Environmental Science & Technology

pubs.acs.org/est

Article

From Model Development to Mitigation: Machine Learning for Predicting and Minimizing Iodinated Trihalomethanes in Water Treatment

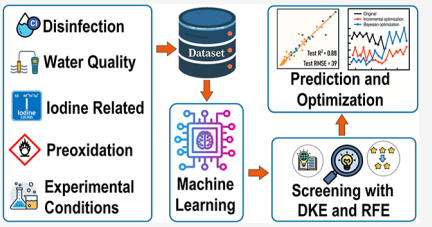
Md. Mahjib Hossain,[#] Rabbi Sikder,[#] Guanghui Hua, and Tao Ye*

Cite This: https://doi.org/10.1021/acs.est.5c05409

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: Disinfection processes in water treatment produce disinfection byproducts (DBPs), such as iodinated trihalomethanes (I-THMs), which pose significant health risks. Mitigating I-THMs remains challenging due to the complex interactions among water quality parameters, disinfectants, and iodine sources, compounded by the difficulty of predicting their formation under varying treatment conditions. This study leverages a data set of 1534 samples from published studies to predict I-THM formation using machine learning (ML). Among five evaluated ensemble models, CatBoost Regression achieved the best performance. Incorporating domain-specific features (iodine/DOC and oxidant/DOC ratios) improved model accuracy and interpretability. Recursive feature elimination revealed that nearly half of the features could be excluded without compromising performance, simplifying model development and reducing experimental effort, an advantage often overlooked in prior research. Feature analysis identified key predictors and mitigation strategies, including minimizing iodine and bromide concentrations, reducing iodine/DOC, UV_{254} and SUVA levels, and optimizing chlorine dose. The model further enabled rapid identification of the optimal chlorine dose to minimize I-THMs using incremental and Bayesian optimization. Achieving an R^2 of 0.67 on an external validation data set, the model demonstrated strong generalizability. This study establishes ML as a powerful tool for predicting and mitigating I-THMs, offering actionable strategies for safer drinking water treatment.



KEYWORDS: *iodinated disinfection byproducts (I-DBPs), iodinated trihalomethanes (I-THMs), machine learning (ML), iodinated contrast media (ICM), feature engineering*

1. INTRODUCTION

Access to safe drinking water is vital for public health, with effective disinfection playing a crucial role in mitigating pathogenic contamination and protecting against waterborne diseases.[1−3] Among various disinfection methods, chlorine-based disinfection (e.g., chlorination and chloramination) remains the most widely used due to its cost effectiveness and operational simplicity. However, a notable drawback of chlorine disinfection is the formation of toxic disinfection byproducts (DBPs), which result from complex reactions between disinfectants, natural organic matter, anthropogenic pollutants, and halides (e.g., bromide and iodide) present in water.[4,5] The most common DBPs in drinking water are trihalomethanes (THMs) and haloacetic acids (HAAs), which are typically regulated by national and international agencies to safeguard public health.[6] For example, in Europe, the guideline value for total THM concentration (sum of four species) in drinking water is 100 μg/L,[7] whereas the United States imposes stricter regulatory limits of 80 μg/L for THMs and 60 μg/L for HAAs (sum of five species).[8] To meet current regulatory standards, many water utilities in the United States have transitioned from chlorination to chloramination.[9,10]

Chloramines, particularly monochloramine (NH_2Cl), are less reactive with dissolved organic matter (DOM), resulting in the formation of less regulated DBPs.[11,12] However, numerous studies have reported that chloramination can still produce several unregulated DBPs, including iodinated DBPs (I-DBPs, e.g., iodinated trihalomethanes [I-THMs], haloacetic acids [I-HAAs]).[13−17] During chloramination, monochloramine can oxidize iodide (I^-) from natural sources, such as seawater intrusion or brine, into hypoiodous acid (HOI), which subsequently reacts with DOM to form I-DBPs (Figure S1).[5,18,19] Interestingly, I-DBPs have been detected in drinking water treatment plants (DWTPs) with very low or undetectable levels of iodide, suggesting that alternative iodine sources may also contribute to I-DBP formation.[20] Among these alternative sources, iodinated X-ray contrast media (ICM, e.g., iopamidol, iohexol, iodixanol, diatrizoate, and iophromide) have been confirmed to act as precursors for I-DBP formation during disinfection.[21] ICM, widely used in

Received: April 24, 2025
Revised: May 26, 2025
Accepted: May 28, 2025

medical imaging of soft tissues (e.g., organs, veins, and blood vessels), are chemically designed as inert drugs with low environmental toxicity and minimal risk to human health.[22,23] However, their chemical stability and hydrophilicity render them resistant to removal by conventional wastewater treatment, leading to elevated concentrations in wastewater effluents, surface waters, and drinking water.[20,21] Several previous studies have demonstrated substantial I-THM formation in water containing iodide and ICM.[20,21,24−31] Despite this recognition, there remains a knowledge gap regarding which iodine source, iodide or ICM, plays a more significant role in I-THM formation. Since iodine sources are indispensable for I-DBP formation, identifying whether iodide or ICM is the dominant precursor is critical. This knowledge will allow water treatment facilities to design targeted removal strategies for the specific iodine source, thereby optimizing treatment efficiency and minimizing I-THM formation risks more effectively.

