Machine Learning for Structure-Agnostic Chemical Analysis from Chromatographic Data

Adam Lahouar

Thesis submitted to the Faculty of the
Virginia Polytechnic Institute and State University
in partial fulfillment of the requirements for the degree of

Master of Science
in
Computer Science & Application

Hoda M. Eldardiry, Chair
Gabriel Isaacman-VanWertz
Pinar Yanardag Delul

December 17, 2025
Blacksburg, Virginia

Keywords: Gas Chromatography, Machine Learning, Compound Classification

Machine Learning for Structure-Agnostic Chemical Analysis from Chromatographic Data

Adam Lahouar

(ABSTRACT)

Environmental monitoring relies heavily on gas chromatography (GC) to measure airborne contaminants such as volatile organic compounds (VOCs), yet many detected compounds lack structural or spectral references, limiting identification, property estimation, and quantitative analysis. This thesis investigates how machine learning (ML) can extract chemically meaningful information directly from chromatographic data to overcome these limitations. First, ML models are developed to establish a bidirectional relationship between chromatographic retention behavior on orthogonal GC phases and key physicochemical properties (vapor pressure, Henry's law constant, and solubility). Using XGBoost regression models trained on the NIST retention index database, a structure-agnostic "Index-to-Property" model predicts physicochemical properties from paired retention indices, while a complementary "Property-to-Index" model predicts retention behavior from known properties, achieving predictive performance up to $R^2 = 0.98$. Second, this work demonstrates that compound identity and concentration can be inferred directly from chromatographic peak shape, bypassing manual peak integration. ML classification and regression models trained on peaks from ambient atmospheric samples achieve 89% identification accuracy and a mean absolute error of 0.085 ppbv in concentration prediction. Together, these results show that machine learning can address key identification and data reduction challenges in environmental GC, enabling faster, structure-independent interpretation of complex mixtures.

Machine Learning for Structure-Agnostic Chemical Analysis from Chromatographic Data

Adam Lahouar

(GENERAL AUDIENCE ABSTRACT)

Gas chromatography is an important method for monitoring air pollution, but many detected chemicals cannot be fully identified because reference information is missing or incomplete. This makes it difficult to understand what these compounds are, how they behave in the environment, and how much of them are present. This thesis explores how machine learning can help extract useful chemical information directly from chromatographic data. First, machine learning is used to relate chemical behavior in a gas chromatograph to important physical properties, allowing unknown compounds to be characterized without knowing their chemical structures. Second, machine learning is used to analyze the shape of chromatographic signals to identify compounds and estimate their concentrations automatically, reducing the need for time-consuming manual data processing. Overall, this research shows how machine learning can expand the capabilities of gas chromatography for environmental monitoring, improving both the speed and depth of chemical analysis over traditional methods.

Acknowledgments

Thank you to my family for supporting me and for always being there for me.

Thank you to my advisor, Dr. Hoda Eldardiry, for her guidance and direction, and to my committee members, Dr. Gabriel Isaacman-VanWertz and Dr. Pinar Yanardag, for their time and feedback.

This research has been supported by the Intelligence Advanced Research Projects Activity (IARPA) through the Department of the Interior/Interior Business Center (DOI/IBC) contract number 140D0424C0002.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Environmental monitoring is the systematic collection, analysis, and interpretation of data on environmental parameters and contaminants to track the quality of natural resources and assess the impact of human activities. The importance of environmental monitoring and modeling has grown in parallel with increasing concern over pollution and climate change [33]. Airborne contaminants such as volatile organic compounds (VOCs) pose significant environmental and health risks [10], and effective understanding of how these pollutants behave depends on accurate measurement and interpretation [33].

1.1 Chromatography

A common analytical technique used in environmental monitoring is chromatography, a family of techniques that separate a mixture into its constituent components. Among these, gas chromatography (GC) is especially relevant for atmospheric analysis. GC, often combined with mass spectrometry (GC-MS), enables the detection and quantification of trace VOCs with high sensitivity and specificity, making it a fundamental tool for air-quality monitoring [47].

Gas chromatography operates by introducing a sample of gaseous compounds into a long, narrow column coated with a stationary phase. A chemically inert carrier gas, known as the mobile phase, transports the analytes through the column. Separation occurs based on the

differential interaction of the compounds with the stationary phase: compounds that interact more strongly are retained longer, while those with weaker interactions continue through the chromatograph. At the end of the column, a detector measures the components as they exit. The detector signal plotted over time produces a chromatogram, which contains a peak corresponding to each compound in the sample. The retention time at which a peak appears, its shape, and its integrated area provide key information used in compound identification and quantification [47].

1.2 Challenges in Environmental Chromatographic Analysis

While GC-MS is the gold standard for atmospheric analysis, traditional workflows face distinct limitations regarding compound identification and data reduction.

Compound identification in environmental samples relies heavily on matching retention times or mass spectra to entries in reference libraries, yet these libraries often lack corresponding spectra for many detected compounds, making identification difficult or impossible [31, 52]. This limitation also affects environmental fate modeling, which requires compound-specific physicochemical properties (e.g., vapor pressure, solubility, Henry's law constant). Current methods for predicting these properties generally rely on known molecular structures [8, 18, 32, 41], and therefore cannot be applied to the majority of samples for which structural information is unavailable [31, 52].

The processing of the data itself is also a challenge. While modern instrumentation allows for high-frequency sampling ("fast GC") [34], the translation of raw signals into usable concentration data is often dependent upon subjective and time-consuming manual peak

integration [26]. This reliance on human interpretation represents a significant throughput bottleneck, limiting the speed and scale of comprehensive monitoring.

1.3 Machine Learning in Gas Chromatography

Machine learning (ML) offers a potential solution to both the identification and data reduction limitations inherent in gas chromatography. ML methods are designed to uncover patterns in complex, high-dimensional data, making them capable of exploiting latent information within chromatographic signals that traditional physics-based models overlook.

By applying ML to chromatography, we can address the identified gaps in two specific ways. First, ML enables structure-agnostic property prediction by learning the relationships between retention behavior on orthogonal phases and physicochemical properties. This capability allows for the prediction of vapor pressure, solubility, and Henry's law constant for unknown compounds directly from retention indices, bypassing the need for structural identification. Second, ML can use the raw shape of a chromatographic peak (e.g., fronting and tailing behavior, width, etc.) as a unique fingerprint, allowing for compound classification and concentration prediction from raw signal data, bypassing the need for manual peak integration or curve fitting.

1.4 Research Questions

This thesis investigates the central question: **How can machine learning be used to identify and characterize compounds from gas chromatographic data?**

To address this, we explore two specific sub-questions:

1. Can retention behavior on orthogonal gas chromatographic columns be used to predict fundamental physicochemical properties of compounds in the absence of structural information?

2. Does the shape of a chromatographic peak contain sufficient latent information to allow for compound classification and concentration prediction without the use of retention time or peak integration?

1.5 Contributions of this Thesis

To answer these questions, this thesis develops ML-based approaches that extract chemically meaningful information from two different aspects of GC data: retention behavior and peak shape.

In Chapter 2, we develop two complementary ML models that operate exclusively on retention indices measured on paired polar and nonpolar stationary phases. Because these phases provide orthogonal information about compound interactions, their combined retention indices capture meaningful chemical signatures. We introduce an "Index-to-Property" model that predicts physicochemical properties (vapor pressure, solubility, and Henry's law constant) directly from paired retention indices, without requiring structural information, and a "Property-to-Index" model that predicts retention indices on a given stationary phase from known physicochemical properties. Together, these models enable bidirectional inference between retention behavior and chemical properties, demonstrating that retention indices can serve as structure-agnostic descriptors for characterizing unknown compounds in environmental mixtures.

In Chapter 3, we investigate whether peak shape contains compound-specific information

that ML can exploit to automate data reduction. To support this investigation, we construct a large labeled dataset of isolated GC peaks from calibration chromatograms and use it to develop two ML models: a classification model that distinguishes among 11 compounds plus an Unknown class, using only normalized peak segments with all timing information removed, and a concentration prediction model that estimates compound concentration directly from peak shape. These models demonstrate that machine learning can effectively bypass manual processing and interpretation steps by exploiting the learnable information present within raw chromatographic signals.

Chapter 2

Compound Prediction from Gas Chromatographic Retention Data Using Machine Learning

2.1 Introduction

Gas chromatography-mass spectrometry (GC-MS) is a widely used technique for identifying compounds in atmospheric and other environmental samples. However, while a subset of compounds in these samples can be identified by their retention index and/or mass spectrometric data, a large fraction of compounds present in complex environmental samples cannot be identified. For example, in ambient atmospheric GC-MS datasets, 90% of detected peaks lack known structures or reference spectra in standard libraries [31, 52]. This limits not only identification of sources or potential constituents of concern, but also more broadly limits the ability to understand the physicochemical properties of these samples. Such properties, particularly volatility (described by vapor pressure, VP), air-water partitioning (described as Henry's Law constant, HLC), and solubility (sol) are critical for modeling partitioning between different environmental media and the environmental fate of a compound in natural and built systems (e.g., tendency toward deposition, chemical degradation, transport, etc.)[2, 3, 11, 20, 27, 40]. It is consequently useful not only to be able to identify compounds

in these mixtures but even to categorize or classify them by properties.

Predicting Properties from Structure Instead, most work to connect retention index and properties has relied on predictions of these parameters of interest based on molecular structure. Quantitative Structure-Activity Relationships (QSAR) models use descriptors based on molecular structures (molecular weight, functional groups, bonding, etc.) or graph-based machine learning (ML) to predict parameters for known compounds with high accuracy. These parameters may be physicochemical properties (e.g., [8, 18, 32, 41]) or may be chromatographic information such as retention index or mass spectrum [45, 48, 50, 51]. These methods have been able to predict properties with high accuracy, but by relying on structural information, these approaches cannot provide information about the 90% of ambient peaks that lack known structures [31, 52].

Relating Properties and Retention Index Recognizing the value of physicochemical property prediction even in the absence of identification, some efforts have sought to predict properties based on a combination of retention index and mass spectrometric data. Prior work has shown that retention indices (RIs) from single- or multidimensional GC (GC×GC) correlate directly with chemical properties, even without knowing structural information [17, 22]. These correlations have been used to classify complex chromatographic information into bins based on vapor pressure by incorporating a small amount of mass spectral information to filter data (e.g., [53]). However, Isaacman et al. and co-workers have previously shown that while a single retention index can provide some information about vapor pressure, accurate correlation requires some knowledge of the molecular structure or mass spectrum [21, 25]. Recent work [12] has expanded these ideas by incorporating retention index and mass spectral information into an ML model to improve predictions of vapor pressure and some molecular descriptions (e.g., oxygen-to-carbon ratio). Alternately, the use of orthogonal column phases

in GC×GC provides additional chemical information than estimation of vapor pressure [22]. Unfortunately, the accuracy of all of these approaches remains less than ideal, frequently predicting properties within a fairly narrow range of parameter space. Furthermore, none of these techniques have been generalized to other parameters like HLC or solubility, which are critical terms in understanding environmental fate in atmospheres and soils [13].

Chapter Objectives A structure-agnostic method to predict multiple properties directly from retention data would enhance the modeling of complex atmospheric mixtures containing mostly unknown compounds. While the work described above addresses this gap, these efforts (a) do not address key physicochemical parameters important in environmental fate modeling, and (b) rely on mass spectral information. Though mass spectrometers are widely used, they are not always available due to their cost and lack of portability; instead, "single-channel" detectors such as flame ionization or photoionization detectors that provide only a single value of total signal with time are often used in cheaper or ruggedized instruments (e.g., [36, 46]). This chapter therefore aims to fill these gaps by predicting a broader range of physicochemical properties using only retention indices. To achieve this aim, pairs of retention indices from nonpolar and polar column phases are used to provide orthogonal information that describes relatively unique combinations of physicochemical properties. We introduce two complementary models that interrelate pairs of retention indices and combinations of multiple physicochemical properties: the "Index-to-Property" model predicts physicochemical properties (VP, HLC, sol) from paired retention indices, requiring no structural information. The "Property-to-Index" model predicts retention indices from combinations of physicochemical properties for a given phase. Together, these models enable bidirectional inference between retention behavior and chemical properties.

2.2 Materials and Methods

The primary objective for the Index-to-Property model is to develop a correlation mapping between pairs of retention indices, with each corresponding to a different orthogonal stationary phase, and the physicochemical properties of interest. Conceptually, the paired indices act as a fingerprint for the compound, enabling its location in property space through the prediction of its vapor pressure, solubility, and Henry's law constant. To model these relationships in both the forward and reverse directions, a large dataset containing relevant chromatographic information is needed.

2.2.1 Dataset Description

Primary data in this chapter consists of a large dataset of gas chromatographic retention times curated by the National Institute of Standards and Technology (NIST), specifically the data included in the 2011 version of the mass spectral library [5]. The full dataset contains 317,310 entries with relevant features summarized in Table 2.1. The dataset also includes other structural representations (InChI) and various operational parameters related to the chromatographic analysis (temperature ramps, etc.); these data represent 55,798 unique structures, each with multiple entries containing different chromatographic data.

For each unique compound in the NIST, physicochemical properties are estimated from the molecular structure using the EPI Suite™ Estimation Program Interface provided by the US Environmental Protection Agency [49]. The SMILES string, a machine- and human-readable description of structure, is used to estimate Henry's Law Constants using the bond contribution method of the HenryWin modules, vapor pressures using the MPBPVP module, and solubility using the WSKOW module. Experimentally determined values are used for these properties when provided by EPI Suite™, but in most cases estimation is based on

Table 2.1: Overview of dataset features extracted from the 2011 NIST mass spectral library, including compound identifiers, structural descriptors, and chromatographic parameters.

Feature Name	Description
`RecordNo`	Unique compound identifier
`Name`	Compound name
`Formula`	Chemical formula
`SMILES`	Molecular representation
`ColumnType`	Chromatographic column type
`PhaseType`	Stationary phase type
`IndexType`	Retention index standard
`Phase`	Stationary phase identifier
`Index`	Retention index value

quantitative structure-activity relationships.

