Applied Machine Learning QSPR Model of Aromatic Hydrocarbon Solute Transport in Sandstone with XGBoost

Kennly Weerasinghe

Submitted in partial Fulfillment of the Requirements for the degree of Master of

Arts in Geological and Environmental Sciences in

the Graduate Division of Queens College of the City University of New York

May 2022

Advised by:

Dr. William Blanford

School of Earth and Environmental Sciences, Queens College

Committee Members:

Dr. Adam Kapelner

Department of Mathematics, Queens College

Dr. Chuixiang Yi

School of Earth and Environmental Sciences, Queens College

Dedication

In memory of Sarge, Goober, and Astro. To G.E.H., for the inspiration and courage to continue challenging the unknown.

Acknowledgements

I would like to extend my appreciation and thanks to Dr. Adam Kapelner for his guidance throughout this project. I thank the tireless effort from Hubert Majewski, whose software engineering expertise enabled the completion of this project. My advisor Dr. William Blanford for enabling me the freedom to explore an interdisciplinary project. I thank my mother and my friends for their support.

Abstract

Miscible displacement (MD) tests and their results have commonly been employed to model solute transport in porous media using graphical representations of flow referred to as breakthrough curves (BTC). The advection dispersion equation (ADE), a parametric mathematical model, and its existing analytical solutions are implemented to solve the inverse problem to obtain the transport parameters to approximate system behavior and model BTCs. Given the recent development of high-performance machine learning (ML) algorithms, this work describes a nonparametric method to adapt a data set obtained from a series of MD tests conducted on a homologous series of hydrophobic organic chemicals, 17 mononuclear aromatic hydrocarbons (MAHs), to train a monotonic-in-time XGBoost model that predicts BTCs. The procedure involves preprocessing, feature selection using quantitative structure property relationships (QSPR), and pruning the initial space of features. These are employed to address excessive dimensionality by using least absolute shrinkage and selection operator (LASSO) and backwards deletion prior to developing the final model. Additional visualization and analysis tools including centered c-ICE plots, additivity lineup tests, and SHAP are implemented as an alternative means of evaluating the XGBoost-based QSPR model by enumerating the feature interactions and contributions toward the predictions. Finally, the QSPR model is assessed against simpler models that utilize one QSPR feature, time, and the injection concentration. The resulting models developed using XGBoost, including the QSPR model and the simpler models, are capable of robust solute transport predictions when evaluated using

root mean squared error (RMSE) and R^2 values. We believe our models can in the future predict MAH BTCs on average with R^2 = 0.95, our out-of-sample performance herein. The monotonically increasing prediction BTCs were found to closely approximate the true BTCs.

Contents

1. Introduction

Hydraulic Fracturing (HF) has rapidly expanded throughout the United States since the early 2000s due to innovations in hydrocarbon extraction technologies (Llewellyn et al., 2015; EPA, 2015) and is now the most common means for well development for crude oil and natural gas production (Perrin & Cook, n.d.). HF involves the use of large inputs of water combined with proppants and up to 50 chemicals chosen from a diverse list of over 1000 chemicals. This mixture is then horizontally injected into low permeability geologic units to facilitate the extraction of hydrocarbons (EPA, 2016; Labrecque & Blanford, 2021). Notable to this list of chemicals, is the presence of carcinogens that are regulated under the Safe Drinking Water Act, including BTEX mononuclear aromatic hydrocarbons (MAHs) (EPA, 2015). The Marcellus formation lies approximately 1200 m below the Berea formation in south central Pennsylvania and steadily rises westward where it is 450 m below the Berea formation in eastern Ohio (Ryder et al., 2012). Production of hydrocarbons shifts from dry-gas to wet-gas and then to petroleum from east to west, representing a shift in hydrocarbon maturity, from mature to less mature (Kirschbaum et al., 2012). The less mature wet-gas and petroleum extraction are of greater concern to groundwater resources due to their proximity and near surface origin. Aside from its presence in the injecting fluid, the aforementioned BTEX compounds are found as components within petroleum distillate and are naturally occurring in the Marcellus Shale rock (EPA, 2015). BTEX members have also been detected in the flowback water and water released by mu-

nicipal wastewater treatment systems in Pennsylvania due to incomplete treatment of HF wastewater processed at the facilities (Ferrar et al., 2013). Given these risks, of particular interest is the means to develop methods and models for prediction and risk assessment of the fate and transport of hydrophobic organic chemicals that are present in the injecting fluid used in HF (Yu et al., 2018; Labrecque & Blanford, 2021).

One current method of modeling the flow of solutes is by conducting miscible displacement (MD) tracer tests to assess the processes governing the fate and transport of the solutes of interest in saturated porous media (Hu & Brusseau, 1995; Zimmerman et al., 2002; Labrecque & Blanford, 2021). These tests involve establishing steady-state flow, temperature and chemical equilibrium with a background aqueous ionic-solution in the porous media. Without altering the flow conditions, the tracer of interest is dissolved in the aqueous ionic-solution and passed throught the system. The concentration of the tracers at the effluent or at points of interest within the system is then monitored as a function of time (Schwarzenbach & Westall, 1981; Hu & Brusseau, 1995; Zimmerman et al., 2002; Limousin et al., 2007; Arenas et al., 2018; Labrecque & Blanford, 2021). Processes that influence transport behavior include advection and dispersion and depending upon the nature of the solute and porous media may also include sorption, degradation, and production (Grathwohl, 2012). Once the results of the tracer tests are obtained, the data is then integrated into a parametric model, the advection dispersion equation (ADE) (Van Genuchten, 1982; Jaiswal et al., 2011). While the ADE-based approach has its practical uses in explaining the transport of solutes previously examined in tracer tests, it is limited in its ability to simulate only the solutes where MD tests have been conducted due to the

need to backfit the transport parameters (Van Genuchten, 1982). Absent the transport parameters, the ADE model is constrained in its ability to model or predict the transport of solutes.

Quantitative structure-activity relationships (QSAR), describes a modeling technique that involves utilizing physicochemical properties of a target chemical, and connecting them to the activity of the chemical for the purpose of inference (Dearden, 2002; Asirvatham et al., 2019). The rationale of QSAR modeling is to incorporate the knowledge that biological and chemical activity of a compound in a system of interest is attributable to physicochemical descriptors (Mackay et al., 2006; Mamy et al., 2015; Asirvatham et al., 2019). Once the structure-activity relationship has been established, a model can be used to predict the activity of previously untested compounds that share similar structural descriptors (i.e. homologous chemicals). Analogous compounds tend to possess similar physicochemical attributes and any variance exhibited within a series of homologous compounds can be attributable to distinguishing molecular properties (Mackay et al., 2006; Mamy et al., 2015). QSAR as applied to physicochemical descriptor interactions is referred to as quantitative structure-property relationship (QSPR) modeling, since the chemical analysis is not concerned with in vivo biological interactions (Mackay et al., 2006; Yousefinejad & Hemmateenejad, 2015).

While acknowledging the limitations of the existing parametric ADE model (the need to obtain suitable transport parameter value) and with the goal of incorporating QSPR descriptors; we propose an alternative nonparametric machine learning (ML) model in order to overcome the stated hurdle of predicting the flow of hydrophobic organic chemicals. ML

refers to a range of available algorithms or ”machines” that learn from the available data with the aim of constructing a model that represents the phenomenon of interest (Bhavsar et al., 2017). Nonparametric models are distinguishable from parametric models in that the input parameters or features are not determined a priori, rather they are a result of the algorithm learning from the available training data resulting in a more flexible model (Breiman, 2001; Salvador, 2016). The US Department of Energy (DOE) noted that in the era of big data, with the availability of high speed computing and data storage, there is a need to adapt existing algorithms in the pursuit of solving simulations in high dimensional problems (Baker et al., 2019). ML models must be scalable, efficient, reduce the need for specialized expert knowledge for production of forward simulations, and provide robust solutions with high accuracy (Baker et al., 2019). ML models, although known for their performance and adaptability are viewed as opaque tools by researchers and often referred to as “black boxes” (Breiman, 2001). This presents a need for the development of alternative methods of exploring, visualizing, validating, and interpreting these complex models (Baker et al., 2019). The modeling procedure will build off the preliminary work by Labrecque and Blanford (2021), which involved a series of miscible displacement tracer tests conducted on a conditioned core of Berea Sandstone with a series of seventeen distinct MAHs, including the BTEX compounds. The available results of the tracer tests will be used in tandem with a selection of open sourced QSPR descriptors of the solutes of interest to build a tree-based ML model for the purpose of simulation and prediction of solute breakthrough curves (BTC).

Tree-based ML models have been deployed for analysis in a variety of water quality

studies including estimation of nutrient concentrations, beach water, mapping groundwater contaminants, and surface water quality (Castrillo & García, 2020; Xu et al., 2020; Chen et al., 2020; Huang et al., 2021; Li et al., 2022). While there are a variety of tree-based ML algorithms to select from, extreme gradient boosting (XGBoost) was chosen due to its scalability, capability of enforcing monotonicity, and usability in scenarios with limited computing resources (Chen & Guestrin, 2016). XGBoost has been successfully deployed in real-world production pipelines for ad click rates, machine learning competitions such as those hosted on Kaggle (Chen & Guestrin, 2016). XGBoost models have been evaluated in environmental studies including, but not limited to, abiotic reduction of organic compounds, estimating gaseous pollutants, predicting E.coli levels in agricultural water, and removal of micropollutants (Gao et al., 2021; Hu et al., 2021; Jeong et al., 2021; Weller et al., 2021) .

The nonparametric XGBoost model will integrate a pool of QSPR based features, including features that are traditionally considered in the fate and transport of organic chemicals sourced from Mackay et al. (2006) and non-traditional features sourced from Ballabio et al. (2009). From the initial selection of traditional and non-traditional features, a combination of the least absolute shrinkage and selection operator (LASSO) and backwards deletion is used for final feature selection. We will demonstrate a multi-modal approach for explaining this model, which will be referred herein as the QSPR XGBoost model. The QSPR XGBoost model will then be evaluated against four simpler models built utilizing one QSPR feature; 2 selected from the traditional set of features and 2 from the non-traditional set of features. We hypothesize that the QSPR XGBoost model will be superior

in performance when assessed against the simpler models given the experimentally obtained dataset.

2. Methods

2.1 EXPERIMENTAL DATASET

The porous media used by Labrecque and Blanford (2021) was a highly laminated Berea Sandstone core supplied by Cleveland Quarries. They chose this core due to its uniformity and its wide use in simulating a range of oil and gas extraction techniques and because the unit serves as an important drinking water source lying above the hydrocaron shales being developed by HF in much of Pennsylvania and Ohio (Kareem et al., 2017; Labrecque & Blanford, 2021). The core had a right circular cylindrical shape and was cut with the bedding plane parallel to the axis. The permeability range as supplied by the quarry was 100-200 millidarcy. The dimensions of the core include: length of 76.48 mm, width of 51.41 mm, and dry mass of 342.58 g. For the experiments conducted, the core was housed in a hydrostatic core holder where the rock was maintained at fixed pressurized conditions.

The core was well conditioned prior to conducting any tests, by flushing 60 liters of tracer-free ionic solution in order to extract loose material within the core (Labrecque & Blanford, 2021). The ionic solution used to condition the core was prepared in an attempt to match the major ion concentrations within the Berea formation located near drinking water sources in Ohio (Labrecque & Blanford, 2021). The sandstone was fully saturated and the flow was conducted under steady-state temperature, flow, and pressure conditions.

Measurements per MD test were obtained using a UV detector to measure for UV absorbance at wavelengths in the range of 194 to 225 nm and recorded at rates between 0.2 to 20 Hz (Labrecque & Blanford, 2021). The input concentration per test varied due to the variation in solubility of each tracer. The objective was to target input concentrations of roughly 10% of aqueous solubility, but this mark was deviated from significantly for pentamethylbenzene due to its low solubility in order to obtain decent readings (Labrecque & Blanford, 2021). For additional details on methodology see Labrecque and Blanford (2021).

The data obtained from the MD tests by Labrecque and Blanford (2021) that are utilized in this work was a homologoues series of monoaromatic hydrocarbons. They speifically include: benzene, toluene, ethylbenzene, o-xylene, m-xylene, p-xylene, 1,2,3-trimethylbenzene, 1,2,4-trimethylbenzene, 1,3,5-trimethylbenzene, n-propylbenzene, isopropylbenzene, 1-ethyl-2-methylbenzene, isopropyl-4-methylbenzene, n-butylbenzene, tert-butylbenzene, 1,2,4,5-tetramethylbenzene, and pentamethylbenzene conducted at a velocity of roughly 0.27 cm/min. Tert-butylbenzene and pentamethylbenzene had replicate data available at the same velocity. The data set has a total of 19 distinct MD BTCs where 2 are replicates (tert-butylbenzene and pentamethylbenzene). These BTCs are considered individual observations and the available data contained the target response as tracer concentration in $\mu\mathrm{mol}$ solute/kg of water as a function of time in minutes. Each tracer study had between 10,000 - 50,000 readings. It should be noted that Labrecque and Blanford (2021) had data available for seven additional MD tests which were not used in this work including five bromide tests at various velocities, Toluene and 1,2,4-

trimethylbenzene at a velocity of 0.09 cm/min. The focus of this work is limited modeling MAH compounds and for this reason the bromide data was excluded from the ML models since bromide is not an MAH. Additionally, Toluene and 1,2,4-trimethylbenzene was conducted at 0.09 cm/min which is not the same experimental condition as the other 19 MD tests. Although it is widely understood that properties of the porous media exert considerable influence on solute transport, including properties such as porosity, organic carbon content, and cation exchange capacity; the training data did not include features of the rock core since the experiments were conducted using the same core sample rendering these properties as constants for each MD test conducted (Piwoni & Keeley, 1990; Labrecque & Blanford, 2021). Similarly, properties of the solution, temperature, and flow were all constant (Labrecque & Blanford, 2021).