DBPs have been extensively studied for over 50 years.[32,33] Despite decades of research, their analysis remains a significant challenge due to their chemical diversity, trace-level concentrations, and the complex nature of water matrices, necessitating advanced analytical techniques, such as gas chromatography (GC) and liquid chromatography (LC).[34] These challenges hinder the development of quick, reliable, and cost-effective online measurement methods. I-DBPs, though not currently regulated, exhibit significantly higher cytotoxicity, genotoxicity, and developmental toxicity compared to their brominated and chlorinated counterparts.[29,35−37] Among I-DBPs, six I-THMs, including triiodomethane (iodoform, TIM), bromodiiodomethane (BDIM), dibromoiodomethane (DBIM), bromochloroiodomethane (BCIM), chlorodiiodomethane (CDIM), and dichloroiodomethane (DCIM), are the most frequently detected species in disinfected waters.[38] The frequent detection of I-THMs, combined with their heightened toxicity, raises serious concerns and highlight the urgent need to study their formation mechanisms and environmental prevalence. Addressing these challenges requires not only advanced analytical techniques and skilled expertise but also the development of more efficient, scalable, and cost-effective approaches that overcome the limitations of traditional methods.[3,39,40] One such promising alternative is the use of predictive models to estimate I-THM occurrence. Several statistical models have been developed to predict DBP formation across a wide range of environmental conditions.[38,41] These models can significantly streamline routine monitoring, minimize dependence on labor-intensive laboratory analyzes, and facilitate proactive risk management strategies to protect public health. However, while statistical models can provide valuable insights into DBP formation, their predictive power is often constrained by predefined assumptions and limited adaptability to complex, nonlinear relationships in data.[42] In this context, machine learning (ML) stands out as a powerful and innovative tool to address the inherent challenges associated with conventional DBP analysis.

In recent years, ML has been increasingly and successfully applied across various fields, including water treatment, due to its exceptional ability to extract intrinsic patterns and relationships from large data sets.[43−47] ML models are particularly well-suited for analyzing complex, nonlinear relationships, making them highly suitable for modeling intricate environmental processes.[48−50] Recent studies have demonstrated the feasibility of ML in addressing nonlinear regression problems related to DBP formation in drinking water.[3,6,42,51] However, these studies have predominantly focused on predicting the formation and identifying the contributing features for regulated DBPs, such as THMs and HAAs. To date, no ML studies have specifically investigated I-DBP formation in the presence of iodide (I^-) or iodinated contrast media (ICM) under varying operational conditions during chlorination or chloramination. Ersan et al. proposed a linear statistical model to predict I-THM formation.[38] While this work provided valuable insights, DBP formation is inherently complex, involving nonlinear interactions and mechanisms that linear models often fail to fully capture.[42] A recent study employed an ML model to predict I-DBP formation in the presence of I^-/IO_3^- during UV-chloramine sequential disinfection.[52] However, this study was limited in scope, focusing solely on the UV-chloramine system and utilizing a small data set (<100 samples). A comprehensive study is urgently needed to address these limitations. Such research should aim to investigate I-THM formation mechanisms across diverse disinfection systems, varied water matrices, and a wide range of operational conditions. The integration of ML in this context could transform predictive and analytical capabilities for I-DBPs, paving the way for more effective monitoring and control strategies in water treatment.

In this study, we employed data-driven ML methods to develop predictive models for I-THM formation in drinking water, accounting for diverse water qualities, experimental conditions, and disinfection parameters. The primary objective of this study was to demonstrate how ML algorithms can be leveraged not only to build accurate predictive models but also to uncover valuable insights into the mechanisms governing I-THM formation. To this end, we evaluated the predictive performance of five tree-based ensemble models. We extensively investigated the impact of feature selection on model performance. By incorporating domain knowledge, we enhanced model accuracy and interpretability, while feature elimination effectively reduced less significant variables, simplifying the model and minimizing noise without compromising performance. Given the resource- and time-intensive nature of measuring DBPs, this ML-driven approach aims to identify alternative, readily measurable parameters that can reliably predict I-THM concentrations. By eliminating unnecessary features, the framework can save significant time and resources. To investigate the combined effects of domain knowledge integration and feature screening on model performance, we developed six distinct model configurations that incorporated feature elimination alongside domain knowledge. The performance of these configurations was compared to assessing the impact of losing features on predictive accuracy. To interpret and analyze the models, we employed SHapley Additive exPlanations (SHAP) values and Partial Dependence Plots (PDPs), enabling a quantitative evaluation of the contributions of input variables to I-THM formation. These analyzes identified key factors influencing I-THM formation and offered critical insights into their emergence.