2.2.2 Dataset Preprocessing and Transformation

The following preprocessing steps are taken to prepare the dataset for model training:

1. Filtering: Only entries with a column type of `Capillary` to limit analysis to capillary chromatographic columns, which are most common in environmental studies. Only entries with an index type of `Normal alkane RI`, in which each carbon number increase represents an increase of 100 units, are kept to minimize unnecessary variability and complexity from including less common index types that are often specialized for particular applications. These filters reduce the dataset to 133,847 entries.

2. Log Transformation: The three target physicochemical properties are log-transformed to improve learning stability, reduce the influence of outliers, and accommodate the large range observed in the raw property values (20 orders of magnitude for all properties, with some outliers beyond this).

3. Phase Normalization: Several phases in the dataset are functionally equivalent but

labeled differently due to branding or minor variations. For modeling purposes, these are grouped under unified labels. For example, `Chromsorb 101`, `HP-101`, and `OV-101` are all replaced with `101`. The main categories after this consolidation include `WAX 20M`, `FFAP`, `101`, `100%`, and `5%`. Any phases not fitting these categories are kept as is. Since these phase labels are categorical, but the models require numerical inputs, they are one-hot encoded to convert each category into a set of binary indicator variables that can be passed to the models.

The Property-to-Index model uses this preprocessed dataset, where each entry is formatted as: `RecordNo` | `Chemical Properties` | `Phase` | `Index`

After preprocessing, several derivative datasets are constructed to support training the Index-to-Property model and compound identification experiments. Each dataset is derived from the NIST database but differs in structure and intended use:

- Complete Compound Dataset: Catalog of all 55,798 unique compounds from the original NIST library, each with estimated physicochemical properties, with no specific retention information retained.

- Paired-Index Dataset: This dataset is created by restructuring the preprocessed NIST dataset into compounds with pairs of phases and their corresponding retention indices. For each unique compound, all unique pairwise combinations of different stationary phases are generated. To focus on chromatographically orthogonal relationships, only pairs consisting of one polar and one non-polar phase are retained. Each entry in this dataset therefore contains an orthogonal phase pair with their respective retention indices and the associated physicochemical property. The resulting dataset contains 871,764 rows representing 3,890 unique compounds, each formatted as: `RecordNo` |

`Phase A` | `Phase B` | `Index A` | `Index B` | `Chemical Properties`. Each entry represents a "real-world" pair of indices measured on individual experimental setups. This dataset is used to train the Index-to-Property Model.

- Compounds-with-Pairs Dataset: For each unique compound in the Paired-Index Dataset (i.e., all compounds containing at least one pair of orthogonal indices), a dataset of average retention indices is created. Created by averaging the retention indices for each unique compound on each of its phases with known retention indices alongside the physicochemical properties of the compounds. Each entry represents the best known average indices of a compound, but only for compounds with known information for at least one set of orthogonal pairs (i.e., 3,890 compounds), so is used to evaluate the Index-to-Property Model as well as the Property-to-Index Model.

2.2.3 Model Setup

The eXtreme Gradient Boosting (XGBoost) [6] regressor is used for both the Index-to-Property and Property-to-Index models to predict the relationships between chemical properties and retention indices. This model is chosen due to its flexibility and high performance on structured, tabular data.

For the Property-to-Index model, a Multi-Layer Perceptron (MLP) [43] model is also trained alongside the XGBoost model to support visualization and interpretability. Although initial testing shows that XGBoost achieves better predictive performance compared to MLP (see Section 4.4), its decision tree architecture produces inherently discontinuous output surfaces. However, in order to examine and interpret the results, it is useful to understand retention index predictions as smooth functions of physicochemical properties (e.g., heatmaps and contour plots across vapor pressure and Henry's law constant space), so the MLP architecture

is also used despite its reduction in accuracy. An MLP provides continuous outputs that enable meaningful interpolation across regions of property space. For applications in which this smooth output is not required, the improved accuracy of the XGBoost architecture could be applied, which is discussed in the text.

The Optuna library [1] is used to explore the hyperparameter space and optimize performance for both models. The best-performing hyperparameters for each model are summarized in Tables 2.2 and 2.3.

Table 2.2: Hyperparameters for the Index-to-Property Model (XGBoost)

Parameter	**Value**
Number of Estimators	555
Maximum Tree Depth	8
Subsample Ratio	0.631
Learning Rate	0.299
Gamma	0.033

Table 2.3: Hyperparameters for the Property-to-Index Model (MLP)

Parameter	**Value**
Input Layer	69 Neurons: 3 physicochemical properties + one-hot encoded phase representation (66 unique phases)
Hidden Layer 1	440 Neurons, Rectified Linear Unit (ReLU) Activation
Hidden Layer 2	440 Neurons, ReLU Activation
Output Layer	1 Neuron, Linear Activation
Learning Rate	0.0017896
Max Epochs	20
Early Stopping Patience	3 Epochs

An 80/20 train-test split and 10-fold cross-validation are used to avoid overfitting and ensure model reliability, as recommended for reliable model evaluation [44]. These validation strategies ensure that model performance is both accurate on the training set and generalizable to unseen data.

2.2.4 Evaluation Metrics

Both models are evaluated using the following regression metrics:

- Mean Squared Error (MSE): Average squared difference between predicted and actual values. Lower is better.
- Mean Absolute Error (MAE): Average absolute difference between predicted and actual values. Lower is better.
- R^2 Score: Proportion of variance in the predictions that is explained by the actual values. A score of 1.00 indicates a perfect fit.

For the Index-to-Property model, each predicted property is evaluated separately.

In addition to these quantitative metrics, individual prediction errors (residuals) are also used to qualitatively assess model performance. Here, residuals are defined as the absolute differences between the predicted and actual values for each observation, $r_i = |y_i - \hat{y}_i|$, where y_i and $\hat{y}_i$ represent the true and predicted values, respectively.

2.2.5 Property-Based Identification

In addition to predicting properties from retention indices, the Index-to-Property model facilitates the identification of an unknown compound for which multiple retention indices can be known (e.g., in an instrument with multiple columns). The model can narrow down candidate compounds by comparing predicted physicochemical property profiles with known database entries. The predicted property vector can then be used to rank all entries in the dataset by distance, which results in an ordered list of predicted compounds.

Experimental Setup

To evaluate and compare the efficacy of retention-index-based and property-based compound identification, three experiments are conducted. In each case, a random entry is selected from the held-out 20% test portion of the Paired-Index Dataset and used as the query. Identification accuracy is evaluated by ranking all candidate compounds in the search database according to their Euclidean distance to the query in the relevant feature space.

1. Index-Based Search: The query's pair of retention indices is compared against averaged indices for all compounds in the Compounds-with-Pairs Dataset. Compounds are ranked by their distance in index space. This experiment represents the ideal case where both orthogonal indices are known for the compound of interest.

2. Property-Based Search (Compounds-with-Pairs Dataset): The same query's indices are passed through the Index-to-Property model to predict a vector of physicochemical properties. These predicted properties are then compared against the known property vectors of compounds in the Compounds-with-Pairs Dataset, using Euclidean distance in property space. This experiment evaluates how well predicted properties can reproduce the identification accuracy achievable from direct retention index comparison.

3. Property-Based Search (Complete Compound Dataset): The same query's indices are again passed through the Index-to-Property model to predict a vector of physicochemical properties. The predicted property vector is compared against all entries in the Complete Compound Dataset. This experiment determines whether property-based prediction can enable identification across a larger chemical space, beyond the compounds for which orthogonal retention indices are known.

Experiments 2 and 3 are conducted using both the full set of predicted properties (vapor

pressure, solubility, and Henry's law constant) and a reduced set (vapor pressure and solubility only) to evaluate the impact of the less reliable HLC prediction on identification performance.

Evaluation Metrics

Performance is measured using the rank of the correct compound in the ordered list of candidates, where a rank of 1 indicates a perfect identification. To account for differences in database size between Experiments 2 and 3, a normalized rank (r_{norm}) is also computed, representing the true compound's position as a fraction of the total number of candidates:

$$r_{\text{norm}} = \frac{r_{\text{true}}}{N_{\text{database}}},$$

where N_{database} is the number of compounds in the corresponding search database. This normalization enables direct comparison of identification performance between the Compounds-with-Pairs and Complete Compound datasets.

2.3 Results and Discussion

2.3.1 Index-to-Property Model

Table 2.4: Index-to-Property Model Evaluation Metrics

Property	**Mean Squared Error**	**Mean Absolute Error**	R^2 **Score**
Vapor Pressure	0.058	0.134	0.970
Solubility	0.108	0.203	0.973
Henry's Law Constant	0.278	0.301	0.936

Using pairs of orthogonal retention indices, the Index-to-Property model is able to predict

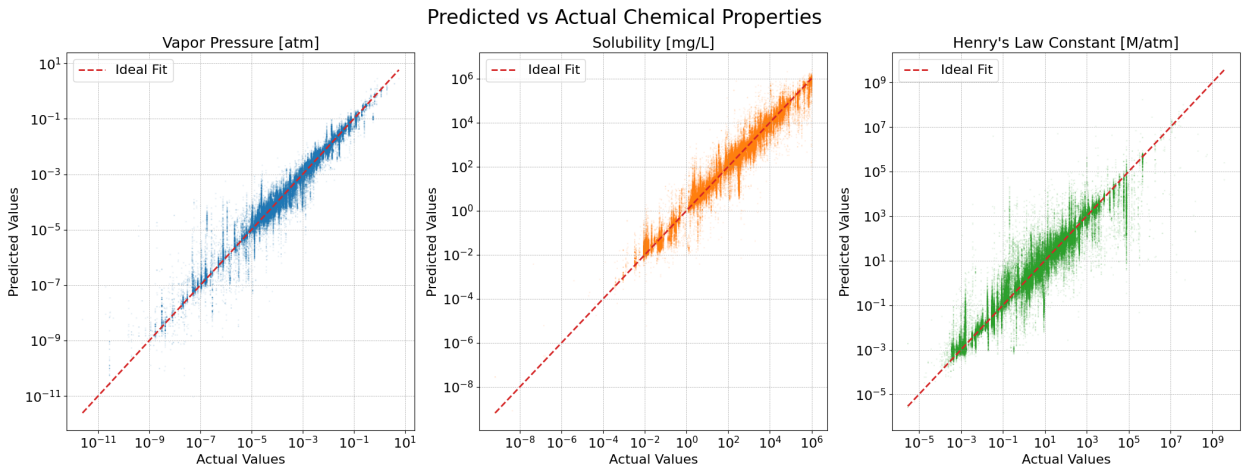

Figure 2.1: Scatterplots of predicted versus actual chemical properties for the Index-to-Property model. The dashed red lines represent ideal fits.

physicochemical properties with R^2 between 0.94 and 0.97, and mean absolute errors of less than 0.3 orders of magnitude for all properties (Table 2.4). The model performs best for vapor pressure prediction, with an R^2 of 0.97 and within 0.15 orders of magnitude (i.e., ~40%). For both VP and HLC (and likely for solubility, though it is less studied), the degree of error is within the uncertainty of the group contribution methods used to estimate the physicochemical properties. The error in the model is a combination of both model inaccuracy and the uncertainty in the underlying data (i.e., it is just as likely that the estimations are 0.2 orders of magnitude off as it is that the prediction is 0.2 orders of magnitude incorrect). Consequently, these metrics suggest the model is performing approximately as well as could be expected or achieved.

Strong predictive performance, especially for VP and sol, is evident from the clustering of predicted and actual values along the identity line (Figure 2.1). The wider spread in HLC predictions is consistent with the higher uncertainty in its underlying structure-based estimations relative to vapor pressure [9, 23, 42].

Figure 2.2 shows the cumulative distributions of prediction errors for the three target prop-

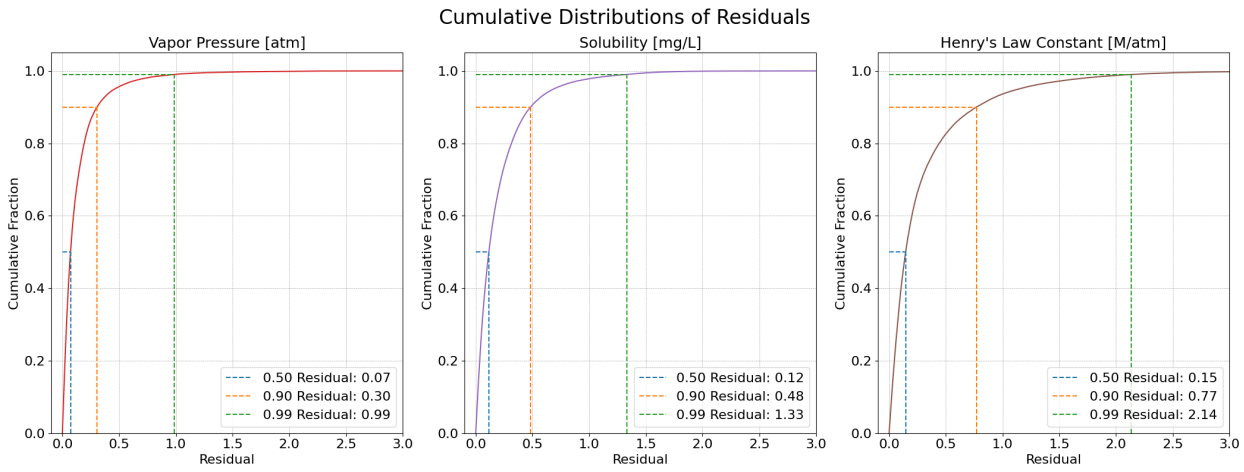

Figure 2.2: Cumulative distributions of residuals for the Index-to-Property model across the three predicted properties. Dashed lines indicate the residual magnitudes corresponding to the 50th, 90th, and 99th percentiles of the absolute error distribution.

erties. As expected, the VP residuals reach higher cumulative probabilities at smaller error magnitudes, indicating higher predictive performance compared to the other properties, which have broader distributions with higher median residuals. The 90-95% threshold is roughly in line with known errors in property prediction for vapor pressure and Henry's Law Constants (roughly 0.5 and 1.5 orders of magnitude, respectively [23]), and suggest that almost all predictions are nearly as accurate as possible.