Table 1
Injection Concentration per MD tracer test

Tracer	C_o (µmol/kg water)
Benzene	2273.17
Toluene	582.44
Ethylbenzene	180.99
o-Xylene	211.03
m-Xylene	152.99
p-Xylene	201.02
1,2,3-Trimethylbenzene	173.18
1,2,4-Trimethylbenzene	51.74
1,3,5-Trimethylbenzene	209.52
n-Propylbenzene	45.68
Isopropylbenzene	206.92
1-Ethyl-2-methylbenzene	176.16
Isopropyl-4-methylbenzene	128.06
n-Butylbenzene	64.9
tert-Butylbenzene	110.14
1,2,4,5-Tetramethylbenzene	13.39
Pentamethylbenzene	84.27

Table 1 : Initial injection concentration per tracer

2.2 MODELING STEPS

2.2.1 Data Preprocessing

BTCs are graphical representations of the effluent temporal concentration data obtained from tracer tests and can be used to compare solute fate and transport in various media and amongst different tracers (Fetter et al., 2018). In order to build and train the model to generate BTC predictions we took the initial dataset, which varied between 10,000 - 50,000 readings per tracer, and standardized them to 4000 measurements per tracer. Standardization was achieved by using a moving average of concentration to 10 second intervals and then locating the nearest times when this averaged concentration superseded in increments of 1/4000 of the 0 to C_o range, similar to the method employed in the previous research for the ADE model (Labrecque & Blanford, 2021). The response variable, the tracer concentration, was converted to a relative fractional value between 0 to 1 by normalizing each observed effluent concentration reading (C_i) by the injected concentration (C_o). Each of the 4000 concentration values per tracer had corresponding time readings, which were converted to minutes (min). The MD test of the lowest molecular weight MAH, benzene, achieved full breakthrough (i.e., C_i = C_o) at 107.97 min for the 4000th data point, whereas the highest molecular weight MAH, pentamethylbenzene achieved full breakthrough at 493.21 min. The other 17 MAH MD tests had data that achieved full breakthrough between these time values. The data was then compiled into one .csv file where each set of 4000 rows of data was labeled with the tracer name ordered from smallest tracer to the largest tracer for a total of 76,000 rows. We used Rstudio and Microsoft Excel for data processing and the construction and implementation

of XGBoost based ML models. The code developed is available on GitHub.

2.2.2 Logit Transformation

Solute transport represented as breakthrough concentrations are bounded between 0 and 1, where 1 would indicate an observed effluent concentration matching that of the input, which is commonly referred to as full breakthrough. In order to enforce the ML model's output space between 0 to 1, we employ a logit transformation of the tracer concentration values $\frac{C_i}{C_o}$ of the initial data set as follows:

$$y = \ln\left(\frac{\frac{C_i}{C_o}}{1 - \frac{C_i}{C_o}}\right) \tag{1}$$

The logit function is a commonly used data transformation technique in regression analysis that maps probabilities (0,1) over all real numbers $(-\infty, +\infty)$. Although the dataset is not a set of probabilities, the tracer concentration is mapped as a fraction of the input concentration, C_i, from (0,1) (Peck et al., 2013; Hartmann et al., 2016). The logit transformation enables the chosen ML algorithm greater distribution space from which to predict, while constraining the predictions within the constrained range (0,1). It further ensures that nonsensical predictions (e.g. netaive values or values greater than 1) are not generated. Transforming or scaling data as a preprocessing step in model construction has been found to improve predictive performance for tree-based algorithms (Ahsan et al., 2021).

2.2.3 Leave One Out Cross Validation

The technique used throughout this work for model validation and generating out-of-sample predictions is leave-one-out cross validation (LOOCV) (Hastie et al., 2009). Although the available data from Labrecque and Blanford (2021) contained 76,000 rows, the true sample size is $n = 19$ distinct MD test observations. Given this relatively limited sample size and the need to maximize the available data for training, we opted for LOOCV rather than the customary k-fold train-test split (Efron & Gong, 1983).

2.2.4 XGBoost and Hyperparameter tuning

XGBoost is a tree-based ensemble algorithm that makes use of weaker learners for improved model accuracy and enables the user a wide range of hyperparameters that can be tuned based on the training data and the predictive target (Chen & Guestrin, 2016; Li et al., 2022). The set of hyperparameters can be delineated into two categories for tuning: (1) hyperparameters assigned manually by the practitioner and (2) hyperparameters that are tuned on the training data. The manually assigned hyperparameters include: `booster`, `monotone_constraints`, `tree_method`, `objective`, `eval_metric`, and `predictor`. The `booster` assigns the learner used in the algorithm to either linear or tree-based functions. The tree-based `booster` settings available include `gbtree` and `dart`, where both build gradient boosted trees. The `booster` used in this work was `dart`, which employs the dropout approach used in deep neural networks, to handle overfitting in tree-based ensemble (Vinayak & Gilad-Bachrach, 2015). The `monotone_constraints` hyperparameter was assigned to enforce monotonicity on the

target response $\frac{C_i}{C_o}$ since BTCs monotonically increase over time until full breakthrough is achieved. The `tree_method` was assigned to `exact`, which evaluates all features in the data for tree splitting. The `objective` was assigned to `reg:squarederror`, which is the mean squared error (MSE) loss function used for modeling problems that predict numerical values. The `eval_metric` was assigned to `RMSE`. The `predictor` was assigned to `cpu_predictor`, but `gpu_predictor` could be selected depending on the computational resources available. The rest of the hyperparameters were tuned using the MLR package in R, which has its functions capable of building and optimizing learning algorithms. The hyperparameters tuned using MLR are generated for each model constructed based on the training data.

2.2.5 Residuals

The dataset was imported into Rstudio as a .csv file and the $\frac{C_i}{C_o}$ values were logit transformed. The XGBoost CRAN package was loaded into Rstudio and the algorithm was trained on the data using LOOCV while simultaneously tuning the hyperparameters that were not manually set. Each model generated predictions for the target response $\frac{C_i}{C_o}$ with time and C_o as the features per tracer. LOOCV resulted in 19 models, each with predictions for 1 tracer obtained in logit form. To obtain the residuals the logit predictions were inverse transformed to convert the distribution of $\frac{C_i}{C_o}$ from $(-\infty, +\infty)$ to the original distribution space (0,1). The predictions were aggregated into the data table. The residuals were obtained by using the following equation:

$$e = y - \hat{y} \tag{2}$$

Table 2
Physicochemical feature values per tracer

Tracer	Physicochemical Features										
	Molecular Weight (g/mol)	Melting point (°C)	Boiling Point (°C)	Fugacity (@ 25°C)	Density ρ (g/cm^3)	Molar Volume MW/ρ (cm^3/mol)	Molar Volume Le Bas (cm^3/mol)	Vapor Pressure (P^L/Pa)	Ksp (mol/m^3)	Kow	Kh
Benzene	78.112	5.49	80.09	1	0.8765	89.12	96	12700	22.788	135	557
Toluene	92.139	-94.95	110.63	1	0.8668	106.3	118.2	3800	5.59	489.8	680
Ethylbenzene	106.165	-94.96	136.19	1	0.867	122.45	140.4	1270	1.431	1349	887
o-Xylene	106.165	-25.2	144.5	1	0.8802	120.61	140.4	1170	2.072	1412.5	565
m-Xylene	106.165	-47.8	139.12	1	0.8842	120.07	140.4	1100	1.507	1584.9	730
p-Xylene	106.165	13.25	138.37	1	0.8611	123.29	140.4	1170	2.024	1513.6	578
1,2,3-Trimethylbenzene	120.191	-25.4	176.12	1	0.8944	134.38	162.6	200	0.582	3548.1	343
1,2,4-Trimethylbenzene	120.191	-43.77	169.38	1	0.8758	137.24	162.6	270	0.474	3981.1	569
1,3,5-Trimethylbenzene	120.191	-44.72	164.74	1	0.88	136.58	162.6	325	0.416	3801.9	781
n-Propylbenzene	120.191	-99.6	159.24	1	0.862	139.43	162.6	450	0.433	4897.8	1040
Isopropylbenzene	120.191	-96.02	152.41	1	0.8618	139.47	162.6	610	0.416	4265.8	1466
1-Ethyl-2-methylbenzene	120.191	-79.83	165.2	1	0.8807	136.47	162.6	330	0.624	4265.8	529
Isopropyl-4-methylbenzene	134.218	-67.94	177.1	1	0.8573	156.56	184.8	204	0.253	12589.3	805
n-Butylbenzene	134.218	-87.85	183.31	1	0.8601	156.05	184.8	137	0.103	18197	1332
tert-Butylbenzene	134.218	-57.8	169.1	1	0.8665	154.9	184.8	286	0.224	12882.5	1280
1,2,4,5-Tetramethylbenzene	134.218	79.3	196.8	0.293	0.838	160.16	184.8	66	0.026	12589.3	2546
Pentamethylbenzene	148.245	54.5	232	0.514	0.917	161.66	207	9.52	0.105	36307.8	129.9

Table 2 : Physicochemical features per tracer

where $\hat{y}$ is the prediction, y is the observed $\frac{C_i}{C_o}$, and e is the residual. The residuals were then exported into a new data table to be used in the next step of QSPR feature selection and pruning. RMSE and R^2 values for the predictions were calculated.

2.2.6 QSPR features

The QSPR feature selection made use of two sources: Mackay et al. (2006) and moleDB (Ballabio et al., 2009). Mackay et al. (2006), contains physicochemical features that are commonly believed to interact with porous media in the aqueous phase and have been found to influence environmental fate and transport of hydrocarbons (Mamy et al., 2015). The physicochemical features chosen for this work are molecular weight (g/mol), melting point (°C), boiling point (°C), fugacity (@ 25°C), density (g/cm^3), molar volume (cm^3/mol) (as determined by molecular weight and density), Le Bas molar volume (cm^3/mol), vapor pressure (P^L/Pa), octanol-water partitioning coefficient (Kow), aqueous solubility (Ksp) (mol/m^3), and Henry's law coefficient (dimensionless) (Kh) (Table 2).

There may be benefits to predictive performance by mixing multiple feature classes

rather than to limit QSPR models to one class Tseng et al. (2012). We expanded the feature space by adding two more classes to the pool of features: geometric and constitutional features. Constitutional features are considered to contain information of molecular composition of the compounds of interest including the bond type and aromaticity (Mamy et al., 2015). Geometric features are considered to contain information regarding molecular size and shape (Mamy et al., 2015). Both classes of features contain spatial information of the chemicals of interest. The constitutional and geometric features were chosen from the moleDB database; values of features found in the database are calculated using the DRAGON software package (Ballabio et al., 2009). The constitutional features selected were referenced with the EPA's "Molecular descriptors Guide" (2008) and include the sum of atomic van der Waals volume (Sv), Kier-Hall electrotopological states (Ss), sum of atomic polarizability (Sp), sum of atomic Sanderson electronegativities (Se), sum of number of bonds (nBT), and number of non-Hydrogen bonds (nBO). The geometric features selected were the three-dimensional Wiener index (W3D), three dimensional Balaban index (J3D), spherosity (SPH), and asphericity (ASP). The EPA molecular descriptor guide, (2008) references the Wiener and Balaban Indices as features that refer to the relative connectivity within the chemical ("Molecular descriptors Guide", 2008). SPH and ASP, both features that reference the geometry of a compound (Ballabio et al., 2009).

The full QSPR space of features once enumerated is $p = 21$. Time and C_o, both non-QSPR features, when included result in p = 23. With n 19 and p = 23, the number of features is greater than the number of distinct observations which may result in an overfit model (Hawkins, 2004). Another consideration is the potential presence of collinearities

Table 3
Constitutional and Geometric feature values per tracer

Tracer	Constitutional Features						Geometric Features			
	Sv	Ss	Sp	Se	nBT	nBO	W3D	J3D	SPH	ASP
Benzene	7.79	12	8.28	11.65	12	6	183.214	2.545	1	0.25
Toluene	9.39	13.67	10.05	14.53	15	7	324.71	2.796	0.939	0.35
Ethylbenzene	10.99	15.17	11.81	17.42	18	8	510.046	3.056	0.788	0.403
o-Xylene	10.99	15.33	11.81	17.42	18	8	499.705	3.122	0.913	0.269
m-Xylene	10.99	15.33	11.81	17.42	18	8	523.51	2.985	0.922	0.319
p-Xylene	10.99	15.33	11.81	17.42	18	8	531.761	2.942	0.926	0.481
1,2,3-Trimethylbenzene	12.59	17	13.57	20.3	21	9	730.069	3.384	0.902	0.245
1,2,4-Trimethylbenzene	12.59	17	13.57	20.3	21	9	760.706	3.255	0.911	0.349
1,3,5-Trimethylbenzene	12.59	17	13.57	20.3	21	9	773.459	3.206	0.914	0.238
n-Propylbenzene	12.59	16.67	13.57	20.3	21	9	774.549	3.19	0.916	0.601
Isopropylbenzene	12.59	17	13.57	20.3	21	9	730.199	3.383	0.688	0.374
1-Ethyl-2-methylbenzene	12.59	16.83	13.57	20.3	21	9	721.709	3.42	0.789	0.261
Isopropyl-4-methylbenzene	14.18	18.67	15.33	23.19	24	10	1067.56	3.452	0.765	0.496
n-Butylbenzene	14.18	18.17	15.33	23.19	24	10	999.308	3.651	0.494	0.292
tert-Butylbenzene	14.18	18.92	15.33	23.19	24	10	966.197	3.816	0.633	0.309
1,2,4,5-Tetramethylbenzene	14.18	18.67	15.33	23.19	24	10	1050.92	3.507	0.907	0.328
Pentamethylbenzene	15.78	20.33	17.09	26.07	27	11	1367.03	3.83	0.901	0.255

Table 3 : Geometric and Constitutional features per tracer

amongst the chosen QSPR features, which is a common problem when modeling complex natural systems (Graham, 2003; Dearden, 2017). In order to evaluate potential collinearities, Pearson correlation analysis was conducted on the 21 features . The next step of feature selection will attempt to address both collinearities and the potential problem of overfitting by pruning the feature space by using least absolute shrinkage and selection operator (LASSO) and backwards elimination. It should be noted that the LASSO was first developed for geophysical applications (Santosa & Symes, 1986) and later rediscovered and refined by Tibshirani (1996).