This study provides a comprehensive and fundamentally novel assessment of I-THM formation through the application of advanced ML methods, addressing critical research gaps previously overlooked in the field. Unlike earlier studies that focused primarily on regulated THMs or were constrained by small, homogeneous data sets, single iodine sources, or

https://doi.org/10.1021/acs.est.5c05409
Environ. Sci. Technol. XXXX, XXX, XXX−XXX

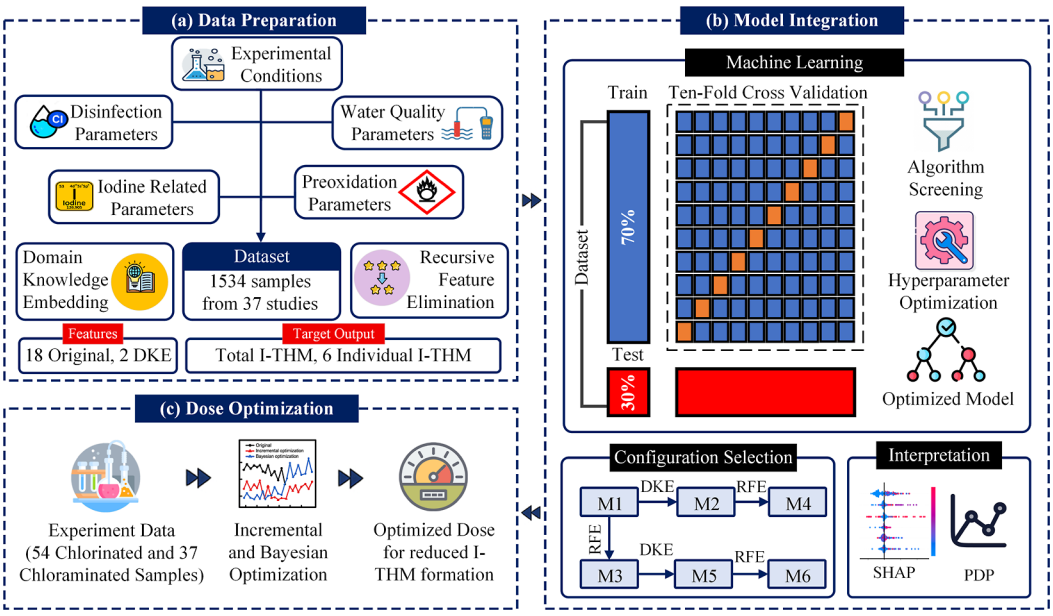


Figure 1. Overview of the study workflow for minimizing I-DBPs using machine learning. (a) Data Preparation: Integration of data from published studies, including disinfection, iodine-related, water quality, and preoxidation parameters, enhanced by domain knowledge embedding (DKE) and refined through recursive feature elimination (RFE). (b) Model Integration: Machine learning pipeline involving data set partitioning, algorithm screening, 10-fold cross-validation, and hyperparameter optimization to generate an optimized predictive model. Configuration selection incorporates RFE and DKE. Model interpretation is conducted using SHAP and PDP. (c) Dose Optimization: Application of the optimized model to experimental data for predicting lower I-DBP levels and guiding optimized dose selection.

simplistic linear modeling, this research pioneers the use of an extensive, diverse data set derived from a wide range of water matrices, multiple iodine sources (both iodide and ICM), and varied disinfection scenarios (chlorination and chloramination). Critically, this represents the first ML-driven effort explicitly tailored to address the unique chemical complexity and higher toxicity of I-THMs. Additionally, we introduce novel, chemically informed features, such as iodine/DOC and oxidant/DOC ratios, strategically embedding domain knowledge into the modeling framework, significantly enhancing both predictive accuracy and interpretability. By employing recursive feature elimination, a methodological advancement not previously applied to I-THMs, our approach demonstrates that numerous traditional analytical parameters can be eliminated without compromising predictive performance. This innovative reduction greatly streamlines the modeling process, substantially reduces analytical demands, and facilitates practical implementation in real-world water treatment settings. Ultimately, the novel insights and actionable optimization strategies generated by this research set a new standard in ML-driven environmental research, establishing ML not only as a predictive tool but as a proactive and prescriptive framework to mitigate emerging DBP threats and protect public health.