The relatively strong predictive performance demonstrated in Table 2.4 and Figures 2.1–2.2 confirms that paired retention indices from orthogonal phases serve as effective proxies for estimating physicochemical properties. This capability is especially valuable for environmental samples where many compounds lack structural identification, as it allows for property estimation directly from chromatographic data without requiring reference spectra or known structures. The high accuracy for VP prediction suggests that retention behavior strongly correlates with volatility, while the performance with ranges of known error for the other properties demonstrates the model's ability to capture more complex relationships. These predicted properties enable a novel approach to compound identification for analytes that

are not identified but are expected to appear within a list of possible candidates. This can be demonstrated best through a case study, as follows.

Chemical Identification Using Property Prediction

The strong predictive performance of the Index-to-Property model enables potential use for compound identification, especially for analytes lacking reference information.

To illustrate this approach, consider the analyte ethylbenzene, a common industrial solvent and constituent of petroleum products and thus a potential analyte of interest within environmental or industrial samples. Estimated physicochemical properties for this compound are: vapor pressure of $10^{-1.90}$ atm, solubility of $10^{2.23}$ mg/L, and HLC of $10^{-0.90}$ M/atm. Having a retention index of 1126 on a polar phase (Carbowax 20M) and of 843 on a non-polar phase (DB-1), the Index-to-Property model predicts as properties of ethylbenzene: vapor pressure of $10^{-1.94}$ atm (error of 0.04 log units), solubility of $10^{2.38}$ mg/L (error of 0.15 log units), and HLC of $10^{-0.75}$ M/atm (error of 0.15 log units). Mean squared distance between the predicted property set and the properties of every other compound in the Complete Compound dataset can be calculated to generate a rank-order list of the most likely candidate.

The set of properties for this compound identifies ethylbenzene as the 24th most likely candidate. This is illustrated in Figure 2.3 for only the vapor pressure and solubility properties, for which 23 compounds in the dataset (orange dots) are closer to the predicted properties than the target, which drops to 13 using all three properties but is harder to visualize in 3-property space. Though not perfect, this approach narrows the full dataset of 55,798 compounds down to a handful of compounds, which would be valuable, particularly in the case of a more limited candidate list.

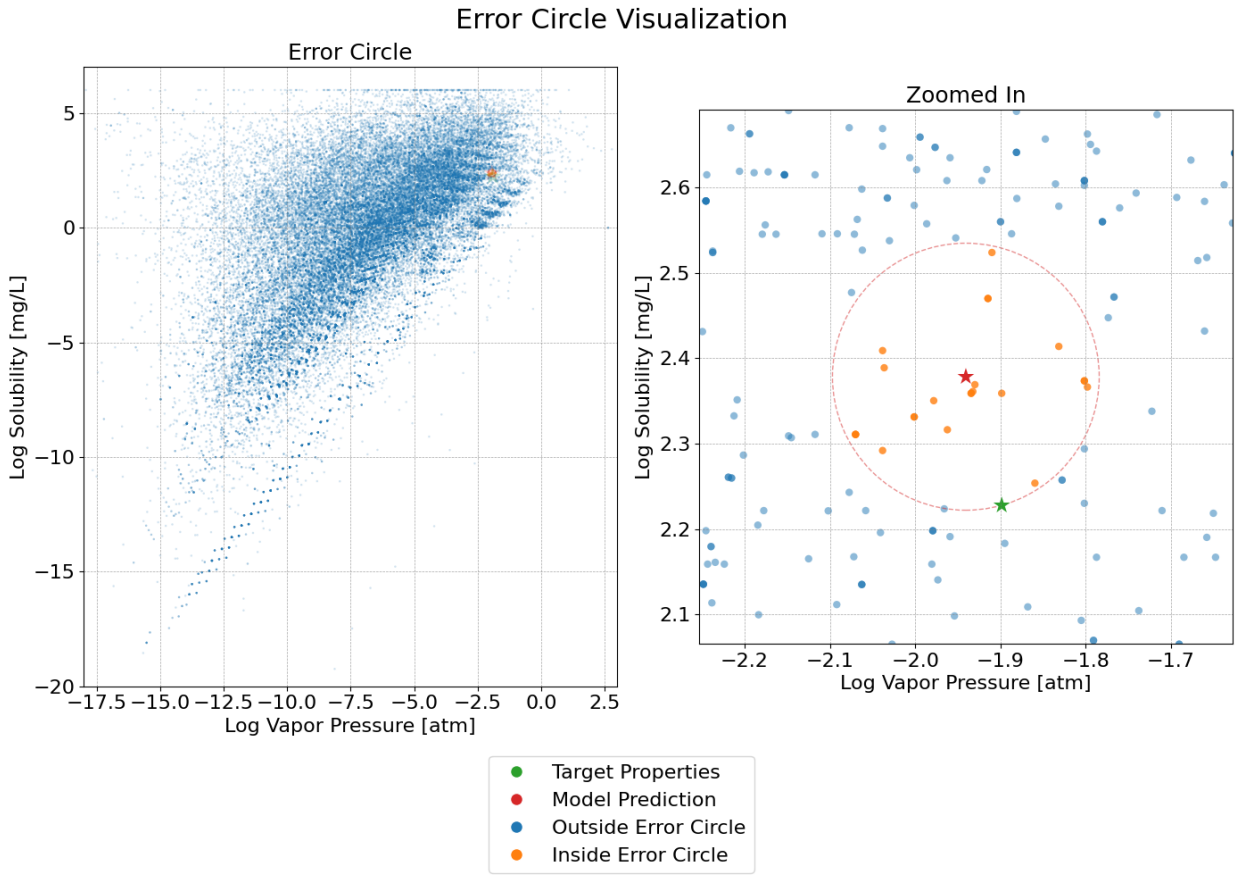

Figure 2.3: Chemical identification using vapor pressure and solubility (HLC not included for visualization purposes). The graphs show database chemicals arranged by log vapor pressure and log solubility as blue dots. The red star indicates the position of the predicted properties, while the green star indicates the actual target chemical. The dashed red circle encloses chemicals that more closely match the prediction compared to the target, which are shown in orange.

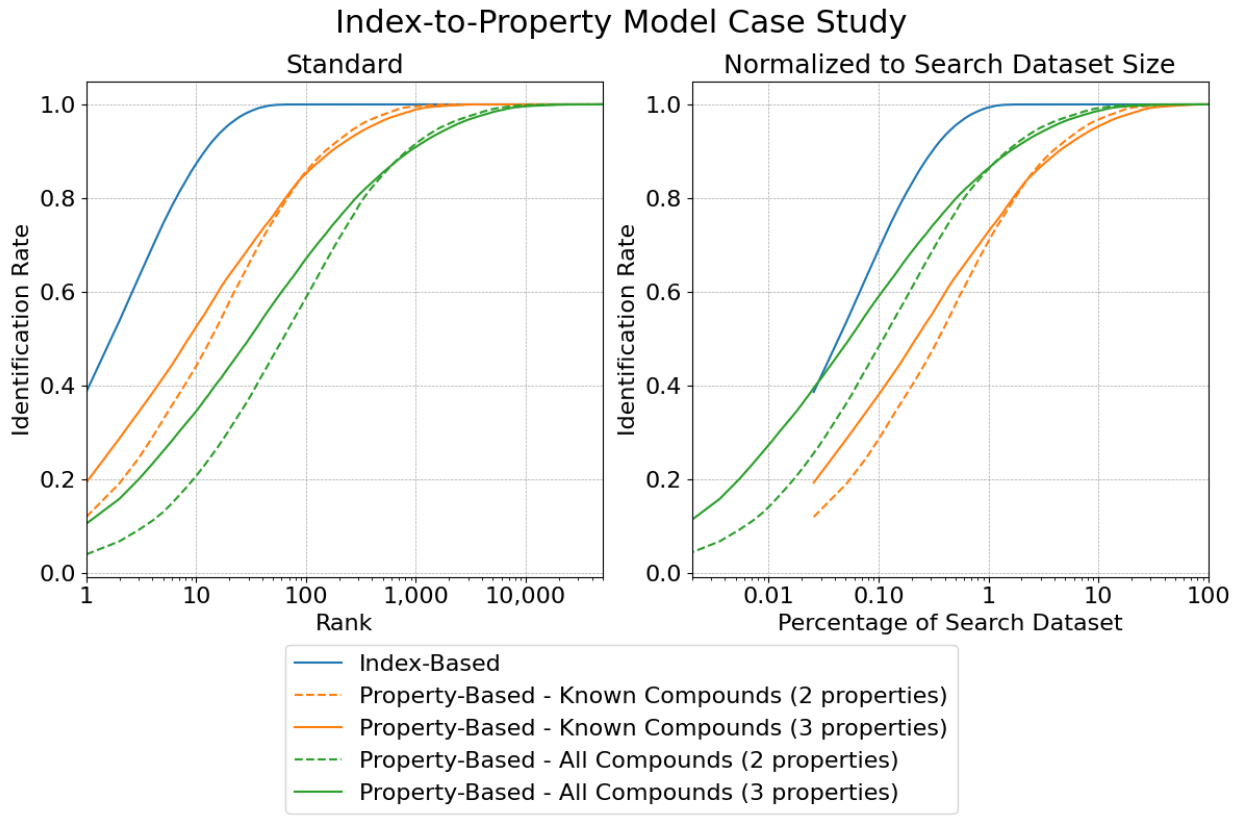

Figure 2.4: Cumulative Match Characteristic (CMC) curves for the index-based and property-based identification experiments. The left pane shows correct identification rate as a function of maximum rank considered, while the right pane shows identification rate as a function of search dataset size.

Building on this example, a series of identification experiments are conducted to evaluate this approach, comparing the effectiveness of direct retention index matching against property-based identification using model predictions. As detailed in Section 2.2.5, these experiments assess how well a query compound can be identified within candidate databases using different feature representations.

The key results are summarized in Figure 2.4 and Table 2.5. Using an index-based search directly, the correct compound is found within the top 12 candidates 90% of the time, and half the time it is one of the top two candidate compounds. In other words, if two orthogonal indices are known for a compound, then a measurement on these two phases is very good at

Table 2.5: Percentile rank results for the Index-to-Property model compound identification case study.

Experiment	Rank Percentiles		
	50th	**90th**	**99th**
Index-Based	2 (0.05%)	12 (0.3%)	35 (0.9%)
Property-Based – Known Compounds			
2 properties	14 (0.36%)	149 (3.8%)	690 (17.7%)
3 properties	9 (0.23%)	170 (4.4%)	1033 (26.6%)
Property-Based – All Compounds			
2 properties	62 (0.11%)	822 (1.5%)	4976 (8.9%)
3 properties	31 (0.06%)	872 (1.6%)	6536 (11.7%)

identifying the compound out of the set of known compounds. This performance represents the upper bound for identification accuracy, as it does not suffer from model prediction errors, and confirms that paired orthogonal indices provide a relatively unique fingerprint for compounds with known retention data.

When retention indices are converted to physicochemical properties using the Index-to-Property model, identification accuracy decreases as expected due to model uncertainty. Using all predicted properties, the 90th percentile rank increases to 170, and the median to 9. Excluding Henry's law constant (the least accurate prediction per Table 2.4) yields slightly better performance in the higher percentiles, likely due to the fact that adding a noisy third dimension to the predictions can expand the search space more than it narrows down candidate compounds. There is no inherent value to using properties derived from indices to query a library for which indices are known, and indeed the accuracy of this approach is found to decrease. However, this approach has a substantial value, which is that it can be used to query a library for which indices are not known. In other words, a measurement of indices on two orthogonal phases can be used to identify compounds in a database that does not include retention time information for orthogonal indices, or indeed any retention time information at all.

The advantage of property-based search can be seen when querying the Complete Compound Dataset (∼55,000 compounds), the majority of which lack measured retention index pairs. Despite a significant increase in database size, the property-based search maintains reasonable accuracy, with the correct compound appearing within the top 31 candidates 50% of the time. Notably, performance relative to is actually better when accounting for database size (right pane of Figure 2.4), coming close to achieving the metrics of a direct index-based search. Half the time, the target compound is within the top 0.1% of the database, and within 1.5% of the database 90% of the time. While less accurate than index-based search, this approach can still narrow the candidate compound list by over 98% with reasonable confidence even in the absence of retention index information.

2.3.2 Property-To-Index Model

The performance of the Property-To-Index Model is summarized in Table 2.6 and visually evaluated in Figure 2.5. The high R^2 score and tight clustering along the ideal fit line indicate strong performance, with an MAE of 25 suggesting that on average this method predicts retention indices within 25 units of observations. More than 90% of predictions are within ±50 units, i.e., the same carbon number bin of the observed value (Figure 1b). For many compounds, multiple retention indices are provided for the same column type from multiple experimental conditions and studies, and in these cases, differences in RI of tens of units are common. The ability of the prediction model to get within this range consequently suggests it is usually as accurate as is possible, though there are a small number of outliers with very high residuals up to a few hundred units.

