739, https://doi.org/10.1021/jo01068a028.
[91] P.G. Boswell, P.W. Carr, J.D. Cohen, A.D. Hegeman, Easy and accurate calculation of programmed temperature gas chromatographic retention times by back-calculation of temperature and hold-up time profiles, J. Chromatogr. A 1263 (2012) 179–188, https://doi.org/10.1016/j.chroma.2012.09.048.
[92] D.D. Matyushin, A.Y. Sholokhova, Large-scale statistical study of the dependence of retention index on heating rate in temperature-programmed gas chromatography, J. Chromatogr. A 1732 (2024) 465223, https://doi.org/10.1016/j.chroma.2024.465223.
[93] D.D. Matyushin, A.E. Karnaeva, A.Y. Sholokhova, Critical evaluation of the NIST retention index database reliability with specific examples, Anal. Bioanal. Chem. 416 (2024) 6181–6186, https://doi.org/10.1007/s00216-024-05562-9.