2. METHODOLOGY

2.1. Data Collection and Preprocessing. The data set was compiled from 37 peer-reviewed studies (2006−2022, Table S1), involving a variety of natural and engineered water sources, including surface water, groundwater, algae organic matter (AOM), and wastewater. These samples were treated under various disinfection strategies such as chlorination, preformed and in situ chloramination, and preoxidation using agents like ozone, $KMnO_4$, Fe(VI), and ClO_2. A total of 1534 data points were collected, capturing key parameters known to influence I-THM formation: pH, DOC, UV_{254}, oxidant doses, iodine concentration, and the type of ICMs (Table S2). Data were extracted from tables or digitized from figures using the GRABIT tool in MATLAB. Studies with incomplete or unconventional data (e.g., use of nonstandard disinfection methods or synthetic water matrices) were excluded. Data distribution for each feature is shown in Figure S2.

Categorical variables were standardized (e.g., water source categories, buffer types), and ICM concentrations were converted to equivalent iodine concentrations. Missing values in numerical variables were imputed using the Multivariate Imputation by Chained Equations (MICE) algorithm with Extremely Randomized Trees (ETR) to maintain data integrity. Categorical variables were ordinally encoded (Table S3). For 118 samples lacking individual I-THM species concentrations, values were estimated using ratio-based inferences from complete data sets. Pearson correlation analysis indicated no strong multicollinearity among features, so all were retained for modeling (Figure S3). Five machine learning models were initially tested using default hyperparameters. The best-performing model, selected based on R^2 and RMSE metrics, was further optimized using Bayesian Optimization, enabling more efficient and effective hyperparameter tuning than conventional methods such as grid or random search.

The details on data collection, preprocessing, hyperparameter tuning, and model interpretation are provided in Text S1−S5. All data preprocessing, analysis, ML model development, and interpretation were performed in Python (version 3.12), with the codes executed in Jupyter Notebook (version 6.5.4). The entire methodology is illustrated in Figure 1.

2.2. Model Development and Evaluation. We evaluated various ensemble tree-based models for predicting I-THM concentrations in water samples. Previous ML studies on DBP

https://doi.org/10.1021/acs.est.5c05409
Environ. Sci. Technol. XXXX, XXX, XXX−XXX

formation have highlighted the effectiveness of ensemble models in capturing the complex relationships between water quality parameters and DBP concentrations, which are often beyond the capability of linear models.[3,6] Tree-based models often outperform neural networks on tabular data sets with individually meaningful features.[53] Therefore, we used five powerful and widely used ensemble tree-based models in this study, including Random Forest Regression (RF), Light Gradient Boosting Machine Regression (LGB), Categorical Boosting (CatBoost) Regression (CAT), Gradient Boosted Regression Trees (GBRT), and Extreme Gradient Boosting (XGBoost) Regression (XGB) for the total I-THM concentration prediction.

We employed a 70−30 cross-validation and testing (CVT) split method to ensure robust model evaluation. The entire data set was randomly partitioned into training (70%) and test (30%) sets, with this process repeated 10 times to account for variability in data set splits. Subsequently, 10-fold cross-validation was conducted on the training sets during model construction. In this process, the training data was further divided into ten subsets; in each iteration, nine subsets were used for training the model, while the remaining subset served as a validation set. This approach allowed comprehensive model training and validation. Finally, the models were evaluated on the independent test data sets to assess their generalization performance. We did not apply any feature scaling methods, such as min−max normalization or z-score standardization, as tree-based ensemble models are inherently robust to variations in feature scales.[54,55] The model performance was evaluated on the test data set based on the coefficient of determination (R^2) and root mean square error (RMSE). R^2 and RMSE were calculated using eqs 1 & 2

$$R^2 = 1 - \frac{\sum_{i=1}^{K}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{K}(y_i - \bar{y})^2} \quad (1)$$

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{i=1}^{K}(y_i - \hat{y}_i)^2}{K}} \quad (2)$$

Here, y_i represents the actual value, $\hat{y}_i$ is the predicted value, and $\bar{y}$ is the average of actual value. A higher R^2 value and a lower RMSE indicate superior model predictive performance. Models were selected based on their performance on the test set, prioritizing those with higher R^2 and lower RMSE values. Initially, all ML models were evaluated using their default hyperparameters. The best-performing model, based on the highest R^2 and lowest RMSE, was selected for further hyperparameter tuning. We employed Bayesian optimization to identify the optimal hyperparameters for the best-performing model, ensuring maximum predictive accuracy and training efficiency.