Table 2.6: XGBoost Property-To-Index Model Evaluation Metrics

Mean Squared Error	**Mean Absolute Error**	R^2 **Score**
4356	25.02	0.9820

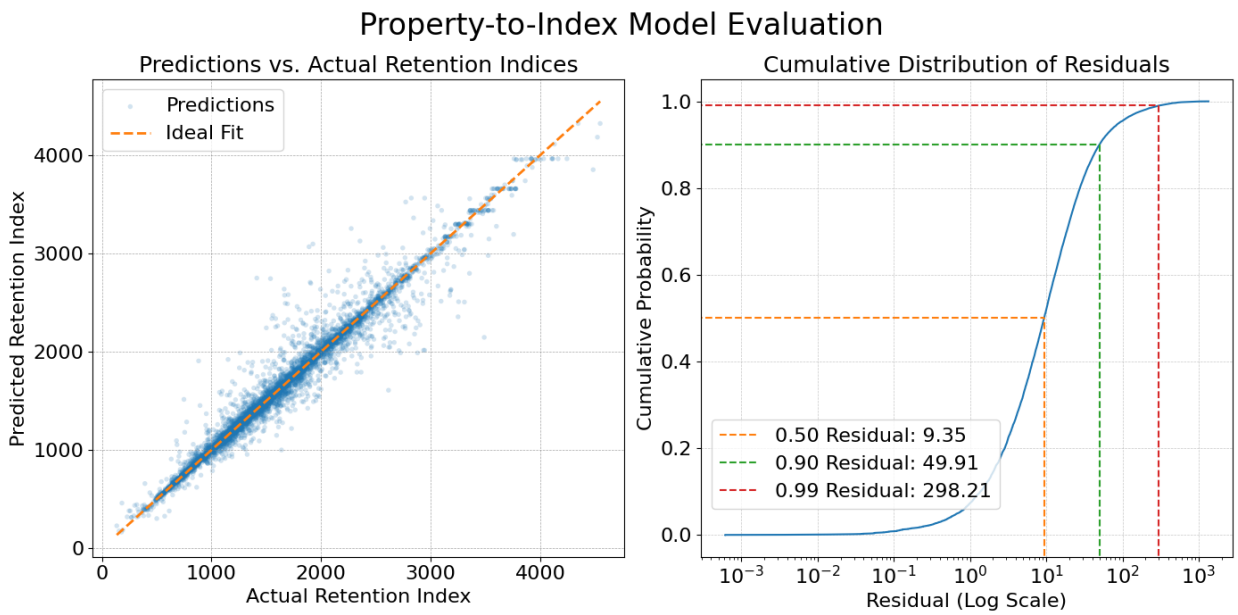

Figure 2.5: Evaluation of the XGBoost version of the Property-To-Index Model on the test dataset. The left pane is a scatterplot of predicted versus actual RIs, with a dashed identity line indicating the ideal fit. The right pane is a cumulative distribution of residuals with the x-axis on a log scale.

Using the Property-To-Index Model, we can conceptually illustrate the value of using paired orthogonal columns for property estimation (Figures 2.6–2.7). Because XGBoost actually generates non-continuous relationship spaces, we use an alternate model to visualize the relationships between properties and paired indices. A Multi-Layer Perceptron (MLP) model generates continuous spaces, which is more physically meaningful, but at the expense of some model performance (see Figure 1 and Table 1). Consequently, it is valuable for interpretation, but not in cases where the sole purpose is property prediction. Using the MLP model, Figure 2.6 shows heatmaps of the predicted RI on a polar column (WAX 20M) and a nonpolar column (5% Phenyl) as a function of VP and HLC. Predictions above an RI of ∼2600, where data are limited, are excluded. The distinct shapes of the two heatmaps confirm that the model has learned the differential effect of each stationary phase on retention behavior. Additionally, prior work has shown that there is a general relationship in the

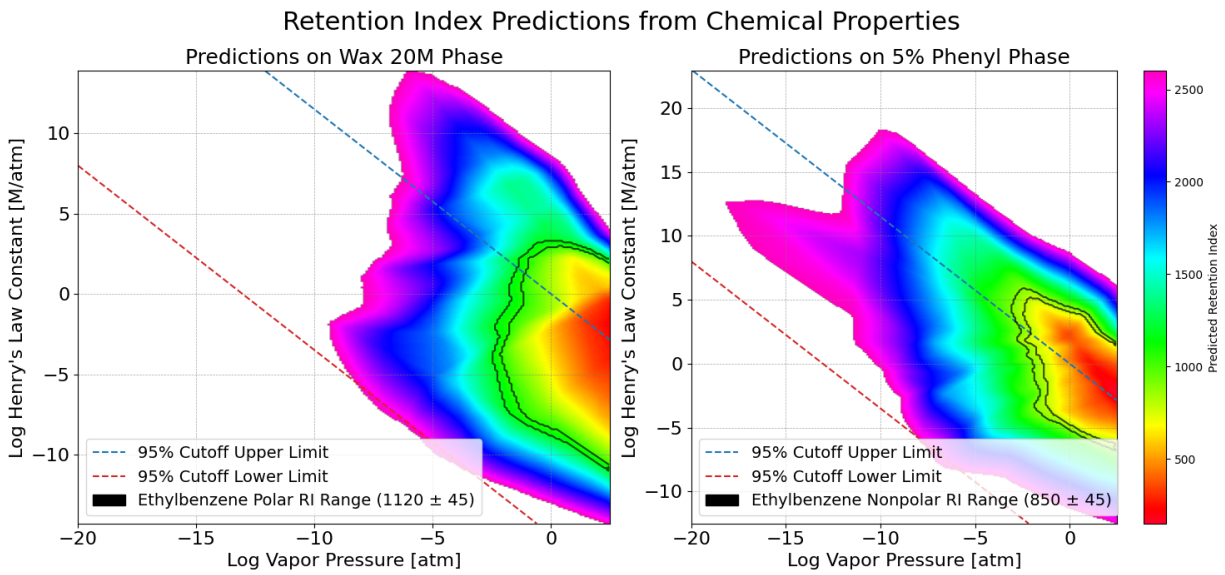

Figure 2.6: Heatmap showing retention index predictions from the Property-To-Index Model as a function of VP and HLC. The outlined RI ranges correspond to ethylbenzene as an example, which has polar and nonpolar RIs of 1120 and 850, respectively. The model median residual of ±45 is used to define a range of interest. Note that predictions above 2600 are clipped because the model extrapolates into non-physical space. Also note that solubility is held constant at its median value for visualization purposes.

atmosphere in which HLC increases by approximately one order of magnitude for every one order of magnitude decrease in vapor pressure, so these physicochemical parameters tend to be roughly related as log(VP) $\propto$ -log(HLC) [16, 24]. Within the training data, less than 2.5% of data is either above the line log(VP) = -log(HLC) or below the line log(VP) = -15-log(HLC), shown by the dashed lines in (Figures 2.6); these regions therefore represent space in which the model is poorly trained and are also generally non-physical, so can be excluded.

The outlined regions in Figure 2.6 correspond to the RI values for an example compound, ethylbenzene (RI of 1120 on WAX 20M and 850 on 5% Phenyl). The banded areas, defined by the model's median residual (45 units), represent the most likely range of possible indices for a given set of chemical properties. These bands demonstrate a significant limitation of

traditional correlations between indices and properties: for a single column, a given retention index corresponds to a large range of possible properties, and also, therefore, a wide range of potential compounds.

The value of orthogonality is demonstrated when the index ranges from both columns are overlaid, as shown in Figure 2.7. The combination constrains the possible candidates to a small overlapping region in property space; though another small region of overlap exists with higher VP and higher HLC, this overlap falls outside the well-trained and physically meaningful region of chemical space. This visualization demonstrates that a pair of orthogonal columns defines a unique region in physicochemical property space that represents a set of physicochemical properties for an analyte even in the absence of an identification. It also supports the conclusions of Section 2.3.1 that a compound can be relatively uniquely identified from a pair of orthogonal retention indices by predicting its physicochemical properties. The overlapping region demonstrates how critical the use of multiple retention indices is in significantly narrowing the possibilities compared to a single column, enabling more precise compound identification and property estimation.

2.4 Chapter Summary

This chapter demonstrates the value of combining orthogonal chromatographic data with machine learning to bridge the gap between retention behavior and physicochemical properties. The core novelty lies in the specific, bidirectional approach that uses *pairs* of retention indices and *sets* of properties, not just in applying ML to chromatographic data. This idea reveals that the combination of these multidimensional data points provides significantly more chemical information than any single measurement. The Index-to-Property and Property-to-Index models together create a structure-agnostic mapping between the chromatographic

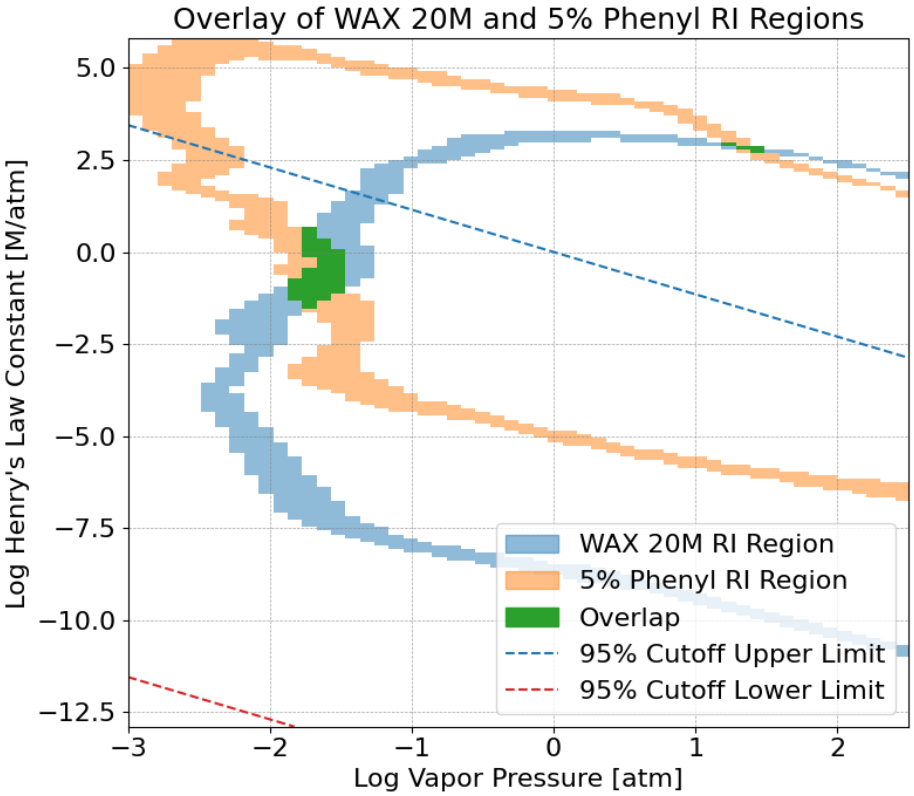

Figure 2.7: Overlay of WAX 20M and 5% Phenyl RI regions of interest highlighted in Figure 2.6, and their overlap.

and physicochemical property spaces.

The success of these models shows that the retention behavior of a compound on orthogonal phases acts as a relatively unique fingerprint for its underlying properties, especially volatility. This insight opens the door for novel environmental applications, enabling analytical approaches or instruments that rely solely on dual-column retention data to describe the collective fate and partitioning behavior of complex samples without the need to identify every constituent. Additionally, this methodology offers a viable path to compound identification and characterization without relying on costly and complex detectors like mass spectrometers, making sophisticated chemical analysis more accessible and deployable in resource-limited or field settings. By enabling property estimation and similarity matching directly from retention data, this work takes a step towards understanding the composition and environmental impact of the unknown fraction inherent in complex mixtures.

While this method addresses identification using derived properties, the next chapter investigates a more direct approach using the raw chromatographic signal itself.

Chapter 3

Machine Learning-Based Prediction of Compound Identity and Concentration from GC Chromatograms

3.1 Introduction

3.1.1 Problem Motivation

Chromatographic analysis of environmental samples presents a unique data reduction challenge. Atmospheric and environmental mixtures are inherently complex, often containing thousands to tens of thousands of distinct compounds [14]. While modern gas chromatography (GC) instrumentation is capable of capturing this complexity, the subsequent step of transforming detector response over time into a time-series of compound concentrations remains a significant bottleneck [26].

Traditional workflows for quantifying known compounds rely heavily on "drawn baseline" integration, where software or analysts manually define peak start and end points. While effective for simple mixtures, this approach is subjective, time-consuming, and scales poorly

to large datasets generated by long-term or continuous collection [26]. As high-throughput "fast GC" [34] instruments increase data generation rates by an order of magnitude, the reliance on manual inspection and user-intensive quality control becomes unsustainable [26, 31].

3.1.2 Background and Prior Work

To address the limitations of drawn baselines, several studies have focused on automating the interpretation of complex chromatograms through mathematical modeling and advanced signal processing.

Mathematical Peak Fitting and Factor Analysis

Automated peak fitting, where experimental data is described using idealized mathematical functions, such as Gaussian or Exponentially Modified Gaussian (EMG) curves, offer a more effective alternative to baseline integration [26]. While this allows for the deconvolution of overlapping peaks and provides quantitative statistical moments (width, skew), it often still requires significant user input to constrain fits or assumes idealized shapes that may not match physical reality [26].

Matrix decomposition methods such as parallel factor analysis (PARAFAC) [19] and positive matrix factorization (PMF) [39] have been used to resolve co-eluting peaks across large datasets [31]. These techniques use the covariance of ions to mathematically separate overlapping signals, effectively deconvolving the mixture into constituent spectra and concentration profiles [31]. However, these approaches generally rely on the multidimensional data provided by mass spectrometry (MS), and are computationally expensive and inapplicable to datasets collected using non-spectral detectors (e.g., flame ionization detectors) or

lower-cost field instrumentation.

Machine Learning in Chromatography

Machine learning (ML) offers a potential solution for automating data reduction with lower computational overhead, and some progress has been made, particularly in liquid chromatography (LC) data. For instance, ML has been successfully applied to distinguish true peaks from noise. Traditional algorithms often produce a significant number of false positives that unnecessarily lengthen the analysis process due to incorrect parameter settings or noise interference. This issue has been addressed through the use of convolutional neural networks (CNNs) to classify peaks, demonstrating superior performance compared to traditional approaches [15, 30, 37].

Two studies preprocess raw LC-MS data into 64×64-pixel grayscale images, which are then fed through a CNN that distinguishes peaks from noise [15, 30]. An Area Under the Curve (AUC) score of 0.988 was achieved in one study, removing 90% of false positives while retaining 98% of true positives [30]. Another achieved 90-95% accuracy across several LC-MS configurations, highlighting the capabilities of image representations for capturing complex signal-noise relationships [15].

Synthetic data generation is used in a separate study to augment training, helping address the lack of labeled data [37]. This work also uses a dual-CNN approach, where the first model distinguishes real peaks from noise and the second model segments and quantifies individual peaks, achieving 97% precision in peak detection while reducing reliance on manually labeled data [37]. However, the authors acknowledged that the generated synthetic data may not accurately represent real-world variability [37].

Several trade-offs and limitations were common across the studies, the first being the balance

of precision and recall. One study sacrifices 11% recall to achieve high precision, while another prioritizes true positives (98% retention) [30, 37]. Logistic regression and random forest models evaluated in one study, while interpretable, sacrifice precision compared to the CNN model (66% false positive removal vs. 90%) [30]. All three studies rely heavily on manual peak labeling for training purposes, limiting scalability, and generalizability across different chromatographic conditions was not evaluated [15, 30, 37].