2.2.7 Feature Selection using LASSO and Backwards Deletion

Osborne et al. (1998) advised using LASSO first in tandem with forward regression or backwards elimination. Zhao and Yu (2006) however state that LASSO may not be able to distinguish between one feature and other features given the threshold. Dearden (2017) stressed the need to handle collinearity in QSPR modeling to avoid statistical distortions. QSPR models developed with ML also run the risk of overfitting when the sample size of

available data to train the model is too small and the number of structural descriptors of compounds analyzed are too large (Ghasemi et al., 2018). Given these previous findings we decided to first use LASSO, followed by backwards elimination to obtain the final set of features to build the integrated QSPR ML model. A csv file containing the residuals and the initial pool of 21 QSPR features was imported into Rstudio. LASSO, a component of the net elastic algorithm that is available in the glmnet CRAN package, provides sparse solutions due to the L1 penalty (Friedman et al., 2010). The net elastic algorithm has two tuning parameters α and λ. To set the net elastic algorithm to run LASSO, α must be set to equal 1 and then cross validation is used to search the grid for the desired λ value by specifying the error metric argument to be minimized. Of the available error metrics in glmnet, mean squared error (MSE) was chosen for consistency with the error metrics used to evaluate performance. Cross validation was performed using the `cv.glmnet` function, which returns 100 λ values. Of the 100 λ values, one unique λ value is chosen based on the one standard error (1-SE) rule (Breiman et al., 1984; Friedman et al., 2010). The `glmnet` function, with the 1-SE λ value and α set to 1, is applied to the data to obtain coefficients. Features that did not return coefficients were dropped from the dataset. For the second step in parameter selection we used backwards deletion, which considers the entire set of remaining features and evaluates the statistical significance of the features as a group by generating a linear model (Graham, 2003; Hastie et al., 2009; Ruengvirayudh & Brooks, 2016; Chowdhury & Turin, 2020). Linear regression using the `lm` function was performed on the dataset with the remaining features. The `summary` function was used to analyze the results including the statistical significance based on the p-value and any detectable collinearities which would return N/A. Features found to be significant were

kept and the rest were dropped. The remaining features were then exported to be used in XGBoost.

2.2.8 Final Model

The conclude the modeling procedure, a csv file containing the logit transformed tracer concentration values, time, and the remaining QSPR features were imported as a data table. XGBoost, using LOOCV, was used to generate 19 predictions. The 19 sets of predictions were then evaluated for performance using RMSE and R^2. BTCs are then plotted using the ggplot CRAN package. The time readings were scaled to pore volumes (PV), a dimensionless time unit that can be calculated by multiplying time by the velocity and then dividing by the distance traveled expressed by equation 5.

$$PV = \frac{v_x t}{L} \tag{3}$$

Conversion of time to PV was done to follow the conventional means of reporting column studies (Fetter et al., 2018). The final model was then constructed using the entire dataset to enable visualization and analysis of the QSPR features by producing ICE plots, additivity lineup test, SHAP, and XGBoost importance plot.

2.2.9 Model Evaluation Error Metrics

The error metrics used to evaluate performance are RMSE and R^2. RMSE is a measure of the average prediction error per tracer and R^2 is a measure of the model fit.

$$RMSE = \sqrt{\frac{\Sigma_{i=1}^{n}(y_i - \hat{y})^2}{n}} \tag{4}$$

RMSE represents the error in terms of the units of the observation $\frac{C_i}{C_o}$ where y_i represents each individual reading of $\frac{C_i}{C_o}$, $\hat{y}$ is the models prediction for that corresponding reading, and n is the total number of readings. R^2 is used to evaluate the overall performance of the model per tracer study where $\bar{y}$ represents the mean of the readings for a given set of n.

$$R^2 = 1 - \frac{\Sigma_{i=1}^{n}(y_i - \hat{y})^2}{\Sigma_{i=1}^{n}(y_i - \bar{y})^2} \tag{5}$$

2.3 INTERPRETING OUR MODEL

Models produced with the XGBoost algorithm are always black boxes which are difficult to interpret due to the large degree of non-linearities among the features and interactions among the features. However, there are ways to understand partially how our model arrives at its predictions through a number of post-processing procedures explain below that produce metrics and visualizations. These post-processing procedures include: XGBoost importance plot, c-ICE-plots, additivity lineup test, SHAP, and a linear model to assess the space of features for statistical significance .

2.3.1 XGBoost Feature Importance

The XGBoost algorithm has a built-in feature importance function that ranks the features according to the contributed improvement in accuracy towards the model. The feature importance values are represented as fractions that sum to 1 (Chen et al., n.d.).

2.3.2 c-ICE plots

ICE plots, an extension of Friedman (2001)'s partial dependence plots (PDPs), is a tool introduced by Goldstein et al. (2015) that enables the visualization of the functional relationship between the analyzed feature and the ML predictions for individual observations. ICE plots enable the practitioner to analyze variability of the prediction across the range of the feature. PDP, defined by the following equation:

$$\hat{y}_s = \frac{\Sigma_{i=1}^{N} \hat{y}(x_s, x_{ci})}{N} \tag{6}$$

where x_s represents the feature of interest and is fixed, x_{ci} represents the set of the remaining features that vary over the entire dataset, and $\hat{y}_s$ represents the ML model average value as a function of x_s. This procedure margins out the effect of x_{ci} features and plots the average partial effect of x_s on $\hat{y}$ the prediction (Friedman, 2001; Goldstein et al., 2015). As an extension of PDPs, Goldstein et al. (2015) developed ICE plots, which disaggregates the average partial effect of x_s, enabling the visualization of any exhibited variance across the range of x_s values and its the partial effect on $\hat{y}$. Centering the ICE plots, referred to as c-ICE involves pinching the prediction lines at a location x^*, which prevents stacking of ICE curves.

2.3.3 Additivity Lineup Test

The additivity lineup test is a procedure used to assess the statistical validity of discoveries through the examination of plots (Buja et al., 2009; Wickham et al., 2010; Goldstein et al., 2015). The procedure involves randomly inserting the observed plot into a lineup of null plots, which are obtained by sampling the data under the null distribution to build the model on all features except the feature that is evaluated for statistical significance (Goldstein et al., 2015). This step is repeated $K-1$ times to generate null plots. K is the total number of plots in the lineup and is determined by the following:

$$\alpha = \frac{1}{K} \tag{7}$$

where alpha is preset and represents the threshold of significance (Goldstein et al., 2015). If the observed plot can be discerned from the lineup test, then the discovery is assigned

the p-value at the threshold of significance, signifying that the feature has an additive effect on the model and rejecting the null hypothesis (Goldstein et al., 2015). We set alpha equal to 0.05, which results in a lineup of $K = 20$ plots: 19 null and 1 observed. We then test the null hypothesis, Ho: the PDP of the QSPR feature of interest is not additive versus Ha: Ho is false, the PDP of the QSPR feature of interest is additive. The ICEbox CRAN package was used to generate c-ICE plots and additivity lineup tests. For additional details on the cICE, PDP, and additivity lineup procedure see Goldstein et al. (2015).

2.3.4 SHAP

SHapley Additive exPlanation (SHAP) is a unified game theoretical approach to interpreting ML models developed by Lundberg and Lee (2017) that quantifies the additivity of each feature using the training data. SHAP is able to enumerate values for both local and global contribution of each feature towards the predictions. The features used to train the model are treated as a coalition of players that contribute to the model output. The equation governing the SHAP approach is defined by the following:

$$\Phi_i(\hat{y}, x) = \sum_{z' \subseteq x'} \frac{|z'|!(M - |z'| - 1)!}{M!} \Big[\hat{y}_x(z') \ - \hat{y}_x(z' \backslash i)\Big] \tag{8}$$

where M is the number of all features, $\Phi_i(\hat{y}, x)$ is the SHAP value for feature i, x is the set of features, x' is the set of all possible feature combinations including feature i, z' is the feature combination as a subset of x', $|z'|$ is the number of non-zero features in z', $z' \backslash i$ removes feature i from z', and $\Big[\hat{y}_x(z') \ - \hat{y}_x(z' \backslash i)\Big]$ are separate predictive models

trained on z' and $z'\backslash i$. SHAP values are obtained by adding the marginal contributions of all possible combinations of z' features using a weighted average. SHAP values are calculated using the formula locally for each feature at each prediction. Given the dataset has 76,000 rows, 76,000 local SHAP values were obtained for each feature. The following equation relates the local SHAP values to the ML model:

$$\hat{y}(x) = g(x') = \Phi_o + \sum_{i=1}^{M} \Phi_i x'_i \tag{9}$$

where $\hat{y}(x)$ is the original model, $g(x')$ is the explanation model, Φ_o is the bias term, M is the number of all features, $z'\backslash i$ is feature i, Φ_i is the SHAP value for feature i. Equation 9, the summation of all the SHAP values per feature and the bias term equals the prediction output of the ML model for each row. Since the predictions are in the logit distribution, the output of the SHAP values are also in the space of the logit distribution. To convert the logit SHAP values, we first calculate the percentage contribution of each feature, including the bias term as a percent relative to the logit SHAP value for the local prediction. Then we take the inverse logit transformation of the prediction and then calculate the values per feature by multiplying the percent contribution to obtain the local SHAP values per feature. The global SHAP scores per feature for the model were calculated by averaging the absolute value of each local SHAP value. We calculated global SHAP values and generated force plots per each tracer to visualize the changing interactions amongst the features towards the prediction of the model using the SHAPforxgboost CRAN package (Liu & Just, 2020).

2.3.5 Model Summary

We can use linear regression to find the best linear approximation of the QSPR XGBoost model, which usually explains a fair amount of the performance. A linear regression was thus performed on the predictions generated from the QSPR model against the set of features. A summary with the coefficients for each feature including the p-value for statistical significance was obtained for comparative analysis against the features of importance, SHAP, and additivity tests.

2.4 MODELS WITH 1 QSPR FEATURE

4 additional XGBoost based models were built, each using time, C_o, and one QSPR feature from the pool of features. The features chosen include molecular weight, Ksp, Ss, and W3D. RMSE and R^2 values were calculated for each tracer and evaluated against the performance of the QSPR model.

3. Results and Discussion

3.1 QSPR FEATURE SELECTION

The feature selection step involved the the use of the residuals, obtained via the LOOCV predictions per tracer with time and C_o as features. Since the values of time and C_o were readily available from the original dataset, the rationale was to evaluate a model built on these two features and obtain the residuals to use for QSPR feature selection. In the context of modeling, residuals provide all the information regarding the fit of the model from the available data (Box & Draper, 1987). The initial Pearson correlation analysis (figure 1) shows significant correlations amongst the features, which must be addressed. The residuals when used in tandem with the LASSO procedure, with $\lambda = 1.966111e-05$, selected using the 1-SE rule, resulted in molecular weight, molar volume, Le Bas molar volume, Se, and nBO returning no coefficients (Appendix A: Table A1.1) and were subsequently dropped from the initial feature pool. The backwards deletion step found the remaining features significant with the exception of Sv, which was also dropped as a feature (Appendix A: Table A1.2). The remaining features included: melting point, boiling point, fugacity, density, vapor pressure, Ksp, Kow, Kh, Sp, Ss, nBT, W3D, J3D, SPH, and ASP. Including time and C_o, the total space of features p equals 17. With our sample size n = 19 and p = 17, the goal of reducing p such that $n > p$ was accomplished.

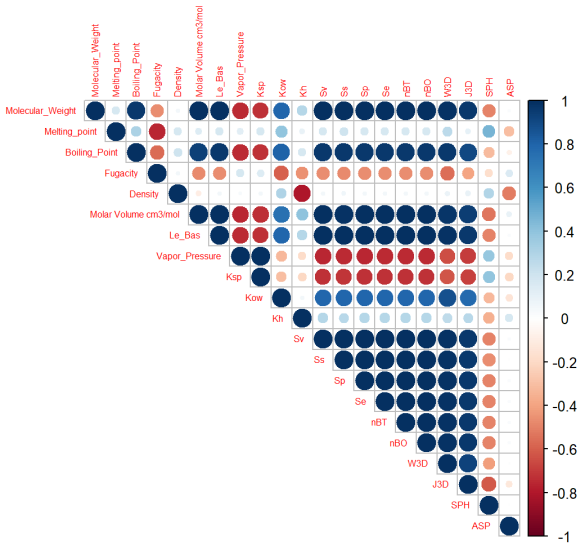
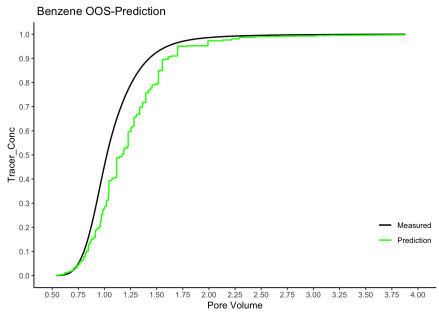


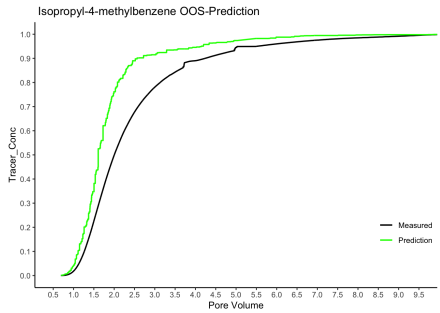
Figure 1 : Pearson correlations values amongst the initial QSPR features

3.2 QSPR XGBOOST MODEL PERFORMANCE

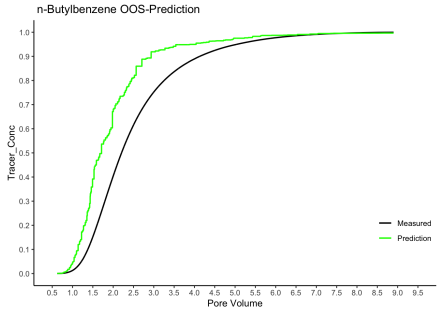
The QSPR XGBoost model fit out of sample (oos) resulted in an average of R^2 = 0.9548 (table 4). The oos average predicted RMSE = 0.0768 (table 4). RMSE describes the error in the units of the target response $\frac{C_i}{C_o}$. A rule of thumb for RMSE interpretation is within ± 2 RMSE's are approximately 95% of predictions. Thus, we are approximately 95% of the time predicting within $\pm 14\%$ BTC. Amongst the tracers, two outliers that resulted in both high RMSE values and lower R^2 relative to the rest of the tracers were n-Butylbenzene (RMSE = 0.1887 and R^2 = 0.5664) and 1,2,4,5-Tetramethylbenzene (RMSE = 0.1261 and R^2 = 0.8162). Benzene and Isopropyl-4-methylbenzene also exhibited RMSE values greater than the average at 0.0825 and 0.1887 respectively. An explanation for high performance R^2 values for benzene and Isopropyl-4-methylbenzene, yet low performance relative to RMSE, can be attributed to the differences in $\bar{y}$ for each tracer, since R^2 is a measure of model fit, and not necessarily reduction in RMSE. These differences are visible when the BTCs are plotted against the measured values from the MD tests (Figure 2). Of the remaining BTCs generated Ethylbenzene, m-Xylene, n-Propylbenzene, and Pentamethylbenzene plots display higher performance, with low RMSE values and high R^2 values (Figure 3). We now turn to understanding how our model performs so well using the procedures discuss in Section 2.3.