2.3. Feature Selection. Feature selection plays a pivotal role in ML, as the choice of relevant features can significantly influence a model's effectiveness. In this study, we initially developed 18 input features categorized into five groups: preoxidation, disinfection, iodine source, water quality, and experimental conditions (Table S2). These features were carefully chosen for their strong relevance to I-THM formation, ensuring a robust foundation for predictive modeling. To enhance the model with domain-specific knowledge before training, we derived two additional features from the data set. Domain-specific features often uncover relationships and patterns that generic algorithms might otherwise miss, enriching both the model's predictive performance and interpretability. Previous research has demonstrated that embedding domain knowledge improves ML model accuracy and clarity.[56−58] In our study, we generated two derived features: the oxidant dose normalized by dissolved organic carbon (Chlorine or Monochloramine Dose/DOC) and iodine dose normalized by dissolved organic carbon (Iodine/DOC). The oxidant dose accounted for both chlorine and monochloramine concentrations, while the iodine dose included contributions from both iodide and ICM. This approach aligns with findings in a related study on DBPs, which identified Chlorine Dose/DOC as a critical predictor for THM formation.[3] Given that our data set focuses on I-THMs, we aimed to assess the influence of both Oxidant Dose/DOC and Iodine/DOC ratios on I-THM formation.

To further evaluate the model's effectiveness, we assessed its performance with a reduced feature set. Reducing the number of features in ML models not only reduces complexity but can also improve generalization while maintaining strong performance.[59] While prior ML studies often include extensive features to maximize accuracy, they frequently overlook the considerable effort required for experimental measurement and analysis, particularly in water treatment applications.[46,60] This oversight is especially relevant in the study of DBPs, where experiments are inherently time-intensive and resource-demanding. By eliminating noninformative or less critical features, researchers can achieve significant practical advantages, including streamlined workflows, reduced costs, and minimized efforts for experimental analysis of nonessential parameters. These benefits enable more efficient and cost-effective water treatment practices. Therefore, we employed recursive feature elimination (RFE) to systematically reduce the number of features while maintaining high model performance. RFE is a well-established wrapper method that iteratively builds models and removes the least important features based on model coefficients.[61] This process continues until the optimal subset of features is identified.[62] The selected features are then ranked according to their importance in predicting the outcome. Previous studies have demonstrated the effectiveness of RFE when applied to CatBoost models,[61−63] further validating its utility in our analysis. To assess the impact of domain-specific features and feature reduction on model performance and interpretability, we constructed six different model configurations. These configurations varied in their inclusion of new domain-derived features and the exclusion of features through RFE (Table S4). For deeper analysis, we focused on the best- and worst-performing configurations, examining how feature inclusion or removal influenced importance rankings and predictive accuracy. This analysis offers valuable insights into the trade-offs between model complexity, interpretability, and performance. By integrating domain-specific knowledge with RFE, we successfully optimized both the accuracy and interpretability of our I-THM formation prediction model.

3. RESULTS AND DISCUSSION

3.1. Characteristics of Collected Data. Descriptive statistics for the input and output features of the compiled data set are presented in Figure S2. The figure illustrates the distribution of key water quality parameters, disinfection conditions, and I-THM concentrations, highlighting the substantial variability across the data set. Many parameters

https://doi.org/10.1021/acs.est.5c05409
Environ. Sci. Technol. XXXX, XXX, XXX−XXX