These existing ML approaches focus primarily on peak detection, which typically serves as a preprocessing step that still feeds into a traditional integration workflow, rather than replacing the integration process itself.

3.1.3 Identification Power of Peak Shape

Current automated methods rely heavily on retention indices or MS spectral matching for identification. It is widely known that retention index is a strong identifier for a compound. However, less frequently considered is that the shape of a chromatographic peak is also a function of the chemical identity and its interaction with the stationary phase. This suggests that the raw shape of a peak (fronting and tailing behavior, width, etc.) contains latent information that can be exploited for both identification and quantification [29], potentially serving as a proxy for the resolving power of MS in simpler detector setups. While the potential utility of peak shape has been recognized for decades, with detailed quantitative models relating shape parameters to chemical properties and identification [7, 28, 38], and has been proposed as a quality assessment indicator in methods reliant on peak fitting [26], using peak shape has been hard to deploy at scale because individual mathematical parameters or simplified peak descriptors are not sufficiently unique for identification.

3.1.4 Chapter Objectives

In this chapter, we introduce a method that processes chromatographic data directly to automate the labor-intensive data reduction process for known compounds. Our approach (a) exploits the identifying power of peak shape to classify compounds without the need for mass spectrometry, and (b) provides calibrated concentration output without the need for time-consuming peak integration or explicit curve fitting. To achieve this, we first construct a large labeled peak dataset derived from calibration chromatograms that enables supervised learning without manual annotation. We then use this dataset to develop and evaluate two machine learning models: a peak-shape-based classification model that distinguishes among 11 compounds, plus an "Unknown" class, using only 150-point peak segments with all timing information removed, and a concentration prediction model that maps peak shape directly to concentration.

3.2 Materials and Methods

The primary objective of this chapter is to develop two ML models for the analysis of raw chromatographic data: a compound prediction model to identify the compound associated with an individual chromatographic peak and a concentration prediction model to quantify its concentration. Conceptually, these models identify patterns and subtle differences in chromatographic peak shapes to learn what visually distinguishes one compound from another and how changes in concentration manifest in the raw data. Critically, a major goal of this work is to examine the ability of a machine learning model to reduce data based on features not necessarily recognizable to humans, so we focus here on providing data that has been stripped of its retention time information (i.e., isolated peaks). To learn these relationships effectively, a large dataset of labeled example peaks is required.

3.2.1 Dataset Description

We use a dataset of gases measured in a forest in central Virginia, collected from 2019 to 2024. Gases are sampled from the mid-canopy at a height of ∼20 m through a heated inlet into an automated gas chromatograph equipped with a multi-bed gas adsorbent trap. This instrument and sampling configuration are described in detail by McGlynn and co-workers [35]. The raw dataset consists of 35,150 chromatograms. The majority of these are unlabeled field samples, but a subset of 2553 chromatograms are calibration runs where the trap is supplied with a mixture of 11 known compounds at known concentrations, detailed in Table 3.1. These calibration chromatograms serve as the basis for constructing a labeled peak dataset. Data pre-processing, such as retention time alignment as described below, is performed using the publicly available analysis package "TERN" [4, 26] in the Igor Pro 9 programming environment (Wavemetrics, Inc.).

Table 3.1: Known Compounds in Calibration Chromatograms

Compound	Retention Time [s]	Concentration [ppbv]			
		callevel1	callevel2	callevel3	tracking
Pentane	254	0.11	0.51	1.03	0.26
Isoprene	287	0.29	1.34	2.69	0.67
Hexane	434	0.07	0.33	0.66	0.17
Methyl Vinyl Ketone	554	0.13	0.58	1.17	0.29
Benzene	751	0.11	0.50	1.00	0.25
a-Pinene	1389	0.13	0.59	1.17	0.29
1,3,5-Trimethylbenzene	1472	0.09	0.41	0.83	0.21
Limonene	1567	0.06	0.28	0.55	0.14
Nopinone	1860	0.06	0.30	0.59	0.15
α-Cedrene	2216	0.03	0.15	0.29	0.07
α-Humulene	2274	0.03	0.15	0.29	0.07

3.2.2 Peak Segmentation

The labeled dataset is constructed via a five-step peak segmentation process, performed on each chromatogram separately:

1. Retention Time Alignment: Each chromatogram is aligned using parameter-optimized warping in an automated process incorporated into TERN. A fourth-order polynomial is applied to the raw retention time $w(x) = ax^4 + bx^3 + cx^2 + dx + e$. The polynomial coefficients are fit to maximize the correlation coefficient between the aligned chromatogram and a reference calibration chromatogram; first- through fourth-order polynomial fits are tested, and the final coefficients are selected as those that maximize the correlation coefficient. Initial guesses for coefficients are determined through correlation-optimized warping: the retention time is divided into n even lengths, each of which is shifted to maximize correlation with the reference, and an n-order polynomial is fit through these shifts to yield initial guesses for the final n-order polynomial fit using parameter-optimized warping. This process reduces retention-time misalignment so that downstream peak detection and matching steps operate on consistently aligned chromatograms.

2. Peak Detection: Potential peaks are identified within each chromatogram by first smoothing the chromatogram with a Gaussian filter to reduce high-frequency noise, after which local maxima are identified as candidate peaks. Peak prominence is then used to rank these candidates, and only the 30 most prominent peaks are retained as detected peaks.

3. Peak Matching: The detected peaks are matched to the 11 known compounds by constructing a cost matrix using the absolute difference between each known compound's retention time and each observed peak's retention time and computing an optimal

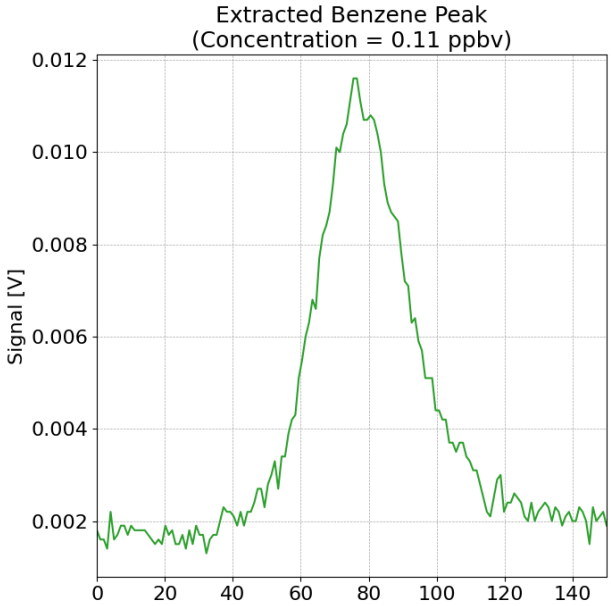

Figure 3.1: Example of a single extracted peak using the peak segmentation process. Note that the original time axis is discarded and replaced by a generic index axis (see Section 3.2.3).

one-to-one assignment that minimizes the total cost (the sum of all retention time differences) (Figure 3.2).

4. Peak Filtering: Matched peaks are only kept if their observed retention times fall within ±10 s of the compound's known retention time (∼5% of one carbon number). This tolerance accounts for normal retention time variability while removing mismatches caused by occasional detection errors that would otherwise create mislabeled training examples.

5. Peak Extraction: For each successfully matched peak, a 150-point data segment centered on the peak's apex (i.e., ± ∼20 s) is extracted from the chromatogram (Figure 3.1). This data segment, along with its assigned compound label and known concentration, forms a single entry in the new dataset.

This process, illustrated in Figure 3.3, results in a new dataset of 17,513 labeled peaks. We

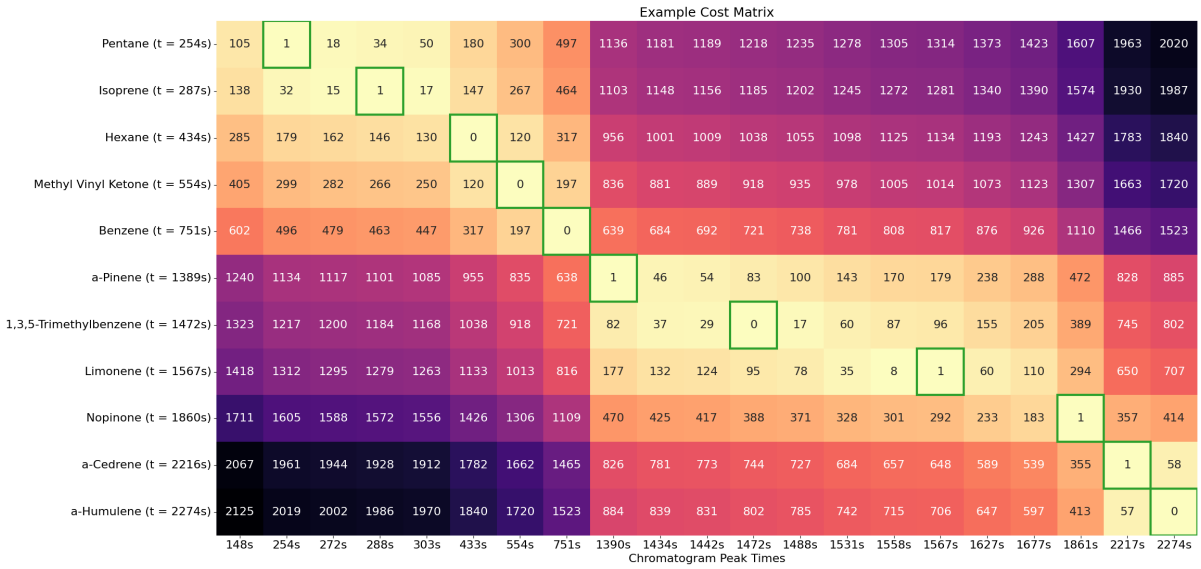

Figure 3.2: Example cost matrix from the peak segmentation process. The 11 known compounds with known retention times are on the left, while the peaks detected in the chromatogram are on the bottom. The values in the matrix represent the absolute difference between each known peak time and each detected peak time. The optimal assignments are outlined in green.

also randomly sample 2553 peaks from the calibration chromatograms that were detected but not assigned to a known compound during the matching step (presumed contaminants) and label them as "Unknown." These unknown peaks are included for compound prediction but not for concentration prediction, as their concentration is undefined. This brings the dataset to a total size of 20,066 peaks. For the compound classification model, each extracted peak is individually background subtracted and normalized to a minimum of 0 and a maximum of 1 to prevent the model from using differences in absolute peak height as a shortcut for classification. For a given peak with raw signal values x, the normalized values x_{norm} are computed as:

$$x_{\text{norm}} = \frac{x - \min(x)}{\max(x) - \min(x)}.$$

This min-max normalization preserves the relative shape of each peak while removing all information related to absolute signal amplitude, ensuring that the classification model learns

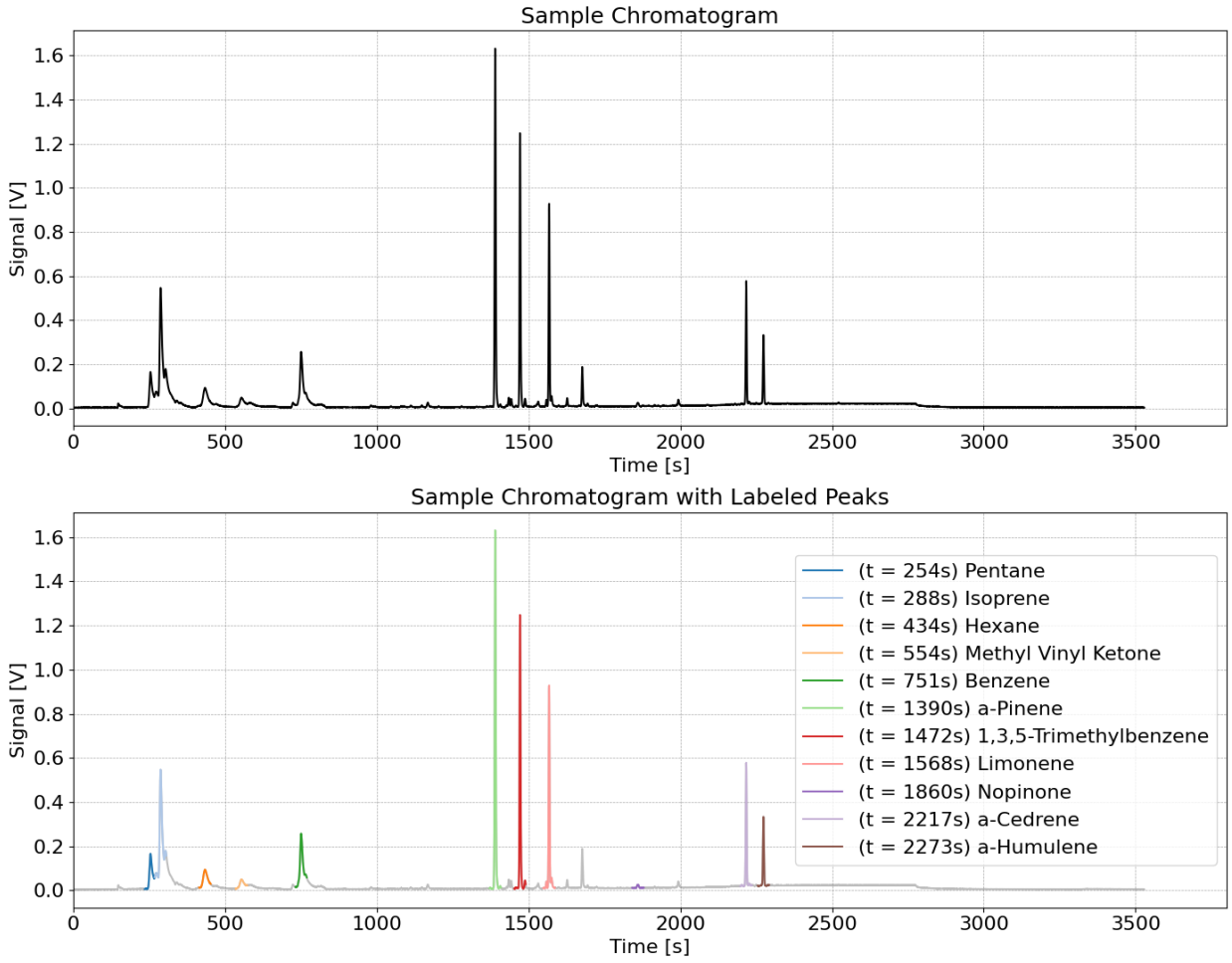

Figure 3.3: Example of the peak segmentation process on a sample chromatogram. The top panel shows the unlabeled chromatogram, while the bottom panel shows the chromatogram with peaks labeled according to the peak segmentation process.

exclusively from peak shape.