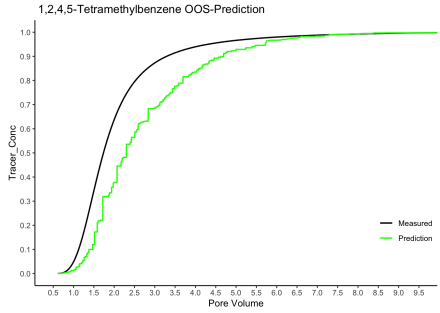
(a) Benzene BTC



(b) Isopropyl-4-methylbenzene BTC

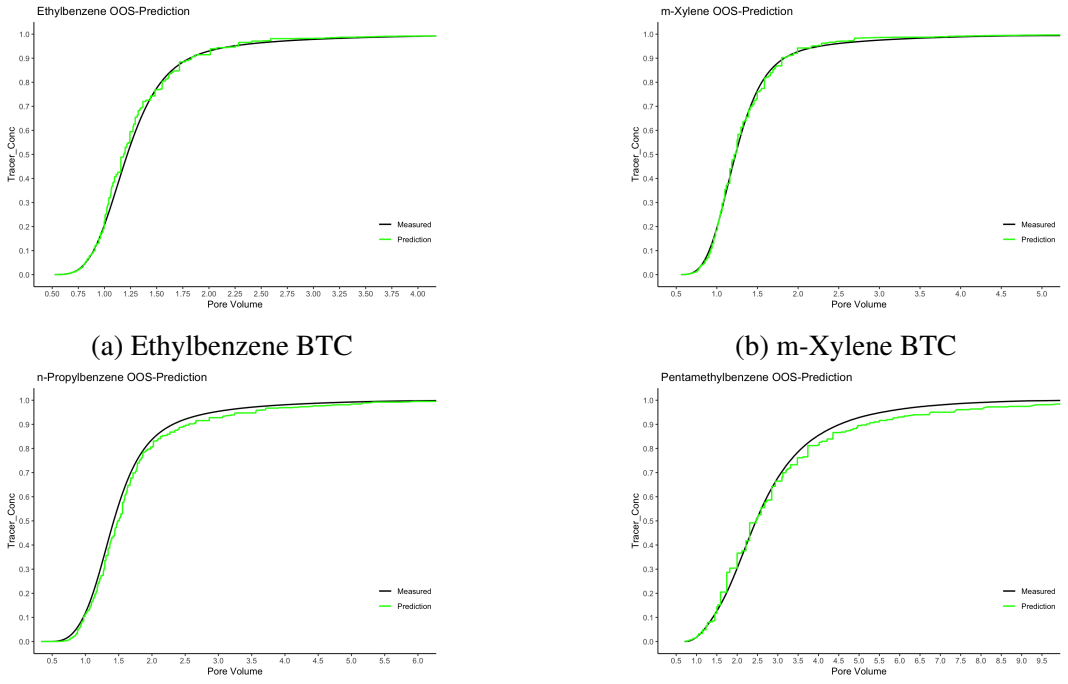


(c) n-butylbenzene BTC



(d) 1,2,4,5-Tetramethylbenzene BTC

Figure 2 : Measured (black lines) versus predicted (green lines) BTCs



(a) Ethylbenzene BTC

(b) m-Xylene BTC

(c) n-Propylbenzene BTC

(d) Pentamethylbenzene BTC

Figure 3 : Measured (black lines) versus out-of-sample predictions (green lines) BTC

3.3 EXPLAINING THE MODEL

3.3.1 XGBoost Feature Importance

The features that resulted in the highest gain or importance in contribution towards the predictions by the model in descending order were time, boiling point, J3D, Ksp, C_o, Kow, and W3D (Figure 4 and Appendix A:Table A2.2). The contributions of vapor pressure, Ss, melting point, density, SPH, ASP, and Kh were minimal according to the feature importance plot. The model did not use nBT, fugacity, or Sp towards any of the predictions. The linear model summary of the features on the prediction values found all features except for nBT, fugacity, and Sp to be significant (Appendix A: Table A2.1). The features found to not be significant by the linear model are in agreement with the features found to not be useful by XGBoost.

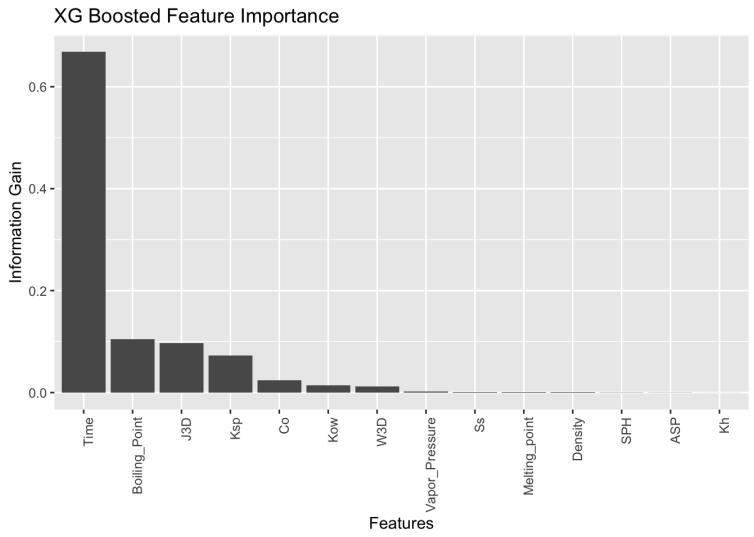


Figure 4 : Feature importance represented by the contribution gained by each feature for the QSPR model

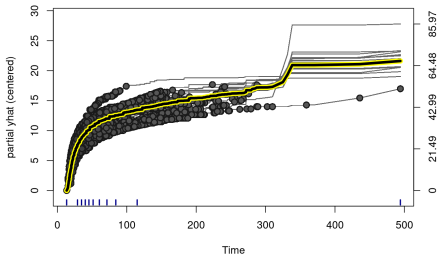
3.3.2 QSPR Model c-ICE

From the c-ICE plots generated (Figure 5 and Appendix B: Figure B2), disaggregating the PDP enable the visualization of the variability of the features impact on the prediction. The primary y-axis indicates the range of the predicted values per given value of the feature. The secondary y-axis displays the variance in the prediction over the baseline as a fraction of the observed range $(\frac{C_i}{C_o})$. Fugacity (Figure 5e) and Vapor Pressure (Figure 5g) both exhibit PDP lines starting at or near 0 on the primary y-axis, which would inaccurately lead one to believe that both features impact on the model are similar. The c-ICE curves clearly show that fugacity has 0 impact on the model due to the absence of c-ICE curves which is in agreement with the results of both XGBoost-importance metric gain and the SHAP procedure. In contrast, Vapor Pressure shows numerous predictions that contribute to the overall model output once the PDP has been disaggregated through the c-ICE procedure. The cumulative range however is 1.86, which in the scope of the entire

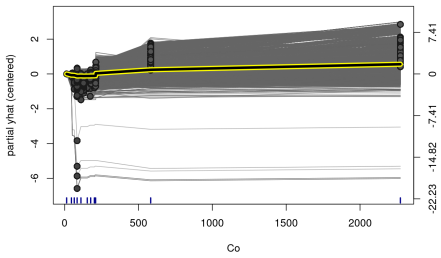
model is small. The c-ICE curve for time (Figure 5a) indicates significant contribution towards the model output when examining the secondary y-axis. The interaction between time and the other features result in cumulative differences in the fitted values of roughly 86% of the range of the prediction. C_o (Figure 5b) has the second highest cumulative differences for the fitted values of the prediction with a range of roughly 30%, which is noteworthy because without c-ICE curves, it would not be possible with just the XGBoost gain values to see the wide range of interaction for the feature C_o.

3.3.3 QSPR Model Additivity Lineup Test

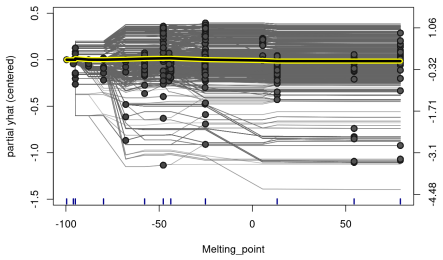
The additivity lineup test performed on the QSPR features nBT, Fugacity, and Sp were visually not significant which resulted in failing to reject the null hypothesis for these features (figure 9). Boiling point, J3D, Ksp, and W3D are the most visually significant of the QSPR features (figure 6). Kow, density, vapor pressure, and melting point subjectively are moderately significant since visually the true plot can be seen relatively easily (figure 7). Ss, SPH, Kh, and ASP are slightly challenging to discern from the null plots and thus can be characterized as exhibiting low visual statistical significance but sufficient enough to meet the lineup test threshold criteria (figure 8).



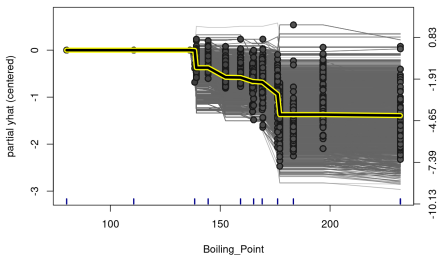
(a) Time



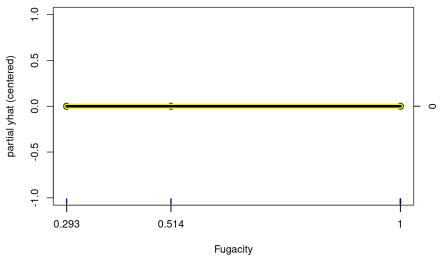
(b) C_o



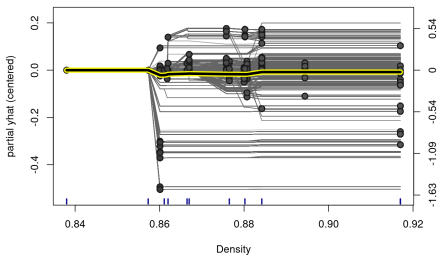
(c) Melting Point



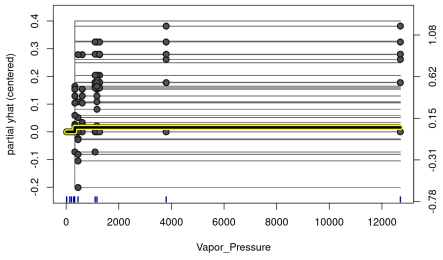
(d) Boiling Point



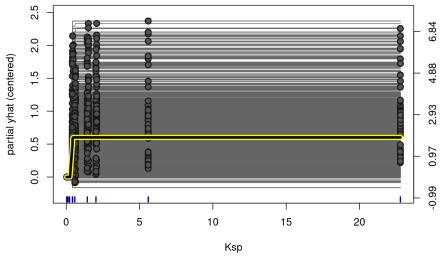
(e) Fugacity



(f) Density

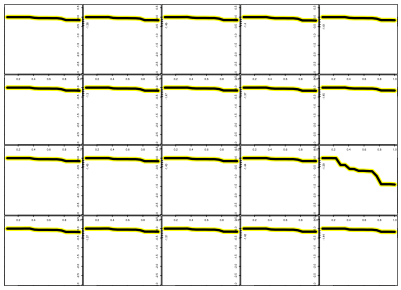


(g) Vapor Pressure

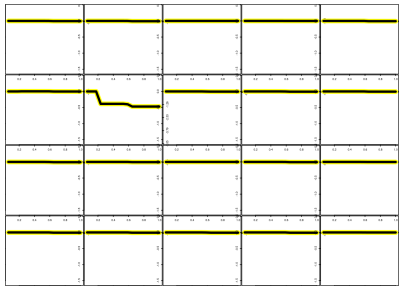


(h) Ksp

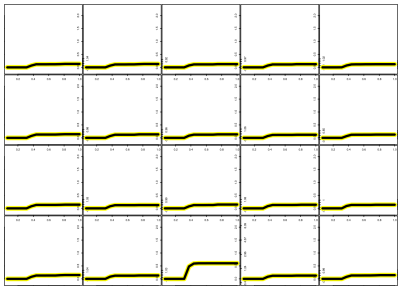
Figure 5 : c-ICE curves for (a) Time, (b) C_o, (c) Melting Point, (d) Boiling Point, (e) Fugacity, (f) Density, (g) Vapor Pressure, and (h) Ksp. Bold black line outlined in yellow represents the PDP plot, the average of the the ICE curves



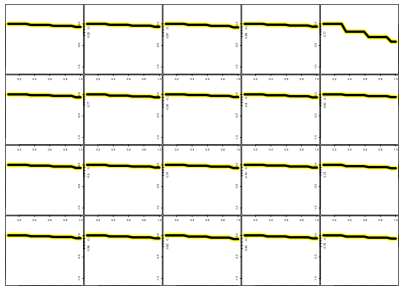
(a) Boiling Point



(b) J3D

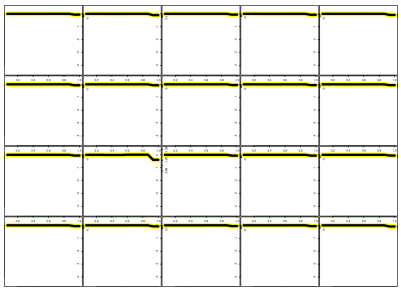


(c) Ksp

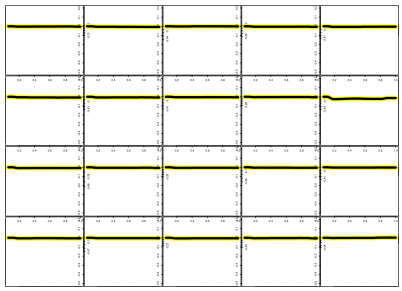


(d) W3D

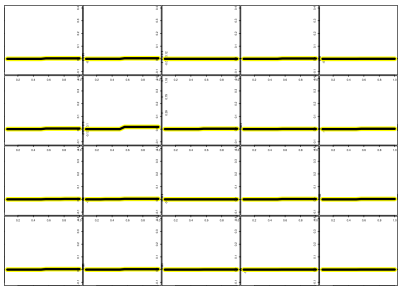
Figure 6 : Additivity Lineup Test - Highest visually significant features