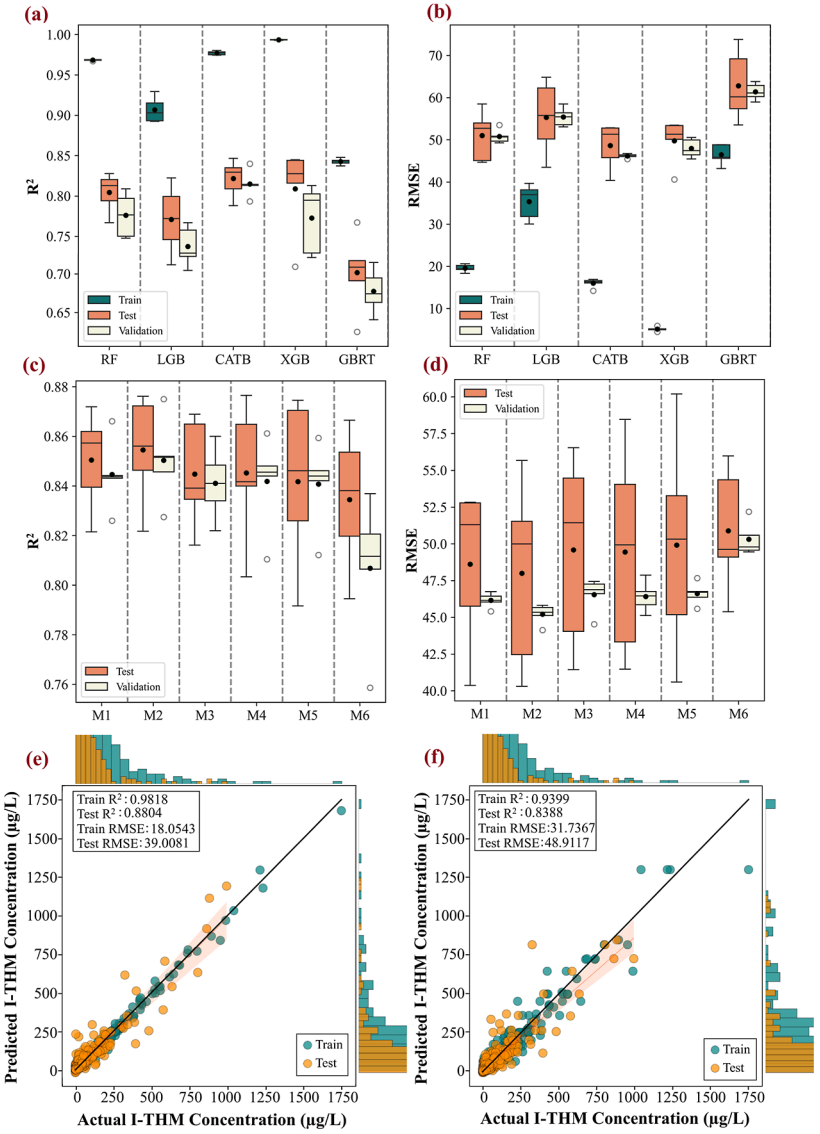


Figure 2. Predictive performance R^2 and RMSE of different ML models and configurations. (a) R^2 of different models; (b) RMSE of different models; (c) R^2 of different model configurations; and (d) RMSE of different model configurations for total I-THM concentration prediction. The lines within the boxes are the median values, the black dots represent the mean value, the tops and bottoms of the boxes are the 75th and 25th percentile values, respectively, the error bars extend to the 90th and 10th percentiles, and the circles indicate outlier points. Predictive performance of the optimized CAT model configurations after hyperparameter tuning (e) M2 Configuration; and (f) M6 Configuration. The black solid line represents the perfect prediction ($y = x$). The orange shade represents the 95% confidence intervals on the regression line for the test set.

display right-skewed distributions, characterized by a high concentration of lower values and a few pronounced outliers. For instance, both chlorine and chloramine doses are skewed toward lower concentrations, indicating that the majority of experimental setups utilized relatively modest disinfectant levels. However, the presence of high outliers in bromide concentration (e.g., >20 mg/L) suggests that certain studies included elevated bromide levels, which are known to enhance the formation of brominated DBPs.

The iodine source is treated as a categorical variable and exhibits a bimodal distribution, corresponding to iodide and iopamidol, reflecting the inclusion of both inorganic iodide and ICM as sources in the data set. Iodine concentrations exhibit a skewed distribution, with a few studies applying substantially higher doses (e.g., >10 mg/L as iodine). Water quality indicators such as pH (5−9), DOC (0−10 mg/L), UV_{254} (0−0.5 cm^{-1}) and SUVA (0−10 L/mg·m) span a broad range, reflecting the heterogeneity of the water sources examined. In particular, the variation in DOC and UV_{254} values suggests substantial differences in organic matter content and characteristics, which are critical drivers of DBP formation potential. Reaction time (0−96 h) and preoxidation time (0−2000 min) also show wide variability. In contrast, temperature displays a relatively narrow distribution (20−25 °C), indicating that most experiments were conducted within a typical laboratory range (e.g., room temperature).

The distributions of individual and total I-THM species (Figures S2s−2y) reveal that most samples contain relatively low concentrations. However, a notable number of samples (approximately 10−15% of the samples with concentrations exceeding 100 μg/L or higher) shows significantly elevated levels, suggesting that certain disinfection conditions can markedly increase I-THM formation. The total I-THM concentration follows a similar pattern, with most values

https://doi.org/10.1021/acs.est.5c05409
Environ. Sci. Technol. XXXX, XXX, XXX−XXX

clustered at the lower end but some samples exhibiting exceptionally high totals. These findings underscore the diversity of experimental conditions captured in the data set and highlight the importance of modeling to understand the influence of key factors on I-THM formation.

3.2. Model Screening and Evaluation. Among the five models evaluated, all demonstrated strong predictive performance for I-THM concentration, as evidenced by their results on both training and test data sets. The performance of the ML algorithms in predicting I-THM is illustrated in Figure 2a,b and Table S5. In general, all models exhibited decent performance (R^2 > 0.7), highlighting the effectiveness of tree-based ensemble methods. Among all the models, the CatBoost Regression (CAT) model delivered superior results (R^2 = 0.8504 and RMSE = 48.61104 in test data set), followed by RF and XGB. CAT employs ordered boosting and permutation-based techniques to mitigate overfitting and target leakage, which is particularly beneficial for moderately sized data sets containing both continuous and categorical variables like the one used in this study.[64] A key strength of CAT lies in its ability to model nonlinear, high-dimensional interactions, which are essential for accurately predicting I-THM formation, a process governed by complex relationships among iodine content, oxidant dose, bromide concentration, UV_{254}, DOC, and various experimental conditions. Fundamentally, CAT combines multiple decision trees in a gradient boosting framework, where each tree is constructed to iteratively correct the residual errors of the previous ones.[65] This sequential error correction process minimizes prediction error and enhances generalization to unseen data. After all trees are trained, their outputs are aggregated, weighted by a learning rate, to produce final predictions.[66] These design principles, along with CAT's resilience to multicollinearity and superior performance on heterogeneous tabular data, likely gave it a distinct advantage over models like XGB and LGB, which required more tuning and preprocessing to achieve similar accuracy. Given its highest average R^2 and lowest average RMSE, the CAT model was selected for further optimization. To refine this approach further, multiple model configurations were developed, incorporating domain knowledge and feature elimination strategies.