In contrast, the concentration prediction model uses the non-normalized peak segments to preserve amplitude information that is physically linked to compound concentration.

3.2.3 Removal of Retention Time Information

Retention times are used to generate the compound labels in the dataset, directly correlating each retention time with its ground-truth compound identity. Therefore, providing retention time as an input feature to the models would lead to a trivial one-to-one mapping between retention time and compound label, achieving artificially high performance without learning any meaningful peak shape characteristics. To instead evaluate the ability to predict compound identities based only on peak shape, all explicit timing information is removed from the input data. Each extracted peak is represented only by its 150-point signal segment, with the original time axis discarded and replaced by a uniform index (0-149).

3.2.4 Model Setup

The eXtreme Gradient Boosting (XGBoost) [6] classifier and regressor models are used for the tasks of compound classification and concentration prediction, respectively. The models take a 150-point peak segment as input and predict the compound and concentration, respectively. In addition to the peak segment, the models receive the collection date of the peak as an auxiliary input feature. Each chromatogram has an associated collection date, which is converted to an integer representing the number of days elapsed since the earliest date in the dataset. This value is stored with every peak extracted from that chromatogram and appended as a 151st feature in the model input. Inclusion of date information allows the model to learn and use changes in data quality over the long collection period ($\sim$5 years),

as slow degradation of the chromatographic column may change peak shapes.

An 80/20 train-test split and 10-fold cross-validation are used to avoid overfitting and ensure model reliability, as recommended for reliable model evaluation [44]. These validation strategies ensure that model performance is both accurate on the training set and generalizable to unseen data.

The Optuna library [1] is used to explore the hyperparameter space and optimize performance for both models. The best-performing hyperparameters for each model are summarized in Tables 3.2 and 3.3.

Table 3.2: Hyperparameters for the compound classification model

Parameter	**Value**
Number of Trees	742
Maximum Tree Depth	15
L1 Regularization (Alpha)	0.37
L2 Regularization (Lambda)	2.35
Column Subsample Ratio (per Tree)	0.58
Row Subsample Ratio	0.83
Learning Rate	0.025
Minimum Child Weight	5

Table 3.3: Hyperparameters for the concentration prediction model

Parameter	**Value**
Number of Trees	763
Maximum Tree Depth	16
L1 Regularization (Alpha)	2.76
L2 Regularization (Lambda)	3.02
Column Subsample Ratio (per Tree)	0.59
Row Subsample Ratio	0.78
Learning Rate	0.073
Minimum Child Weight	9

3.2.5 Evaluation Metrics

The compound classifier is evaluated using the following multiclass classification metrics:

- Accuracy: Proportion of total predictions that are correct.
- Precision: Proportion of predicted positive instances that are correct.
- Recall: Proportion of actual positive instances that are correctly identified.
- F1 Score: Harmonic mean of precision and recall, providing a balance between the two.

The concentration prediction model is evaluated using the following regression metrics:

- Mean Squared Error (MSE): Average squared difference between predicted and actual values. Lower is better.
- Mean Absolute Error (MAE): Average absolute difference between predicted and actual values. Lower is better.
- R^2 Score: Proportion of variance in the predictions that is explained by the actual values. A score of 1.00 indicates a perfect fit.

All metrics are reported overall, with accuracy and R^2 score additionally provided on a per-compound basis to highlight variations in model performance across different compounds.

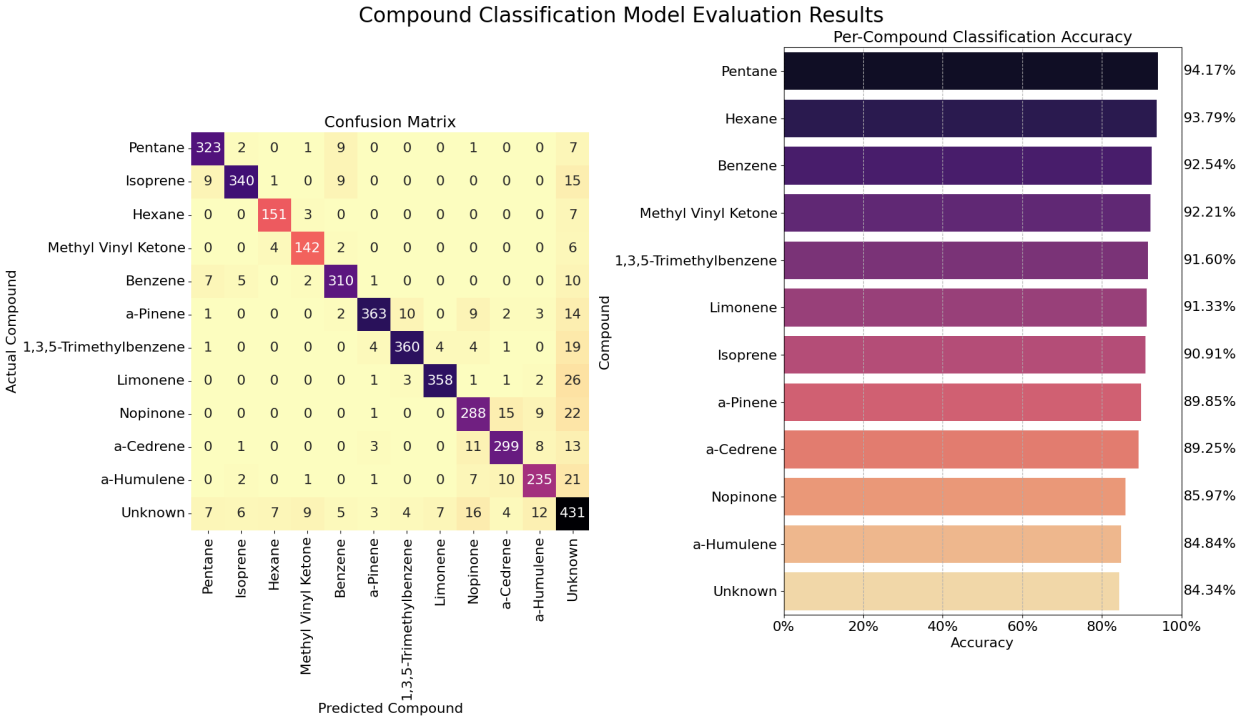

Figure 3.4: Evaluation of the compound classification model on the test dataset. The left panel shows a confusion matrix of predicted (bottom edge) versus actual (left edge) compound labels. Each box counts the number of occurrences of its corresponding predicted and actual labels. The right pane shows classification accuracy broken down by compound.

Table 3.4: Compound Prediction Model Evaluation Metrics

Accuracy	Precision	Recall	F1 Score
0.8969	0.9015	0.8969	0.8983

3.3 Results and Discussion

3.3.1 Compound Prediction Model

Using just peak shape, the compound prediction model is able to label peaks with an accuracy of 0.89 (Table 3.4). The confusion matrix (Figure 3.4a) shows predictions across the 12 classes (11 known compounds and one "Unknown" label). Ideally, all predictions would lie on the diagonal, representing perfect classifications. The matrix shows strong clustering along the diagonal, indicating generally high accuracy, but there are some notable misclassifications.

Pentane and benzene, for example, are correctly identified 323 and 310 times, respectively, but are confused with each other a total of 18 times. Among the known compounds, nopinone, α-cedrene, and α-humulene are the most commonly confused, likely due to their lower concentrations in the dataset and longer retention times, which can reduce signal quality, negatively affecting model performance, and increase peak broadening leading to similar peak shapes.

Performance on the Unknown category is the lowest in the dataset, with more misclassifications than any individual known compound (Figure 3.4b). Both directions of confusion are prevalent: unknown peaks are more likely to be mistaken for known compounds, and known compounds are more likely to be incorrectly predicted as unknown. This behavior is consistent with the construction of the Unknown class, which aggregates a wide variety of unmatched peaks that could include noise features or peaks from compounds present in field samples but absent from the calibration mixture. Because of this, it lacks a common

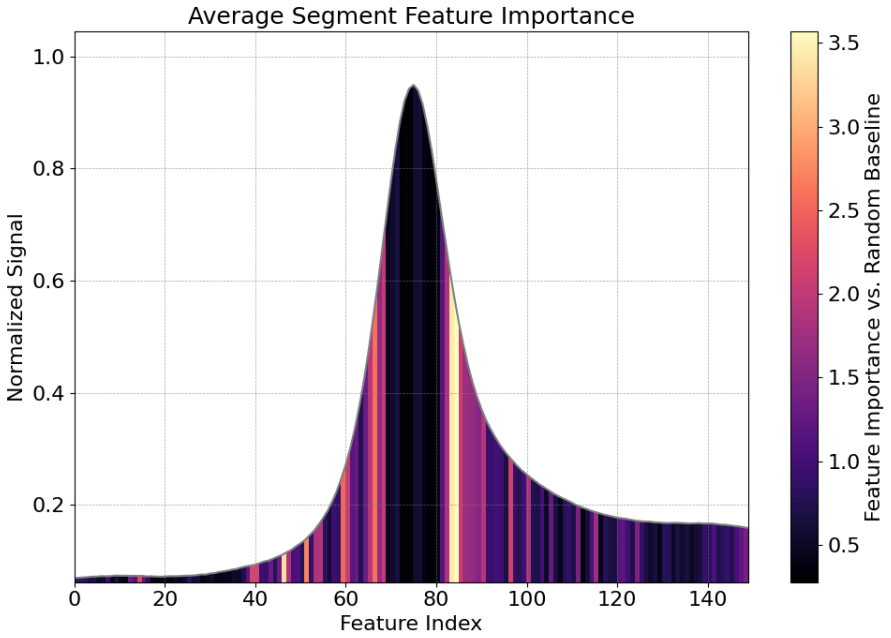

Figure 3.5: Compound Prediction model feature importance plotted on the average chromatogram peak segment. More important areas are shown in lighter colors. "Importance" refers to the XGBoost gain, which measures how much each input position improves the model's predictions when used in a decision-tree split. Higher-gain regions of the peak are relied on the most for distinguishing compounds. Here, the feature importance is reported as the ratio of each feature's importance to the expected importance of an untrained model (1/N), indicating how many times more informative that region is than chance.

peak-shape signature, making learning a coherent decision boundary for the unknown class more difficult, which leads to lower prediction accuracy.

Feature Importance

The feature importance analysis (Figure 3.5) shows which regions of an averaged 150-point peak segment are most informative for the compound classification model. The model assigns the highest importance to areas on the left and right regions of the peak, specifically point 46 on the left and points 83 and 84 on the right, which roughly bound one peak width (i.e., $\pm 1\ \sigma$). The apex region (points 70-80) and the baseline regions at the edges of the segment

(points 0-40 and 110-150) show relatively low importance.

This pattern indicates that the model is learning to distinguish compounds based on the rising and falling edges of a peak, where compound-specific fronting or tailing behavior could provide more discriminative information than the rest of the segment. The especially high importance at points on the tailing edge (83 and 84) suggests that the model has learned to rely especially on the decay characteristics of the peak. Chromatographic peaks are often modeled as an exponentially modified Gaussian (EMG), so the reliance of the model on the tailing region suggests that the coefficient of the exponential in these peaks provides unique information about each peak. The low importance of the baseline regions is expected, as they contain mostly noise. The low importance of the peak apex can be explained by the fact that the normalization step explicitly removes absolute intensity information, limiting discernible differences between peak apexes, and the apex of all of the peaks is roughly Gaussian so likely does not provide unique identifying information. These results demonstrate that the peak parameters are not only useful for peak integration, as used in other chromatographic processing systems, but should also more generally be considered in peak identification and quality control.

3.3.2 Concentration Prediction Model

Table 3.5: Concentration Prediction Model Evaluation Metrics

Mean Squared Error	**Mean Absolute Error**	R^2 **Score**
0.0323	0.0854	0.8241

The concentration prediction model achieves fair performance on the test dataset, with an MAE of 0.09 ppbv (Table 3.5) being larger than the smallest concentrations in the test dataset (Table 3.1), indicating that the model is unable to reliably predict low-concentration values. This is further demonstrated in Figure 3.6a, where lower-concentration predictions

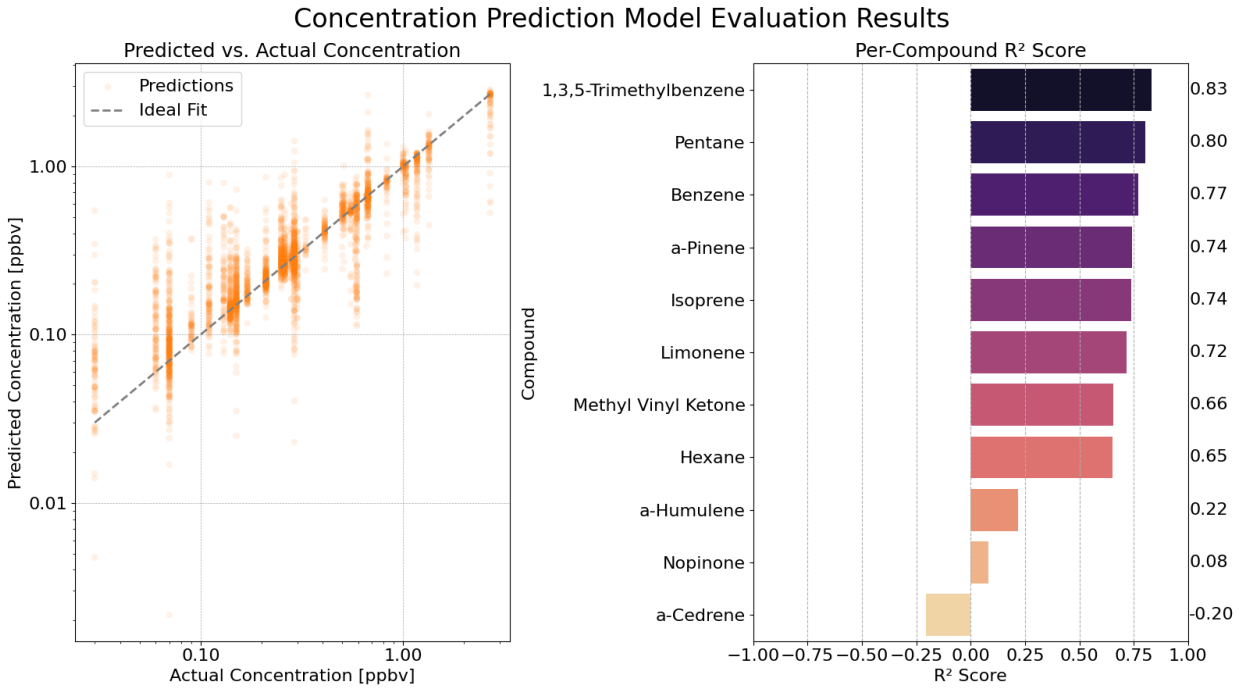

Figure 3.6: Evaluation of the concentration prediction model on the test dataset (with unknown peaks removed). The left panel shows a scatterplot of predicted versus actual concentrations, while the right panel shows R^2 scores broken down by compound.

are more spread out both above and below the ideal fit line.