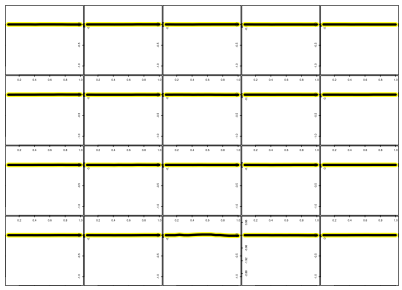
(a) Kow



(b) Density

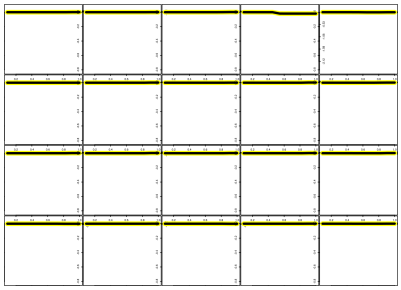


(c) Vapor Pressure

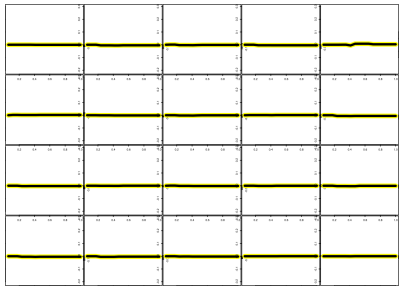


(d) Melting Point

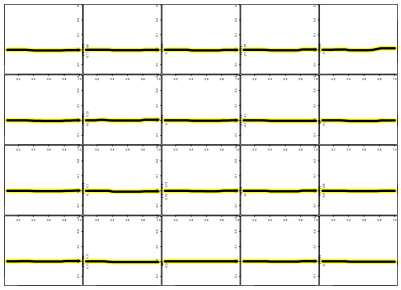
Figure 7 : Additivity Lineup Test - Moderate visually significant features



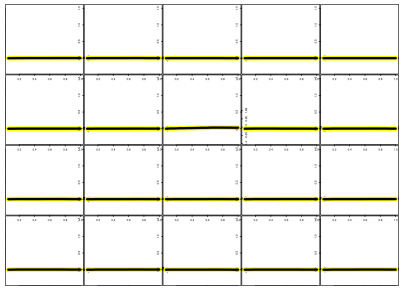
(a) Ss



(b) SPH

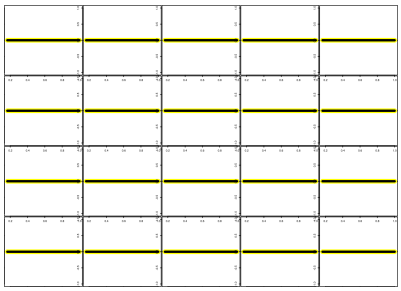


(c) Kh

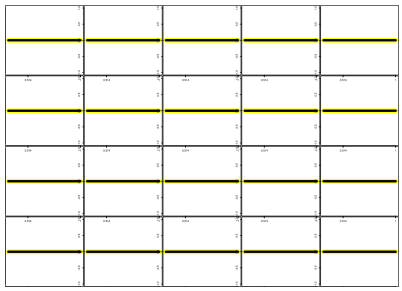


(d) ASP

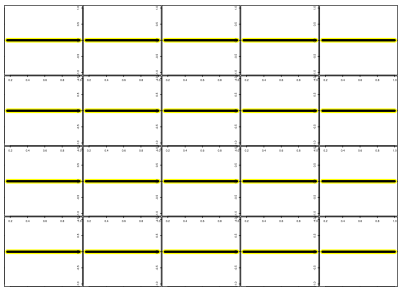
Figure 8 : Additivity Lineup Test - Least visually significant features



(a) nBT



(b) Fugacity



(c) Sp

Figure 9 : Additivity Lineup Test - Features exhibiting no visual significance

3.3.4 QSPR Model SHAP

The SHAP model summary (Figure 5) enumerates the absolute average and the range of SHAP values per feature, also referred to as the global SHAP value. When examining the force plots per tracer (Figure 11 and Appendix B: Figures B3.1 - B3.3) there are some differences in the top 5 features that contribute to predictions. For example, m-Xylene has time, boiling point, Ksp, W3D and C_o as the top 5 contributing features (figure 11b). Benzene had C_o ranked 3rd, J3D ranked 4th, and Ksp ranked 5th with respect to contribution (figure 11a). The replicates for tert-butylbenzene and pentamethylbenzene have identical features of importance, but the interaction amongst these features is variable across the 4000 predictions per tracer test (figure 11c - 11f). The interaction amongst the features contribution towards the prediction would be expected to be similar for the replicates, but based on the force plots this is not the case. Considering that XGBoost is governed by the objective function that it attempts to minimize, this level of variability between MD tests would make sense since the algorithm is tasked with using the available data to achieve the greatest possible reduction in prediction error.

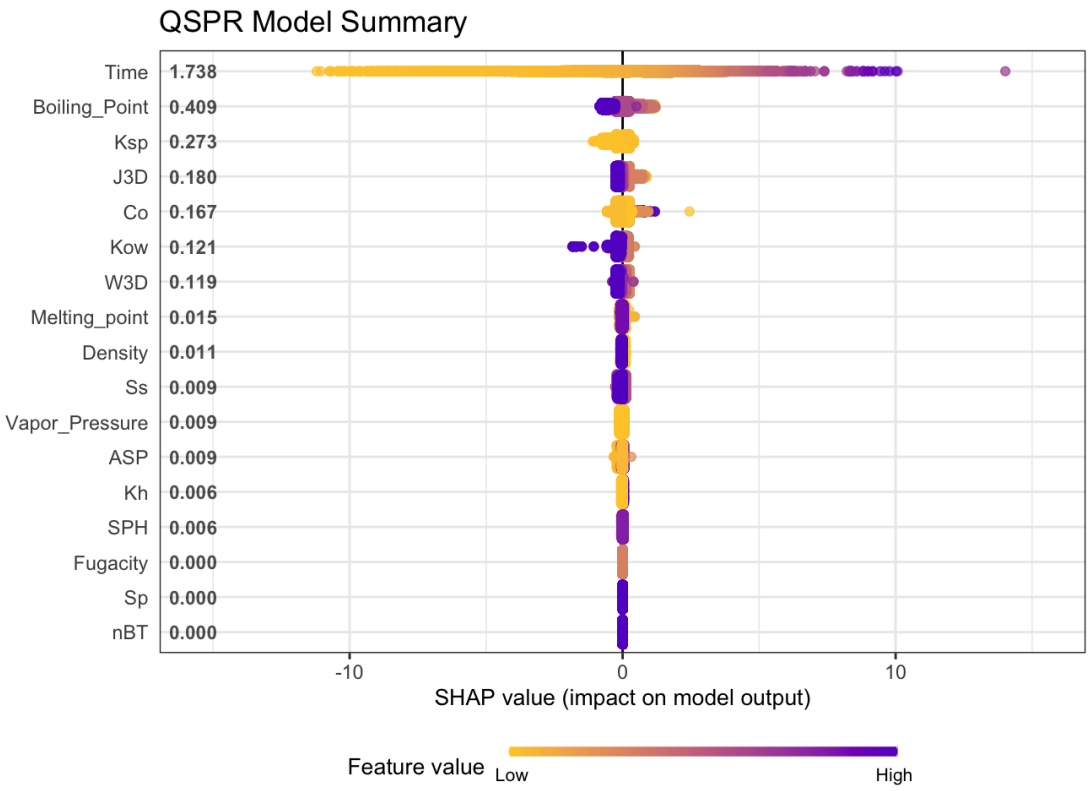
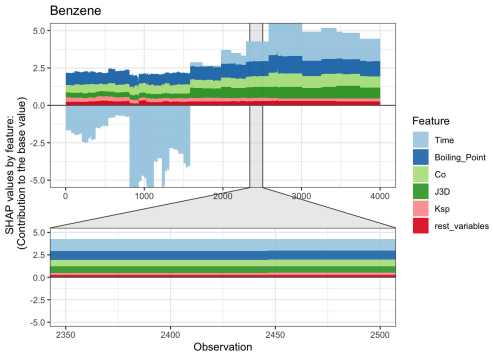
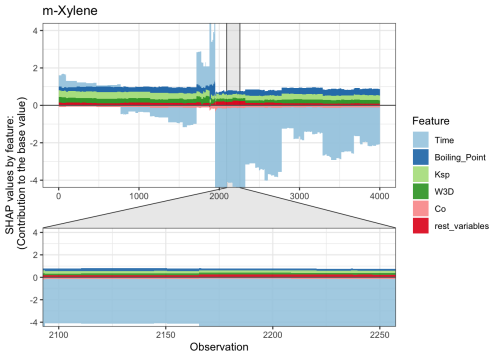


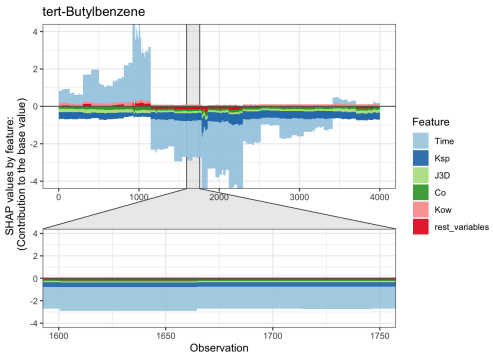
Figure 10 : Global SHAP Values of feature contribution on model output



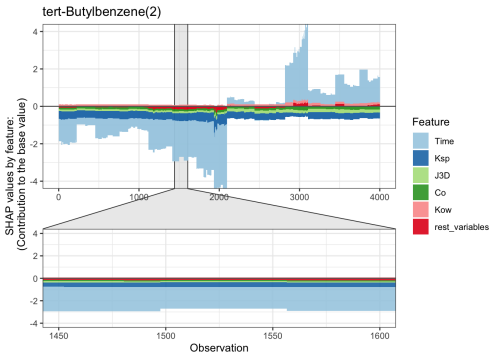
(a) Benzene Forceplot



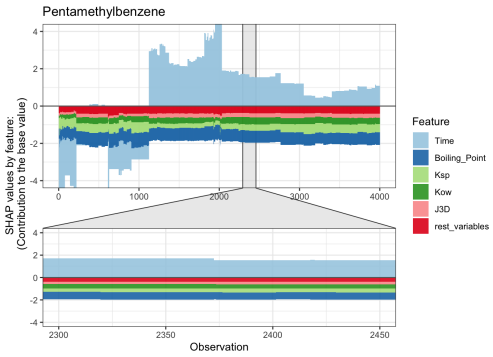
(b) m-Xylene Forceplot



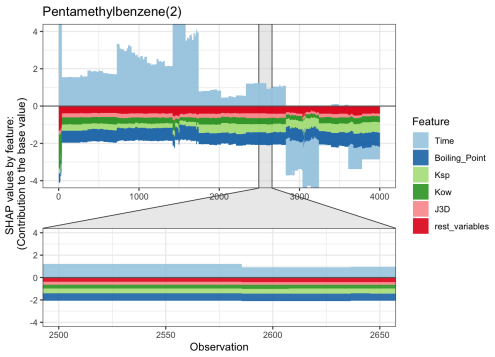
(c) tert-Butylbenzene Forceplot



(d) tert-Butylbenzene(2) Forceplot



(e) Pentamethylbenzene Forceplot



(f) Pentamethylbenzene(2) Forceplot

Figure 11 : Force plots for (a) Benzene, (b) m-Xylene (c) tert-Butylbenzene, (d) tert-Butylbenzene(2), (e) Pentamethylbenzene, and (f) Pentamethylbenzene(2)

Table 4
RMSE and R^2 out-of-sample (oos) predictive performance metrics per tracer

Tracer	Models											
	Time, Co		Time, Co, Ss		Time, Co, W3D		Time, Co, MW		Time, Co, Ksp		QSPR	
	RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2
Benzene	0.0750	0.9332	0.0867	0.9108	0.0793	0.9254	0.0856	0.9131	0.0815	0.9211	0.0825	0.9192
Toluene	0.0565	0.9538	0.0560	0.9545	0.0567	0.9535	0.0567	0.9535	0.0568	0.9532	0.0545	0.9569
Ethylbenzene	0.1030	0.8908	0.0150	0.9977	0.0107	0.9988	0.0143	0.9979	0.0282	0.9918	0.0276	0.9922
o-Xylene	0.2092	0.5080	0.0384	0.9834	0.0243	0.9934	0.0405	0.9816	0.0312	0.9891	0.0317	0.9887
m-Xylene	0.1632	0.7174	0.0147	0.9977	0.0231	0.9943	0.0128	0.9982	0.0180	0.9966	0.0168	0.9970
p-Xylene	0.0589	0.9510	0.0122	0.9979	0.0105	0.9984	0.0149	0.9969	0.0158	0.9965	0.0135	0.9974
1,2,3-Trimethylbenzene	0.0354	0.9861	0.0336	0.9875	0.0342	0.9871	0.0326	0.9882	0.0334	0.9876	0.0402	0.9821
1,2,4-Trimethylbenzene	0.0294	0.9884	0.0564	0.9573	0.0288	0.9889	0.0283	0.9892	0.0289	0.9888	0.0513	0.9647
1,3,5-Trimethylbenzene	0.2034	0.4991	0.0796	0.9234	0.0672	0.9453	0.0780	0.9263	0.0683	0.9436	0.0700	0.9407
n-Propylbenzene	0.0397	0.9844	0.0358	0.9873	0.0380	0.9857	0.0351	0.9878	0.0318	0.9900	0.0162	0.9974
Isopropylbenzene	0.0665	0.9482	0.0659	0.9491	0.0425	0.9788	0.0674	0.9468	0.0694	0.9435	0.0712	0.9406
1-Ethyl-2-methylbenzene	0.0362	0.9844	0.0243	0.9930	0.0368	0.9840	0.0362	0.9845	0.0370	0.9837	0.0354	0.9851
Isopropyl-4-methylbenzene	0.1712	0.9692	0.1653	0.9713	0.0414	0.9982	0.1689	0.9701	0.1701	0.9696	0.1698	0.9697
n-Butylbenzene	0.3392	-0.4010	0.2012	0.5072	0.2410	0.2930	0.2243	0.3872	0.0594	0.9571	0.1887	0.5664
tert-Butylbenzene	0.0303	0.9889	0.0329	0.9870	0.0313	0.9882	0.0326	0.9872	0.0307	0.9887	0.0328	0.9870
tert-Butylbenzene (2)	0.0307	0.9886	0.0313	0.9882	0.0303	0.9889	0.0312	0.9883	0.0321	0.9875	0.0321	0.9876
1,2,4,5-Tetramethylbenzene	0.1134	0.8513	0.1226	0.8263	0.1206	0.8321	0.1158	0.8450	0.1273	0.8126	0.1261	0.8162
Pentamethylbenzene	0.0307	0.9887	0.0331	0.9869	0.0315	0.9881	0.0315	0.9881	0.0322	0.9876	0.0278	0.9907
Pentamethylbenzene(2)	0.0363	0.9842	0.0333	0.9867	0.0331	0.9868	0.0334	0.9866	0.0378	0.9829	0.0315	0.9881
OOS Average	**0.1268**	**0.8766**	**0.0784**	**0.9529**	**0.0726**	**0.9595**	**0.0808**	**0.9498**	**0.0645**	**0.9680**	**0.0768**	**0.9548**

Table 4 : RMSE and R^2 values per tracer prediction

3.4 MODEL COMPARISON

The additional four models built using time, C_o, and one QSPR feature (molecular weight, Ksp, Ss, and W3D) resulted in comparable performance to the QSPR model. The model fit and error were as follows: (1) molecular weight model had average R^2 = .9458 and RMSE = 0.0808, (2) Ksp model had average R^2 = .9680 and RMSE = 0.0645, (3) Ss had average R^2 = .9529 and RMSE = 0.0784, and (4) W3D had average R^2 = .9595 and RMSE = 0.0726 (table 4). These values indicate performance that is comparable to the more complex QSPR model. When examining the individual tracer performance metrics, the Ksp model provides the most improvement for the outlier tracer n-Butylbenzene, with an R^2 = .9571. The reduction in RMSE for n-Butylbenzene in the Ksp model compared to the QSPR model is roughly 67%.