3.3. Feature Selection in Model Design. The results of six distinct model configurations (Table S4) are summarized in Figure 2c,d and Table S6. The first configuration, M1, utilized all the original 18 input features (Table S2), serving as the baseline model. Configuration M2 expanded upon M1 by incorporating two additional domain-knowledge-driven features. Using RFE, six features (preoxidant, preoxidation time, preoxidant dose, buffer, temperature, and water source) were removed from M1 and M2, resulting in configurations M3 and M4, respectively. To further explore optimization, two domain knowledge-based features were added to M3, creating M5. RFE was then applied to M5, eliminating three more features (i.e., pH, iodine source, and DOC), to produce the final configuration, M6, which featured the least number of input variables. Despite the substantial reduction in features, all model configurations exhibited comparable performance, with test R^2 values consistently ranging between 0.83 and 0.86 (Figure 2c). Among the configurations, M2 achieved the highest performance, while M6 demonstrated the lowest. However, the performance differences across configurations were minimal and not statistically significant (ANOVA F: 0.35, p value: 0.88 > 0.05). Interestingly, M3 and M4 showed performance nearly identical to M1 and M2, despite having approximately one-third of their features removed. This finding suggests that the eliminated features contribute minimally to I-THM formation and can be excluded from experimental workflows, significantly reducing the time, resources, and experimental efforts required for data collection and analysis. For example, the use of preoxidation and the measurement of temperature may not be necessary for precise prediction in M3 and M4, highlighting how specific experimental efforts can be minimized without compromising predictive accuracy. The addition of domain knowledge-driven features in M2 slightly improved its performance over M1, emphasizing the value of incorporating expert insights into feature design.[56−58] While M3, M4, and M5 exhibited nearly identical performance, the aggressive feature reduction in M6 resulted in a minor decline, indicating that the three final features removed do have some contribution to model accuracy. Nevertheless, M6 significantly reduced the number of features while maintaining decent model performance, underscoring its potential to minimize experimental efforts required for analyzing nonessential features. Model M6 is highly advantageous as it eliminates several features that typically require extensive laboratory analysis, including the preoxidation process, temperature, pH, and iodine source (Table S4). For instance, measuring temperature and pH becomes not necessary, and distinguishing between iodine sources (e.g., ICM or iodide) is no longer required, significantly simplifying experimental workflows.

To enhance model performance, the hyperparameters for the M2 and M6 models were further optimized using Bayesian optimization, based on Bayes' theorem.[67] This approach begins with an initial assumption about the objective function and iteratively refines this guess by incorporating information from previous samples. A utility function is then used to select the next sampling point, with the goal of maximizing the optimization function.[68] We also built models for the individual I-THM species using the highest performing M2 model configuration. The optimal hyperparameters identified through this process are summarized in Table S7. After tuning the hyperparameters, the models were retrained with optimized settings, and their performance was evaluated on the test data. The results of the optimized CAT model configurations (M2 and M6) are shown in Figure 2e,f. Compared to the preoptimization results (Figure 2c,d), the test R^2 values improved slightly from 0.854 to 0.88 for the M2 configuration and from 0.833 to 0.838 for the M6 configuration. Additionally, the RMSE decreased by 18% and 4% for M2 and M6 configurations, respectively, which helped mitigate overfitting and increased the reliability of the models for prediction.[46,60] The performance improvement for the M6 configuration after hyperparameter tuning was relatively modest (R^2 increased from 0.833 to 0.838; RMSE decreased by ∼ 4%). This is expected given that M6 was designed as a highly simplified model, containing only a minimal set of the most influential features. With fewer input parameters available, the model had limited flexibility for further performance enhancement during tuning.[69,70] In contrast, the M2 configuration retained a more comprehensive set of features, allowing for a greater degree of optimization and resulting in a more substantial improvement in both R^2 (from 0.854 to 0.88) and RMSE (18% reduction). These results highlight the trade-off between model simplicity and maximum predictive accuracy.

https://doi.org/10.1021/acs.est.5c05409
Environ. Sci. Technol. XXXX, XXX, XXX−XXX