Performance varies significantly across compounds, as shown in Figure 3.6. In particular, nopinone, α-cedrene, and α-humulene have low R^2 scores. The poor performance on nopinone is likely due to sampling artifacts introduced by the ozone filter at the inlet, which can also filter out oxygenates like nopinone. This likely results in reduced and inconsistent concentrations in the training data, affecting the model's ability to learn accurate predictive patterns. In contrast, the low scores for α-cedrene and α-humulene seem to suggest that the generalization patterns learned from the rest of the compounds do not translate well to α-cedrene and α-humulene predictions. Peak heights for these three compounds tend to be very low and sometimes absent due to their losses in the sampling inlet cause by their low volatility and the presence of the ozone filter. It is possible that the automated segmentation does not accurately assign the correct peak in all cases, which may lead in part to the poor calibration performance.

The performance of this model suggests that automated chromatographic processing with this approach is not yet able to replace current peak integration methods. However, there is some promise toward this end, with the compounds in the middle of the chromatogram—where data quality is most consistent—achieving moderate predictions.

3.4 Chapter Summary

This chapter demonstrates that machine learning models can extract meaningful chemical information directly from raw chromatographic peak shapes, even in the absence of spectral data, retention times, or manually engineered features. By training on a large, automatically labeled dataset derived from calibration chromatograms, the compound classifier achieves high accuracy using normalized peak shapes alone. Feature importance analysis reveals

that the model learns primarily from characteristics of the rising and falling edges of a peak, suggesting that these regions encode compound-specific shape signatures. The concentration prediction model similarly shows that explicit peak integration or calibration curve fitting are not required for concentration estimation.

Together, these results illustrate a viable path toward automated and scalable interpretation of complex GC data. Unlike traditional workflows that often rely heavily on manual peak annotation or derived features like integrated area, the presented approach operates directly on the earliest, most fundamental form of the data: raw signal segments. The method's reliance on automatically generated labels from calibration runs reduces the major bottleneck of manual human preprocessing, enabling the possibility of a fully automated end-to-end system that continuously interprets incoming data from field-deployed chromatographs to identify and quantify compounds with minimal intervention. Ultimately, this work provides a proof-of-concept that aims to advance the throughput and scalability of long-term atmospheric and environmental monitoring.

Chapter 4

Discussion

4.1 Summary of Findings

This thesis demonstrates that machine learning can augment the information extracted from gas chromatographic data, addressing two persistent challenges in environmental analysis: (i) the characterization of unknown compounds and (ii) the labor-intensive reduction of complex chromatograms.

Traditionally, the analysis of complex environmental mixtures has relied on reference libraries for compound identification and manual processes for quantification. The work presented here outlines two potential paths forward. First, we show that retention behavior on two orthogonal stationary phases serves as a structure-agnostic chromatographic fingerprint that enables the prediction of physicochemical properties (vapor pressure, solubility, and Henry's law constant) directly through machine learning models, bypassing the need for mass spectral identification for property estimation. Second, we show that the raw shape of a chromatographic peak contains latent, compound-specific information that can be exploited for automated compound classification and quantification without manual peak integration.

Together, these contributions show the potential for machine learning to enable more automated and chemically meaningful interpretation of chromatographic signals, even when individual constituents remain formally unidentified.

4.2 Synthesis

The methodologies developed in Chapters 2 and 3, although different in their inputs, outputs, and immediate objectives, are conceptually complementary. Both operate on the idea that machine learning can map latent signatures in GC data directly to chemically relevant endpoints.

Using both approaches within a single analytical workflow could be particularly valuable. In a system equipped with two orthogonal stationary phases (polar and nonpolar), the methodology from Chapter 2 could estimate the volatility and partitioning behavior of an unknown compound from its retention indices, providing initial information for fate modeling. Subsequently, once the compound is identified and added to a monitoring library, the approach in Chapter 3 could automate its future detection and quantification directly from raw chromatograms. Importantly, both approaches reduce dependence on expensive mass spectrometers, making sophisticated analysis more accessible. Collectively, these ideas point towards a future in which GC systems contain embedded ML models that provide real-time characterization and quantification directly from the chromatographic signal.

4.3 Limitations

While this work shows promising results, several limitations must be acknowledged.

First, all models are inherently constrained by the chemical space and instrumental conditions represented in their training data. This is particularly true for the peak-shape models in Chapter 3, which are only trained on 11 compounds and on data generated using a single GC column and a specific set of instrumental conditions (e.g., temperature program, carrier gas flow).

Similarly, generalizability across instruments is not guaranteed. Although the 2011 NIST mass spectral dataset [5] in Chapter 2 includes retention indices measured on a wide range of instrument types, the models may perform poorly on systems whose behavior deviates significantly from the overall trends represented in the training data. In such cases, system-specific retraining may be necessary.

Third, model interpretability remains limited, especially when compared with more traditional process-based models, where coefficients offer more direct insight. While the feature-importance analysis in Chapter 3 highlights the relevance of peak edges, it does not fully explain the underlying mechanisms driving this relationship.

Additionally, the accuracy of the concentration prediction model in Chapter 3 varies across compounds and concentration ranges. Performance differs among the 11 compounds considered and degrades at lower concentrations, where reduced signal-to-noise ratios lead to less reliable peak shapes. Given the limited chemical diversity considered, these performance differences may become more pronounced with more candidate compounds.

Lastly, the work focuses on "ideal" data conditions: Chapter 2 relies on precomputed retention indices, and Chapter 3 focuses on segmented, isolated peaks. The complexities of real-world samples (e.g., co-elution, peak distortion, etc.) remain unaddressed.

4.4 Future Work

Several directions for future advancements can be derived from these limitations. A potential first step would be to incorporate GC system parameters (e.g., temperature program, carrier gas flow, etc.) as explicit model inputs. This could improve cross-system generalizability by enabling models to learn the relationships between instrument configuration and resulting

chromatographic behavior.

Additionally, an exploration of multitask learning architectures for the physicochemical property prediction models (Chapter 2), in which one regression head predicts physicochemical properties while a parallel classification head assigns compound identity, could lead to an improvement in data efficiency and strengthen the shared representation learned from retention behavior.

Finally, the integration and evaluation of these approaches on real-world field instruments, as described in Section 4.2, would serve to test their effectiveness under operational conditions.

Bibliography

[1] Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. Optuna: A next-generation hyperparameter optimization framework. *CoRR*, abs/1907.10902, 2019. URL http://arxiv.org/abs/1907.10902.

[2] D.H. Bennett and Davis University of California. *Environmental Fate of Low Vapor Pressure-volatile Organic Compounds from Consumer Products: A Modeling Approach*. California Air Resources Board, Research Division, 2015. URL https://books.google.com/books?id=1I_v5BAzpSMC.

[3] Chenyang Bi and Gabriel Isaacman-VanWertz. Formation of late-generation atmospheric compounds inhibited by rapid deposition. *Nature Geoscience*, 18(3):213–218, Mar 2025. ISSN 1752-0908. doi: 10.1038/s41561-025-01650-2. URL https://doi.org/10.1038/s41561-025-01650-2.

[4] bmlerner. aerodyneresearch/tern: Tern 2.3.0a - autotern, May 2025. URL https://doi.org/10.5281/zenodo.15757836.

[5] NIST Mass Spectrometry Data Center and William E. Wallace. Retention indices. In Peter J. Linstrom and William G. Mallard, editors, *NIST Chemistry WebBook, NIST Standard Reference Database Number 69*. National Institute of Standards and Technology, Gaithersburg, MD, 20899, 2011. doi: 10.18434/T4D303. URL https://doi.org/10.18434/T4D303. Retrieved November 2, 2025.

[6] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. *CoRR*, abs/1603.02754, 2016. URL http://arxiv.org/abs/1603.02754.

[7] Stephen N. Chesler and Stuart P. Cram. Iterative curve fitting of chromatographic peaks. *Analytical Chemistry*, 45(8):1354–1359, 1973. doi: 10.1021/ac60330a031. URL https://doi.org/10.1021/ac60330a031.

[8] S Compernolle, K Ceulemans, and J-F Müller. EVAPORATION: a new vapour pressure estimation methodfor organic molecules including non-additivity and intramolecular interactions. *Atmos. Chem. Phys.*, 11(18):9431–9450, September 2011.

[9] S Compernolle, K Ceulemans, and J-F Müller. EVAPORATION: a new vapour pressure estimation methodfor organic molecules including non-additivity and intramolecular interactions. *Atmos. Chem. Phys.*, 11(18):9431–9450, September 2011.

[10] Elena David and Violeta-Carolina Niculescu. Volatile organic compounds (vocs) as environmental pollutants: Occurrence and mitigation using nanomaterials. *International Journal of Environmental Research and Public Health*, 18(24):13147, December 2021. ISSN 1660-4601. doi: 10.3390/ijerph182413147. URL http://dx.doi.org/10.3390/ijerph182413147.

[11] Clara M A Eichler, Elaine A Cohen Hubal, Ying Xu, Jianping Cao, Chenyang Bi, Charles J Weschler, Tunga Salthammer, Glenn C Morrison, Antti Joonas Koivisto, Yinping Zhang, Corinne Mandin, Wenjuan Wei, Patrice Blondeau, Dustin Poppendieck, Xiaoyu Liu, Christiaan J E Delmaar, Peter Fantke, Olivier Jolliet, Hyeong-Moo Shin, Miriam L Diamond, Manabu Shiraiwa, Andreas Zuend, Philip K Hopke, Natalie von Goetz, Markku Kulmala, and John C Little. Assessing human exposure to SVOCs in materials, products, and articles: A modular mechanistic framework. *Environ. Sci. Technol.*, 55(1):25–43, January 2021.

[12] Emily B Franklin, Lindsay D Yee, Bernard Aumont, Robert J Weber, Paul Grigas, and Allen H Goldstein. Ch3MS-RF: a random forest model for chemical character-

ization and improved quantification of unidentified atmospheric organics detected by chromatography–mass spectrometry techniques. *Atmos. Meas. Tech.*, 15(12):3779–3803, June 2022.

[13] Emily B Franklin, Lindsay D Yee, Bernard Aumont, Robert J Weber, Paul Grigas, and Allen H Goldstein. Ch3MS-RF: a random forest model for chemical characterization and improved quantification of unidentified atmospheric organics detected by chromatography–mass spectrometry techniques. *Atmos. Meas. Tech.*, 15(12):3779–3803, June 2022.

[14] Allen H. Goldstein and Ian E. Galbally. Known and unexplored organic constituents in the earth's atmosphere. *Environmental Science & Technology*, 41(5):1514–1521, March 2007. ISSN 0013-936X, 1520-5851. doi: 10.1021/es072476p. URL https://pubs.acs.org/doi/10.1021/es072476p.

[15] Jian Guo, Sam Shen, Shipei Xing, Ying Chen, Frank Chen, Elizabeth M. Porter, Huaxu Yu, and Tao Huan. Eva: Evaluation of metabolic feature fidelity using a deep learning model trained with over 25000 extracted ion chromatograms. *Analytical Chemistry*, 93 (36):12181–12186, 2021. doi: 10.1021/acs.analchem.1c01309. URL https://doi.org/10.1021/acs.analchem.1c01309. PMID: 34455775.

[16] A Hodzic, B Aumont, C Knote, J Lee-Taylor, S Madronich, and G Tyndall. Volatility dependence of henry's law constants of condensable organics: Application to estimate depositional loss of secondary organic aerosols. *Geophys. Res. Lett.*, 41(13):4795–4804, July 2014.

[17] Michal Hoskovec, Dana Grygarová, Josef Cvačka, Ludvík Streinz, Jiří Zima, Sergey P Verevkin, and Bohumír Koutek. Determining the vapour pressures of plant volatiles

from gas chromatographic retention data. *Journal of Chromatography A*, 1083(1):161–172, 2005. ISSN 0021-9673. doi: https://doi.org/10.1016/j.chroma.2005.06.006. URL https://www.sciencedirect.com/science/article/pii/S0021967305011775.

[18] Yuanyuan Hou, Shiyu Wang, Bing Bai, H C Stephen Chan, and Shuguang Yuan. Accurate physical property predictions via deep learning. *Molecules*, 27(5):1668, March 2022.

[19] Mia Hubert, Johan Van Kerckhoven, and Tim Verdonck. Robust parafac for incomplete data. *Journal of Chemometrics*, 26(6):290–298, 2012. doi: https://doi.org/10.1002/cem.2452. URL https://analyticalsciencejournals.onlinelibrary.wiley.com/doi/abs/10.1002/cem.2452.

[20] James F Hurley, Elizabeth Smiley, and Gabriel Isaacman-VanWertz. Modeled emission of hydroxyl and ozone reactivity from evaporation of fragrance mixtures. *Environ. Sci. Technol.*, 55(23):15672–15679, December 2021.