4. Conclusions

XGBoost algorithm is highly adaptive and sensitive to the given dataset on which it is trained on. The BTC predictions were both monotonically increasing and closely approximated the true BTCs. The XGBoost algorithm was able to enforce monotonicity due to its native hyperparameter montone_constraints. With the exception of benzene, n-butylbenzene, 1-isopropy-4-methylbenzene, and 1,2,4,5-tetramethylbenzene; the predictions for the remaining MAHs resulted in RMSE values less than 0.08 (Figure 3 and Appendix B: B1.1 - B1.2). The R^2 values for 16 of the 19 predictions were greater than 0.91, with 11 of the predictions achieving R^2 values greater than or equal to 0.98. The resulting models exhibit variability in feature contribution towards the model's predictions. The results and analysis of the feature interactions as they pertain to the predictions echo a common conundrum in modeling complex systems, the trade off between performance and interpretability (Breiman, 2001). The adaptability of XGBoost enables high performance but at the cost of interpretability, especially in high dimensional systems. Given that QSPR models are known to also suffer from collinearities when increasing the space of features (figure 1), the results suggest that the simpler model is preferred if interpretability is the goal. Tools such as SHAP and c-ICE are useful in understanding the degree of feature contribution towards the predictions and feature interactions. SHAP enables the quantification of feature interaction both locally at each prediction point and globally for the entire set of predictions, both of which are useful when analyzing how XGBoost ar-

rives at the prediction endpoints. Additivity lineup test enable an alternate visual measure to test for significance, reaffirming the features of importance values from XGBoost (table 6) and SHAP (figure 9). However, feature interactions within the context of the model's predictions and insights as to the system are not necessarily the same. The SHAP force plots can visually display the adaptable nature of the algorithm in predicting, including rapid shifts in how the features interact towards the generated output. There was strong agreement between the additivity lineup test, SHAP global values, the linear summary (Appendix A: Table A2.1), and the gain values of XGBoost when comparing features that had zero contribution to the model. Amongst the remaining features the impact of time was substantially greater than the other features. When comparing the top five most useful features, the SHAP global values and the gain values of XGBoost are in agreement, both of which found time, boiling point, Ksp, J3D, and C_o to be the most important. However, SHAP found Ksp to be the third most important feature for predictions and J3D to be the fourth most important. The gain values enumerated J3D to be the third most important followed by Ksp as the fourth most important feature indicating a discrepancy in how the model internally quantifies feature importance and feature interactions. But any definitive conclusion about a features impact on transport using the QSPR model, given the low sample size inherent in the available dataset, is difficult based on the exploratory analysis conducted.

These results however strongly suggest that time and a coalition of features such as boiling point, Ksp, J3D, and C_o will have a sizable impact on any future predictions. The results from the modeling step that involved obtaining the residuals resulted in high R^2

values for most of the tracers, suggesting that majority of solute transport conducted in column studies through Berea sandstone may be adequately explained with time and injection concentration (table 4). To further reduce the residuals, XGBoost needs one additional useful feature to improve predictive performance. This still leaves an open question as to which feature is ideal to select to model solute transport. When comparing all the models, the results from the model built using using Ksp, time, and C_o appear to have the highest performance based on RMSE and R^2 values. The difference in performance however is not large significant enough to assume that Ksp as a feature will be sufficient in all future prediction scenarios for unseen tracer tests. Further work needs to be conducted to obtain more data for the MAH family of compounds, including expanding the list of compounds. Additional tests that include using multiple rock cores with varying porous media properties, varying the length of the rock core, varying injection velocities, and multiple injection concentrations per tracer are worthwhile to include in future ML modeling since these features were not included in this work.

A. Supplementary Tables

A.1 FEATURE SELECTION

Lasso Coefficients	
	s0
(Intercept)	-8.096992e+00
Molecular_Weight	.
Melting_point	-3.984572e-03
Boiling_Point	-1.261063e-03
Fugacity	-1.177316e+00
Density	7.993539e+00
Molar_Volume	.
Le_Bas	.
Vapor_Pressure	-9.248944e-04
Ksp	4.892919e-01
Kow	1.861004e-05
Kh	1.435615e-04
Sv	-1.853164e-02
Se	.
Sp	-8.546750e-07
Ss	7.977446e-01
nBT	-4.449481e-01
nBO	.
W3D	-1.251845e-03
J3D	-2.103378e-01
SPH	-6.441685e-01
ASP	1.078609e+00

Table A1.1 : Lasso Coefficients

Backwards Deletion Coefficients

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	1.214e+02	2.888e+00	42.031	<2e-16	***
Melting_point	-6.706e-03	5.368e-05	-124.924	<2e-16	***
Boiling_Point	4.111e-03	1.452e-04	28.314	<2e-16	***
Fugacity	-1.244e+00	1.382e-02	-90.015	<2e-16	***
Density	1.900e+00	1.176e-01	16.157	<2e-16	***
Vapor_Pressure	-1.246e-03	8.234e-06	-151.275	<2e-16	***
Ksp	6.117e-01	5.371e-03	113.898	<2e-16	***
Kow	4.858e-05	5.130e-07	94.700	<2e-16	***
Kh	1.555e-04	2.931e-06	53.066	<2e-16	***
Sv	3.542e-01	2.788e-01	1.271	0.204	
Sp	-1.014e+02	2.282e+00	-44.428	<2e-16	***
Ss	1.268e+00	9.855e-03	128.619	<2e-16	***
nBT	5.858e+01	1.309e+00	44.765	<2e-16	***
W3D	-2.591e-03	1.143e-04	-22.678	<2e-16	***
J3D	-4.932e-01	3.286e-02	-15.010	<2e-16	***
SPH	-2.933e-01	1.278e-02	-22.956	<2e-16	***
ASP	8.862e-01	7.888e-03	112.341	<2e-16	***

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.06102 on 75983 degrees of freedom
Multiple R-squared: 0.7674, Adjusted R-squared: 0.7674
F-statistic: 1.567e+04 on 16 and 75983 DF, p-value: < 2.2e-16

Table A1.2 : Backwards Deletion Features of Significance

A.2 QSPR SUMMARY

Linear Regression Summary

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	-1.775e+01	1.376e+01	-1.289	0.197271	
Melting_point	1.354e-03	2.316e-04	5.844	5.10e-09	***
Boiling_Point	2.367e-02	6.612e-04	35.801	< 2e-16	***
Fugacity	2.110e-01	5.677e-02	3.716	0.000202	***
Density	-1.706e+01	5.584e-01	-30.554	< 2e-16	***
Vapor_Pressure	1.168e-03	3.950e-05	29.578	< 2e-16	***
Ksp	-5.123e-01	2.536e-02	-20.197	< 2e-16	***
Kow	3.065e-05	2.313e-06	13.255	< 2e-16	***
Kh	-3.975e-04	1.332e-05	-29.844	< 2e-16	***
Sp	1.732e+01	1.066e+01	1.625	0.104238	
Ss	3.898e-01	4.309e-02	9.046	< 2e-16	***
nBT	-9.457e+00	6.237e+00	-1.516	0.129447	
W3D	-1.075e-02	3.954e-04	-27.186	< 2e-16	***
J3D	-1.885e+00	1.188e-01	-15.865	< 2e-16	***
SPH	5.648e-01	6.118e-02	9.233	< 2e-16	***
ASP	1.601e-01	3.679e-02	4.351	1.36e-05	***

Signif codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’

Residual standard error: 0.2927 on 75984 degrees of freedom
Multiple R-squared: 0.1698, Adjusted R-squared: 0.1696
F-statistic: 1036 on 15 and 75984 DF, p-value: < 2.2e-16

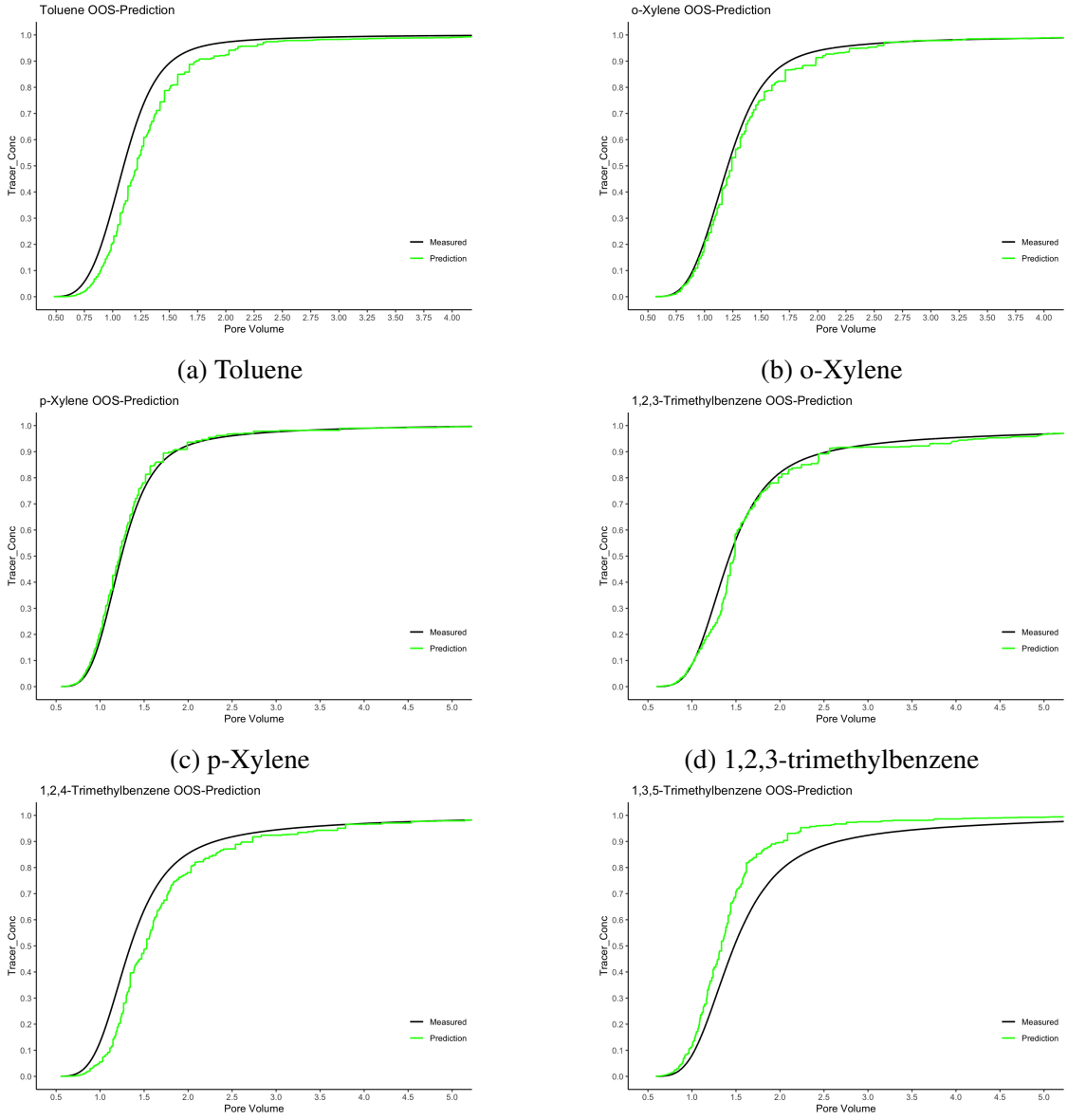
Table A2.1 : Linear Summary for QSPR model

Feature	Percent_Gain
Time	66.850
Boiling Point	10.043
J3D	9.686
Ksp	7.244
Co	2.457
Kow	1.449
W3D	1.233
Vapor Pressure	0.183
Ss	0.146
Melting Point	0.144
Density	0.099
SPH	0.042
ASP	0.032
Kh	0.009
Sp	0.000
nBT	0.000
Fugacity	0.000