[21] G Isaacman, D R Worton, N M Kreisberg, C J Hennigan, A P Teng, S V Hering, A L Robinson, N M Donahue, and A H Goldstein. Understanding evolution of product composition and volatility distribution through in-situ GC × GC analysis: a case study of longifolene ozonolysis. *Atmos. Chem. Phys.*, 11(11):5335–5346, June 2011.

[22] G. Isaacman, D. R. Worton, N. M. Kreisberg, C. J. Hennigan, A. P. Teng, S. V. Hering, A. L. Robinson, N. M. Donahue, and A. H. Goldstein. Understanding evolution of product composition and volatility distribution through in-situ gc × gc analysis: a case study of longifolene ozonolysis. *Atmospheric Chemistry and Physics*, 11(11):5335–5346, 2011. doi: 10.5194/acp-11-5335-2011. URL https://acp.copernicus.org/articles/11/5335/2011/.

[23] Gabriel Isaacman-VanWertz and Bernard Aumont. Impact of organic molecular structure on the estimation of atmospherically relevant physicochemical parameters. *Atmos. Chem. Phys.*, 21(8):6541–6563, April 2021.

[24] Gabriel Isaacman-VanWertz and Bernard Aumont. Impact of organic molecular structure on the estimation of atmospherically relevant physicochemical parameters. *Atmos. Chem. Phys.*, 21(8):6541–6563, April 2021.

[25] Gabriel Isaacman-VanWertz, Lindsay D Yee, Nathan M Kreisberg, Rebecca Wernis, Joshua A Moss, Susanne V Hering, Suzane S de Sá, Scot T Martin, M Lizabeth Alexander, Brett B Palm, Weiwei Hu, Pedro Campuzano-Jost, Douglas A Day, Jose L Jimenez, Matthieu Riva, Jason D Surratt, Juarez Viegas, Antonio Manzi, Eric Edgerton, Karsten Baumann, Rodrigo Souza, Paulo Artaxo, and Allen H Goldstein. Ambient gas-particle partitioning of tracers for biogenic oxidation. *Environ. Sci. Technol.*, 50(18):9952–9962, September 2016.

[26] Gabriel Isaacman-VanWertz, Donna T. Sueper, Kenneth C. Aikin, Brian M. Lerner, Jessica B. Gilman, Joost A. de Gouw, Douglas R. Worsnop, and Allen H. Goldstein. Automated single-ion peak fitting as an efficient approach for analyzing complex chromatographic data. *Journal of Chromatography A*, 1529:81–92, December 2017. ISSN 0021-9673. doi: 10.1016/j.chroma.2017.11.005. URL http://dx.doi.org/10.1016/j.chroma.2017.11.005.

[27] William A Jury, Garrison Sposito, and Robert E White. A transfer function model of solute transport through soil: 1. fundamental concepts. *Water Resour. Res.*, 22(2): 243–247, February 1986.

[28] Akinde F. Kadjo, Hongzhu Liao, Purnendu K. Dasgupta, and Karsten G. Kraiczek. Width based characterization of chromatographic peaks: Beyond height and area. *An-*

alytical Chemistry, 89(7):3893–3900, 2017. doi: 10.1021/acs.analchem.6b04858. URL https://doi.org/10.1021/acs.analchem.6b04858. PMID: 28244321.

[29] Akinde F. Kadjo, Purnendu K. Dasgupta, and Kannan Srinivasan. Shape-based peak identity confirmation in liquid chromatography. *Analytical Chemistry*, 93(8):3848–3856, 2021. doi: 10.1021/acs.analchem.0c04432. URL https://doi.org/10.1021/acs.analchem.0c04432. PMID: 33600150.

[30] Edward D. Kantz, Saumya Tiwari, Jeramie D. Watrous, Susan Cheng, and Mohit Jain. Deep neural networks for classification of lc-ms spectral peaks. *Analytical Chemistry*, 91 (19):12407–12413, 2019. doi: 10.1021/acs.analchem.9b02983. URL https://doi.org/10.1021/acs.analchem.9b02983. PMID: 31483992.

[31] S. Kim, B. M. Lerner, D. T. Sueper, and G. Isaacman-VanWertz. Comprehensive detection of analytes in large chromatographic datasets by coupling factor analysis with a decision tree. *Atmospheric Measurement Techniques*, 15(17):5061–5075, 2022. doi: 10.5194/amt-15-5061-2022. URL https://amt.copernicus.org/articles/15/5061/2022/.

[32] Matteo Krüger, Tommaso Galeazzo, Ivan Eremets, Bertil Schmidt, Ulrich Pöschl, Manabu Shiraiwa, and Thomas Berkemeier. Improved vapor pressure predictions using group contribution-assisted graph convolutional neural networks (GC^2NN). *Geosci. Model Dev.*, 18(20):7357–7371, October 2025.

[33] Yang Li, Biqing Chen, Shuaifei Yang, Zhe Jiao, Meichen Zhang, Yanmei Yang, and Yanhui Gao. Advances in environmental pollutant detection techniques: Enhancing public health monitoring and risk assessment. *Environment International*, 197:109365, 2025. ISSN 0160-4120. doi: https://doi.org/10.1016/j.envint.2025.109365. URL https://www.sciencedirect.com/science/article/pii/S0160412025001163.

[34] Eva Matisová and Milena Dömötörová. Fast gas chromatography and its use in trace analysis. *Journal of Chromatography A*, 1000(1):199–221, 2003. ISSN 0021-9673. doi: https://doi.org/10.1016/S0021-9673(03)00310-8. URL https://www.sciencedirect.com/science/article/pii/S0021967303003108. A Century of Chromatography 1903-2003.

[35] Deborah F. McGlynn, Namrata Shanmukh Panji, Graham Frazier, Chenyang Bi, and Gabriel Isaacman-VanWertz. An autonomous remotely operated gas chromatograph for chemically resolved monitoring of atmospheric volatile organic compounds. *Environ. Sci.: Atmos.*, 3:387–398, 2023. doi: 10.1039/D2EA00079B. URL http://dx.doi.org/10.1039/D2EA00079B.

[36] Deborah F McGlynn, Namrata Shanmukh Panji, Graham Frazier, Chenyang Bi, and Gabriel Isaacman-VanWertz. An autonomous remotely operated gas chromatograph for chemically resolved monitoring of atmospheric volatile organic compounds. *Environ. Sci. Atmos.*, 3(2):387–398, February 2023.

[37] Arsenty D. Melnikov, Yuri P. Tsentalovich, and Vadim V. Yanshole. Deep learning for the precise peak detection in high-resolution lc–ms data. *Analytical Chemistry*, 92(1): 588–592, 2020. doi: 10.1021/acs.analchem.9b04811. URL https://doi.org/10.1021/acs.analchem.9b04811. PMID: 31841624.

[38] Sandra D. Mott and Eli Grushka. Chromatographic solute identification using peak shape analysis. *Journal of Chromatography A*, 126:191–204, 1976. ISSN 0021-9673. doi: https://doi.org/10.1016/S0021-9673(01)84072-3. URL https://www.sciencedirect.com/science/article/pii/S0021967301840723.

[39] Pentti Paatero and Unto Tapper. Positive matrix factorization: A non-negative factor model with optimal utilization of error estimates of data values. *Environ-*

metrics, 5(2):111–126, 1994. doi: https://doi.org/10.1002/env.3170050203. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/env.3170050203.

[40] Demetrios Pagonis, Derek J Price, Lucas B Algrim, Douglas A Day, Anne V Handschy, Harald Stark, Shelly L Miller, Joost de Gouw, Jose L Jimenez, and Paul J Ziemann. Time-resolved measurements of indoor chemical emissions, deposition, and reactions in a university art museum. *Environ. Sci. Technol.*, 53(9):4794–4802, May 2019.

[41] T Raventos-Duran, M Camredon, R Valorso, C Mouchel-Vallon, and B Aumont. Structure-activity relationships to estimate the effective henry's law constants of organics of atmospheric interest. *Atmos. Chem. Phys.*, 10(16):7643–7654, August 2010.

[42] T Raventos-Duran, M Camredon, R Valorso, C Mouchel-Vallon, and B Aumont. Structure-activity relationships to estimate the effective henry's law constants of organics of atmospheric interest. *Atmos. Chem. Phys.*, 10(16):7643–7654, August 2010.

[43] David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. Learning representations by back-propagating errors. *Nature*, 323(6088):533–536, October 1986.

[44] scikit-learn developers. Common pitfalls and recommended practices. URL https://scikit-learn.org/stable/common_pitfalls.html. Accessed November 2, 2025.

[45] Stephen E Stein, Valeri I Babushok, Robert L Brown, and Peter J Linstrom. Estimation of kováts retention indices using group contributions. *J. Chem. Inf. Model.*, 47(3):975–980, May 2007.

[46] Nipun Thamatam, Jeonghyeon Ahn, Mustahsin Chowdhury, Arjun Sharma, Poonam Gupta, Linsey C Marr, Leyla Nazhandali, and Masoud Agah. A MEMS-enabled portable gas chromatography injection system for trace analysis. *Anal. Chim. Acta*, 1261(341209):341209, June 2023.

[47] Kyaw Thet and Nancy Woo. Gas chromatography, Aug 2023. URL https://chem.libretexts.org/Bookshelves/Analytical_Chemistry/Supplemental_Modules_(Analytical_Chemistry)/Instrumentation_and_Analysis/Chromatography/Gas_Chromatography.

[48] Eva Tyteca, Mohammad Talebi, Ruth Amos, Soo Hyun Park, Maryam Taraji, Yabin Wen, Roman Szucs, Christopher A Pohl, John W Dolan, and Paul R Haddad. Towards a chromatographic similarity index to establish localized quantitative structure-retention models for retention prediction: Use of retention factor ratio. *J. Chromatogr. A*, 1486: 50–58, February 2017.

[49] U.S. Environmental Protection Agency. *Estimation Programs Interface Suite™ for Microsoft® Windows, Version 4.11*. United States Environmental Protection Agency, Washington, DC, USA, 2019. Available from: https://www.epa.gov/tsca-screening-tools/epi-suitetm-estimation-program-interface.

[50] Tomáš Vrzal, Michaela Malečková, and Jana Olšovská. DeepReI: Deep learning-based gas chromatographic retention index predictor. *Anal. Chim. Acta*, 1147:64–71, February 2021.

[51] Fei Wang, Jaanus Liigand, Siyang Tian, David Arndt, Russell Greiner, and David S Wishart. CFM-ID 4.0: More accurate ESI-MS/MS spectral prediction and compound identification. *Anal. Chem.*, 93(34):11692–11700, August 2021.

[52] David R. Worton, Monika Decker, Gabriel Isaacman-VanWertz, Arthur W. H. Chan, Kevin R. Wilson, and Allen H. Goldstein. Improved molecular level identification of organic compounds using comprehensive two-dimensional chromatography, dual ionization energies and high resolution mass spectrometry. *Analyst*, 142:2395–2403, 2017. doi: 10.1039/C7AN00625J. URL http://dx.doi.org/10.1039/C7AN00625J.

[53] Yunliang Zhao, Ngoc T. Nguyen, Albert A. Presto, Christopher J. Hennigan, Andrew A. May, and Allen L. Robinson. Intermediate volatility organic compound emissions from on-road gasoline vehicles and small off-road gasoline engines. *Environmental Science & Technology*, 50(8):4554–4563, 2016. doi: 10.1021/acs.est.5b06247. URL https://doi.org/10.1021/acs.est.5b06247. PMID: 27023443.

Appendices

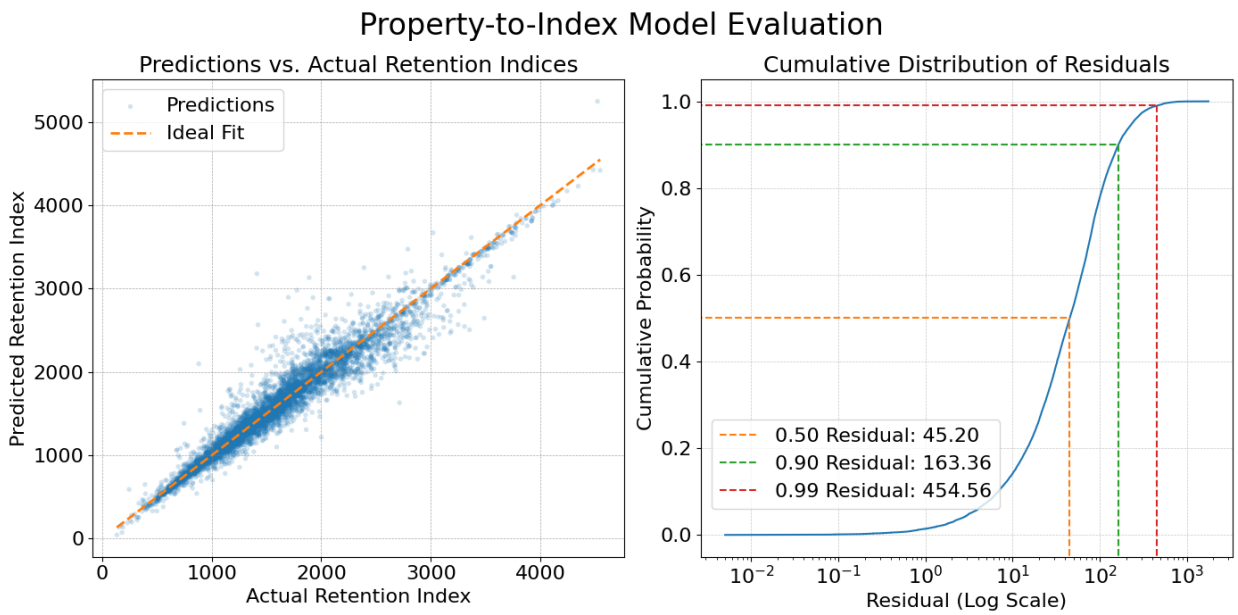

Figure 1: Evaluation of the MLP version of the Property-To-Index Model on the test dataset. The left pane is a scatterplot of predicted versus actual RIs, with a dashed identity line indicating the ideal fit. The right pane is a cumulative distribution of residuals with the x-axis on a log scale.

Table 1: MLP Property-To-Index Model Evaluation Metrics

Mean Squared Error	**Mean Absolute Error**	R^2 **Score**
13304	72.6	0.945