Table A2.2 : Feature gain represented by percentage for QSPR model

B. Supplementary Figures

B.1 BTC



(a) Toluene

(b) o-Xylene

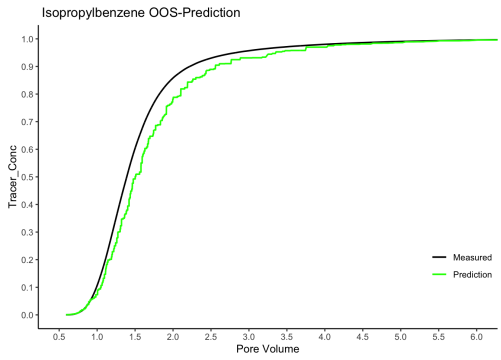
(c) p-Xylene

(d) 1,2,3-trimethylbenzene

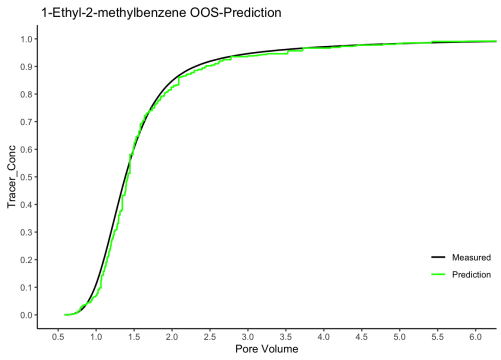
(e) 1,2,4-trimethylbenzene

(f) 1,3,5-trimethylbenzene

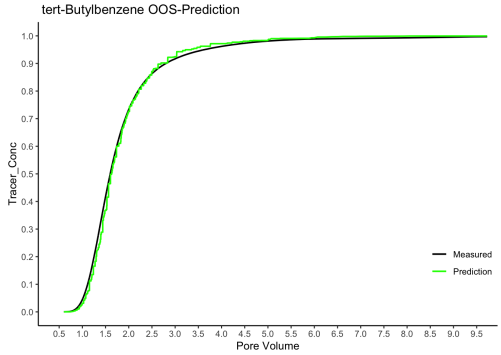
Figure B1.1 : BTCs for (a) Toluene, (b) o-Xylene, (c) p-Xylene, (d) 1,2,3-trimethylbenzene, (e) 1,2,4-trimethylbenzene, and (f) 1,3,5-trimethylbenzene



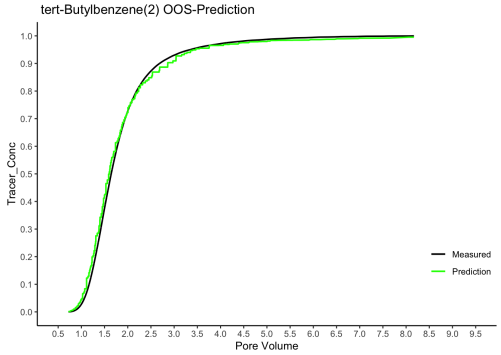
(a) Isopropylbenzene



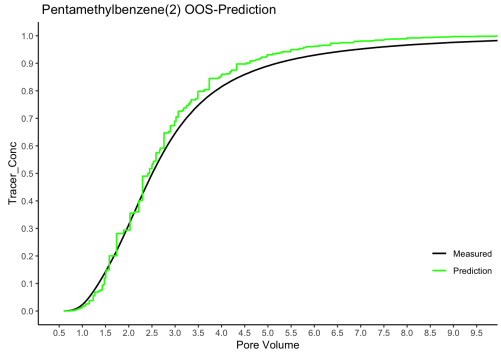
(b) 1-ethyl-2-methylbenzene



(c) tert-Butylbenzene



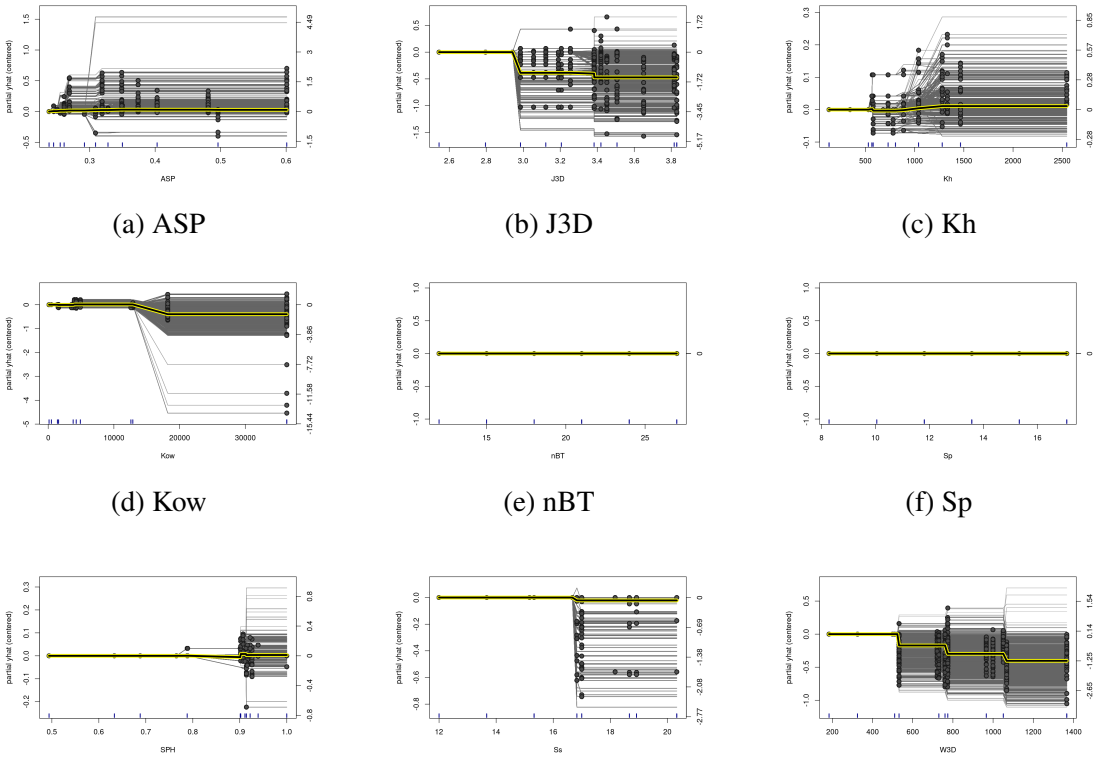
(d) tert-Butylbenzene(2)



(e) pentamethylbenzene(2)

Figure B1.2 : BTCs for (a) Isopropylbenzene, (b) 1-ethyl-2-methylbenzene, (c) tert-Butylbenzene, (d) tert-Butylbenzene(2), and (e) Pentamethylbenzene(2)

B.2 C-ICE PLOTS



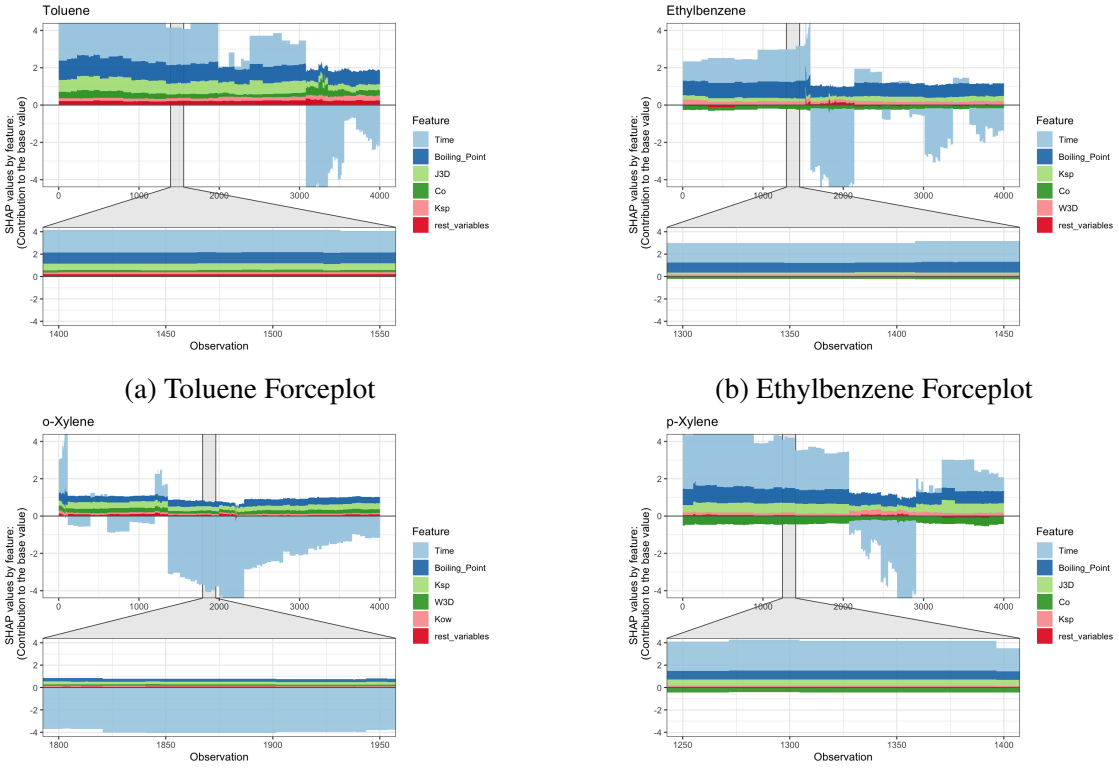
(a) ASP (b) J3D (c) Kh

(d) Kow (e) nBT (f) Sp

(g) SPH (h) Ss (i) W3D

Figure B2 : c-ICE plots per feature

B.3 SHAP FORCE PLOTS



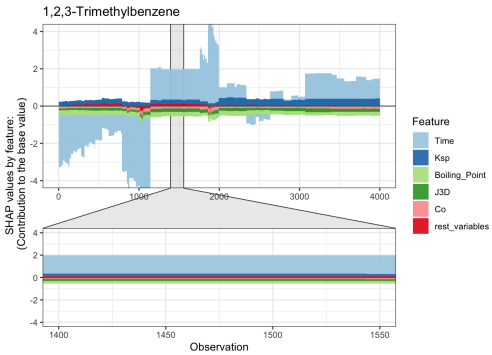
(a) Toluene Forceplot

(b) Ethylbenzene Forceplot

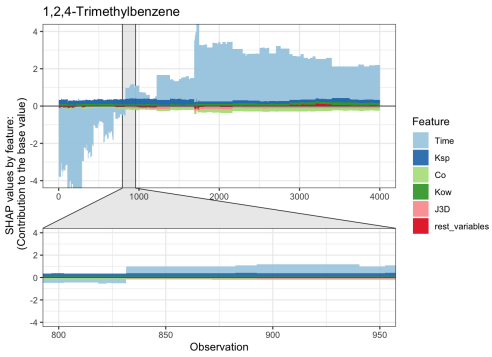
(c) o-Xylene Forceplot

(d) p-Xylene Forceplot

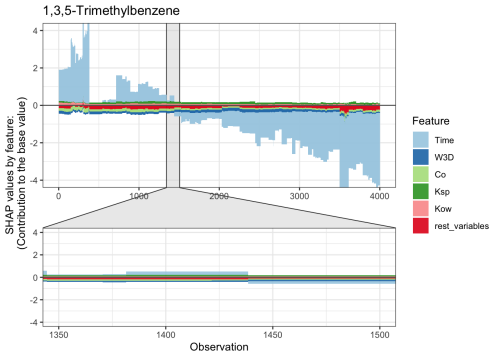
Figure B3.1 : Force plots for (a) Toluene, (b) Ethylbenzene (c) o-Xylene, and (d) p-Xylene



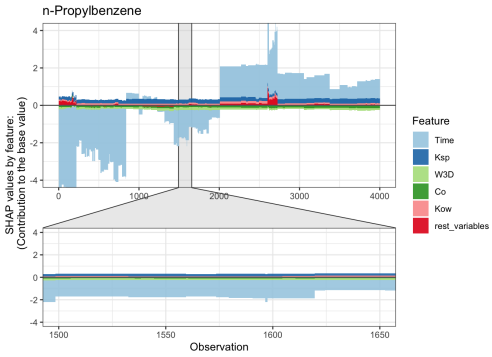
(a) 1,2,3-trimethylbenzene Forceplot



(b) 1,2,4-trimethylbenzene Forceplot

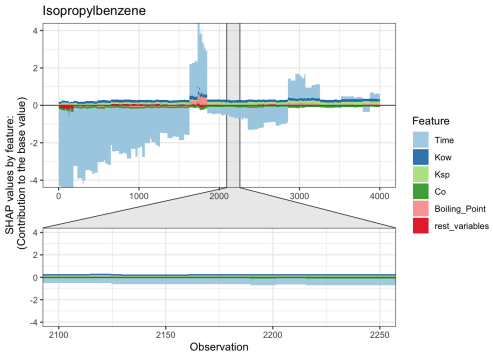


(c) 1,3,5-trimethylbenzene Forceplot

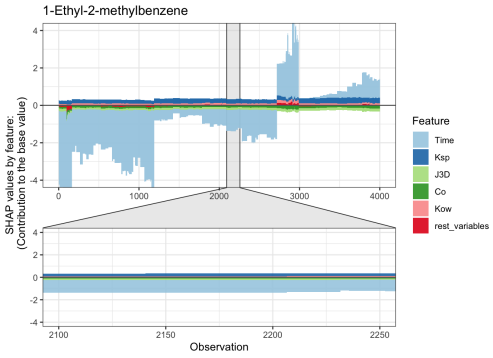


(d) n-propylbenzene Forceplot

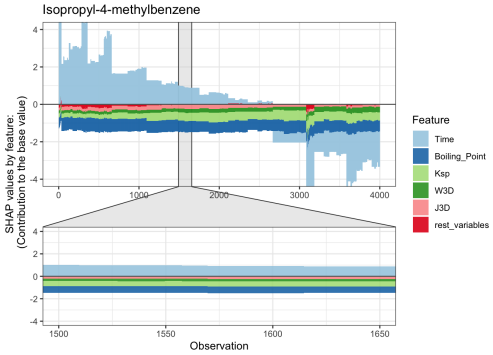
Figure B3.2 : Force plots for (a) 1,2,3-trimethylbenzene, (b) 1,2,4-trimethylbenzene (c) 1,3,5-trimethylbenzene, and (d) n-propylbenzene



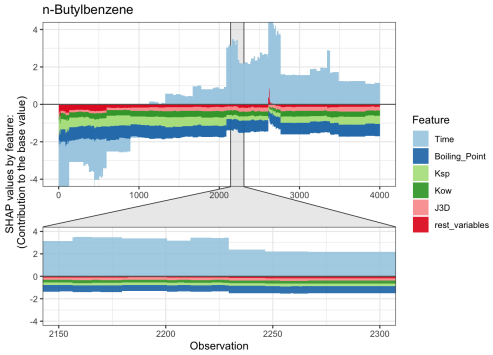
(a) Isopropylbenzene Forceplot



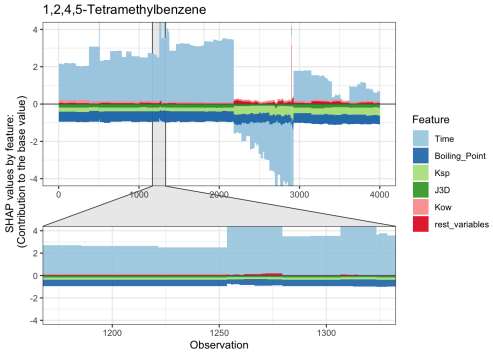
(b) 1-ethyl-2-methylbenzene Forceplot



(c) Isopopyl-4-methylbenzene Forceplot



(d) n-butylbenzene Forceplot



(e) 1,2,4,5-tetramethylbenzene Forceplot

Figure B3.3 : Force plots for (a) Isopropylbenzene, (b) 1-ethyl-2-methylbenzene (c) Isopopyl-4-methylbenzene, (d) n-butylbenzene, and (e) 1,2,4,5-tetramethylbenzene

Bibliography

Ahsan, Md Manjurul et al. (2021). "Effect of Data Scaling Methods on Machine Learning Algorithms and Model Performance". In: *Technologies* 9.3. ISSN: 2227-7080. DOI: 10.3390/technologies9030052. URL: https://www.mdpi.com/2227-7080/9/3/52.

Arenas, Alejandro Yopasa et al. (2018). "Mobility of polivinylpyrrolidone coated silver nanoparticles in tropical soils". In: *Chemosphere* 194, pp. 543–552.

Asirvatham, Sahaya, Bharat V Dhokchawle, and Savita J Tauro (2019). "Quantitative structure activity relationships studies of non-steroidal anti-inflammatory drugs: A review". In: *Arabian Journal of Chemistry* 12.8, pp. 3948–3962.

Baker, Nathan et al. (2019). *Workshop report on basic research needs for scientific machine learning: Core technologies for artificial intelligence*. Tech. rep. USDOE Office of Science (SC), Washington, DC (United States).

Ballabio, Davide et al. (2009). "Introduction to MOLE DB - on-line Molecular Descriptors Database". In.

Bhavsar, Parth et al. (2017). "Machine learning in transportation data analytics". In: *Data analytics for intelligent transportation systems*. Elsevier, pp. 283–307.

Box, George E.P. and Norman R. Draper (1987). *Empirical Model-Building And Response Surfaces*. John Wiley amp; Sons.

Breiman, Leo (Aug. 2001). "Statistical Modeling: The Two Cultures (with Comments and a Rejoinder by the Author)". In: *Statistical Science* 16, pp. 199–231. ISSN: 0883-4237. DOI: 10.1214/ss/1009213726. URL: https://doi.org/10.1214/ss/1009213726.

Breiman, Leo et al. (1984). *Classification and regression trees*. wadsworth amp; brooks.

Buja, Andreas et al. (2009). "Statistical Inference for Exploratory Data Analysis and model diagnostics". In: *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 367.1906, 4361–4383. DOI: 10.1098/rsta.2009.0120.

Castrillo, María and Álvaro López García (2020). "Estimation of high frequency nutrient concentrations from water quality surrogates using machine learning methods". In: *Water Research* 172, p. 115490.

Chen, Kangyang et al. (2020). "Comparative analysis of surface water quality prediction performance and identification of key water parameters using different machine learning models based on big data". In: *Water research* 171, p. 115454.

Chen, Tianqi and Carlos Guestrin (2016). "XGBoost: A Scalable Tree Boosting System". In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD '16. San Francisco, California, USA: Association for Computing Machinery, 785–794. ISBN: 9781450342322. DOI: 10.1145/2939672.2939785. URL: https://doi.org/10.1145/2939672.2939785.

Chen, Tianqi et al. (n.d.). *Understanding your dataset with XGBoost*. URL: https://cran.r-project.org/web/packages/xgboost/vignettes/discoverYourData.html.

Chowdhury, Mohammad Ziaul Islam and Tanvir C Turin (2020). "Variable selection strategies and its importance in clinical prediction modelling". In: *Family Medicine and Community Health* 8.1. ISSN: 2305-6983. DOI: 10.1136/fmch-2019-000262. URL: https://fmch.bmj.com/content/8/1/e000262.

Dearden, John C (2002). "Prediction of environmental toxicity and fate using quantitative structure-activity relationships (QSARs)". In: *Journal of the Brazilian Chemical Society* 13, pp. 754–762.

Dearden, John C. (2017). "The use of topological indices in QSAR and QSPR modeling". In: *Challenges and Advances in Computational Chemistry and Physics*, 57–88. DOI: 10.1007/978-3-319-56850-8_2.

Efron, Bradley and Gail Gong (1983). "A leisurely look at the bootstrap, the jackknife, and cross-validation". In: *The American Statistician* 37.1, p. 36. DOI: 10.2307/2685844.

EPA, U.S. (2015). "Analysis of hydraulic fracturing fluid data from the FracFocus chemical disclosure registry 1.0". In: p. 31.

— (2016). "Hydraulic fracturing for oil and gas: Impacts from the hydraulic fracturing water cycle on drinking water resources in the United States". In: *EPA's Study Hydraul. Fract. Its Potential Impact Drink. Water Resour.*, pp. 1–666.

Ferrar, Kyle J et al. (2013). "Assessment of effluent contaminants from three facilities discharging Marcellus Shale wastewater to surface waters in Pennsylvania". In: *Environmental science & technology* 47.7, pp. 3472–3481.

Fetter, CW, TB Boving, and D Kreamer (2018). *Contaminant Hydrogeology-ISBN: ISBN 10: 1-4786-3279-8, Publisher*. Waveland Press, Inc.

Friedman, Jerome, Trevor Hastie, and Rob Tibshirani (2010). "Regularization paths for generalized linear models via coordinate descent". In: *Journal of statistical software* 33.1, p. 1.

Friedman, Jerome H. (2001). "Greedy function approximation: A gradient boosting machine." In: *The Annals of Statistics* 29.5, pp. 1189 –1232. DOI: 10.1214/aos/1013203451. URL: https://doi.org/10.1214/aos/1013203451.

Gao, Yidan et al. (2021). "Quantitative structure activity relationships (QSARs) and machine learning models for abiotic reduction of organic compounds by an aqueous Fe (II) complex". In: *Water Research* 192, p. 116843.

Ghasemi, Fahimeh et al. (2018). "Neural network and deep-learning algorithms used in QSAR studies: merits and drawbacks". In: *Drug discovery today* 23.10, pp. 1784–1790.

Goldstein, Alex et al. (2015). "Peeking inside the black box: Visualizing statistical learning with plots of individual conditional expectation". In: *journal of Computational and Graphical Statistics* 24.1, pp. 44–65.

Graham, Michael H (2003). "Confronting multicollinearity in ecological multiple regression". In: *Ecology* 84.11, pp. 2809–2815.

Grathwohl, Peter (2012). *Diffusion in natural porous media: contaminant transport, sorption/desorption and dissolution kinetics*. Vol. 1. Springer Science & Business Media.

Hartmann, K., J. Krois, and B. Waske (2016). *E-Learning Project SOGA: Statistics and Geospatial Data Analysis.* URL: https://www.geo.fu-berlin.de/en/v/soga/Basics-of-statistics/Logistic-Regression/The-Logit-Function/index.html.

Hastie, Trevor et al. (2009). *The elements of statistical learning: data mining, inference, and prediction*. Vol. 2. Springer.

Hawkins, Douglas M (2004). "The problem of overfitting". In: *Journal of chemical information and computer sciences* 44.1, pp. 1–12.

Hu, Liyang et al. (2021). "Estimating gaseous pollutants from bus emissions: A hybrid model based on GRU and XGBoost". In: *Science of The Total Environment* 783, p. 146870. ISSN:

0048-9697. DOI: https://doi.org/10.1016/j.scitotenv.2021.146870. URL: https://www.sciencedirect.com/science/article/pii/S0048969721019409.

Hu, Qinhong and Mark L Brusseau (1995). "Effect of solute size on transport in structured porous media". In: *Water Resources Research* 31.7, pp. 1637–1646.

Huang, Ruixing et al. (2021). "Machine learning in natural and engineered water systems". In: *Water Research* 205, p. 117666.

Jaiswal, Dilip Kumar, Atul Kumar, and Raja Ram Yadav (2011). "Analytical solution to the one-dimensional advection-diffusion equation with temporally dependent coefficients". In: *Journal of Water Resource and Protection* 2011.

Jeong, Nohyeong, Tai-heng Chung, and Tiezheng Tong (2021). "Predicting Micropollutant Removal by Reverse Osmosis and Nanofiltration Membranes: Is Machine Learning Viable?" In: *Environmental Science & Technology* 55.16. PMID: 34342439, pp. 11348–11359. DOI: 10.1021/acs.est.1c04041. URL: https://doi.org/10.1021/acs.est.1c04041.

Kareem, Rikan et al. (2017). "Multi-technique approach to the petrophysical characterization of Berea sandstone core plugs (Cleveland Quarries, USA)". In: *Journal of Petroleum Science and Engineering* 149, pp. 436–455.

Kirschbaum, Mark A et al. (2012). "Assessment of undiscovered oil and gas resources of the Ordovician Utica Shale of the Appalachian Basin Province, 2012". In: *US Geological Survey Fact Sheet* 3116.6.

Labrecque, Steven P and William J Blanford (2021). "Fate and transport of bromide and mononuclear aromatic hydrocarbons in aqueous solutions through Berea Sandstone". In: *Science of The Total Environment* 766, p. 141714.

Li, Lingbo et al. (2022). “Interpretable tree-based ensemble model for predicting beach water quality”. In: *Water Research*, p. 118078.

Limousin, G et al. (2007). “Sorption isotherms: A review on physical bases, modeling and measurement”. In: *Applied geochemistry* 22.2, pp. 249–275.

Liu, Yang and Allan Just (2020). *SHAPforxgboost: SHAP Plots for 'XGBoost'*. R package version 0.1.0. URL: https://github.com/liuyanguu/SHAPforxgboost/.

Llewellyn, Garth T et al. (2015). “Evaluating a groundwater supply contamination incident attributed to Marcellus Shale gas development”. In: *Proceedings of the National Academy of Sciences* 112.20, pp. 6325–6330.

Lundberg, Scott M and Su-In Lee (2017). “A unified approach to interpreting model predictions”. In: *Advances in neural information processing systems* 30.

Mackay, Donald, Wan-Ying Shiu, and Sum Chi Lee (2006). *Handbook of physical-chemical properties and environmental fate for organic chemicals*. CRC press.

Mamy, Laure et al. (2015). “Prediction of the fate of organic compounds in the environment from their molecular properties: a review”. In: *Critical Reviews in Environmental Science and Technology* 45.12, pp. 1277–1377.

Molecular descriptors Guide (2008). URL: https://www.epa.gov/sites/production/files/2015-05/documents/moleculardescriptorsguide-v102.pdf.

Osborne, MR, B Presnell, and BA Turlach (1998). “Knot selection for regression splines via the lasso”. In: *Computing Science and Statistics*, pp. 44–49.

Peck, Elizabeth A, Douglas C Montgomery, and G Geoffrey Vining (2013). *Solutions Manual to accompany Introduction to Linear Regression Analysis*. John Wiley & Sons.

Perrin, J. and T. Cook (n.d.). *Hydraulically fractured horizontal wells account for most new oil and natural gas wells*. URL: https://www.eia.gov/todayinenergy/detail.php?id=37815.

Piwoni, Marvin D and Jack W Keeley (1990). *Ground water issue: basic concepts of contaminant sorption at hazardous waste sites*. Superfund technology Support Center for Ground Water, Robert S. Kerr ...

Ruengvirayudh, Pornchanok and Gordon P Brooks (2016). "Comparing stepwise regression models to the best-subsets models, or, the art of stepwise". In: *General linear model journal* 42.1, pp. 1–14.

Ryder, Robert T et al. (2012). "Geologic cross section CC'through the Appalachian basin from Erie County, north-central Ohio, to the Valley and Ridge province, Bedford County, south-central Pennsylvania". In.

Salvador, Jordi (2016). *Example-Based super resolution*. Academic Press.

Santosa, Fadil and William W Symes (1986). "Linear inversion of band-limited reflection seismograms". In: *SIAM Journal on Scientific and Statistical Computing* 7.4, pp. 1307–1330.

Schwarzenbach, Rene P and John Westall (1981). "Transport of nonpolar organic compounds from surface water to groundwater. Laboratory sorption studies". In: *Environmental Science & Technology* 15.11, pp. 1360–1367.

Tibshirani, Robert (1996). "Regression shrinkage and selection via the lasso". In: *Journal of the Royal Statistical Society: Series B (Methodological)* 58.1, pp. 267–288.

Tseng, Yufeng J, Anton J Hopfinger, and Emilio Xavier Esposito (2012). "The great descriptor melting pot: mixing descriptors for the common good of QSAR models". In: *Journal of computer-aided molecular design* 26.1, pp. 39–43.

Van Genuchten, M Th (1982). *Analytical solutions of the one-dimensional convective-dispersive solute transport equation*. 1661. US Department of Agriculture, Agricultural Research Service.

Vinayak, Rashmi Korlakai and Ran Gilad-Bachrach (2015). "Dart: Dropouts meet multiple additive regression trees". In: *Artificial Intelligence and Statistics*. PMLR, pp. 489–497.

Weller, Daniel L., Tanzy M. T. Love, and Martin Wiedmann (2021). "Interpretability Versus Accuracy: A Comparison of Machine Learning Models Built Using Different Algorithms, Performance Measures, and Features to Predict E. coli Levels in Agricultural Water". In: *Frontiers in Artificial Intelligence* 4. DOI: 10.3389/frai.2021.628441. URL: https://www.frontiersin.org/article/10.3389/frai.2021.628441.

Wickham, Hadley et al. (2010). "Graphical inference for infovis". In: *IEEE Transactions on Visualization and Computer Graphics* 16.6, pp. 973–979.

Xu, Tingting, Giovanni Coco, and Martin Neale (2020). "A predictive model of recreational water quality based on adaptive synthetic sampling algorithms and machine learning". In: *Water research* 177, p. 115788.

Yousefinejad, Saeed and Bahram Hemmateenejad (2015). "Chemometrics tools in QSAR/QSPR studies: A historical perspective". In: *Chemometrics and Intelligent Laboratory Systems* 149, pp. 177–204.

Yu, Rong et al. (2018). "Diffusion-coupled degradation of chlorinated ethenes in sandstone: An intact core microcosm study". In: *Environmental science & technology* 52.24, pp. 14321–14330.

Zhao, Peng and Bin Yu (2006). "On model selection consistency of Lasso". In: *The Journal of Machine Learning Research* 7, pp. 2541–2563.

Zimmerman, Mitchell D et al. (2002). "Experimental determination of sorption in fractured flow systems". In: *Journal of contaminant hydrology* 58.1-2, pp. 51–77.