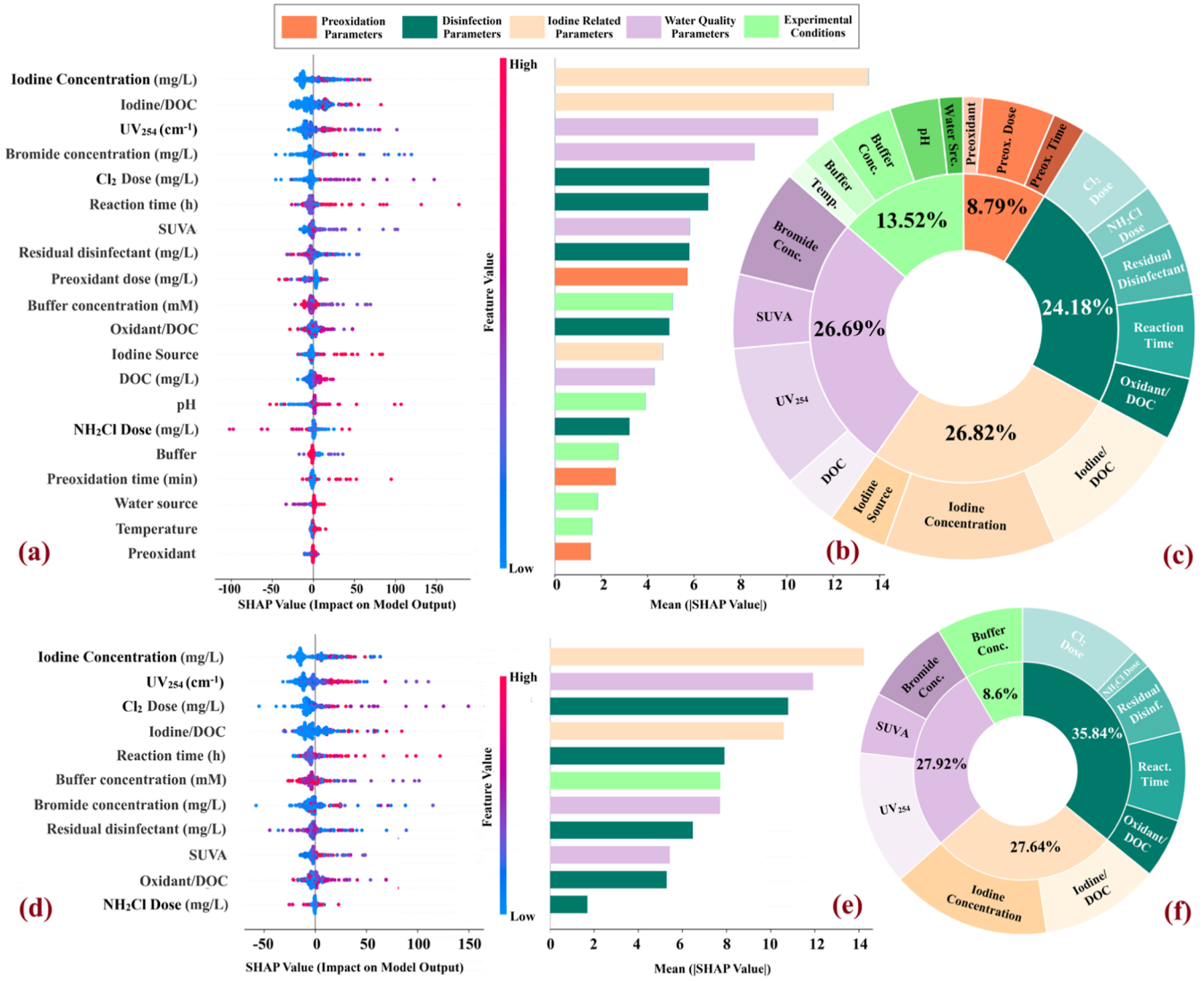


Figure 3. Explainable analysis of input features in the M2 and M6 model configurations. SHAP beeswarm plot showing the specific SHAP values of various input variables for predicting total I-THMs by the (a) M2 configuration and (d) M6 configuration. The *y*-axis represents the features, ordered based on the mean absolute SHAP value. The color bar indicates feature values, with red corresponding to higher values and blue to lower values. The features are color-coded from blue to red based on the SHAP value, where a positive or negative SHAP value indicates the potential to increase or decrease I-THM formation, respectively. Quantitative assessment of the contribution of individual and grouped input features to I-THM formation for (b) and (c) M2 configuration; and (e,f) for M6 configuration.

The prediction performance of the M2 model configuration for individual I-THM species, after hyperparameter optimization is presented in Table S8. The predictive performance of the M2 configuration varied among individual I-THM species. TIM achieved the highest test R^2 (0.8947), followed closely by BCIM (R^2 = 0.8719), CDIM (R^2 = 0.8553), and DBIM (R^2 = 0.8156). These results indicate strong model reliability for these species. In contrast, DCIM and BDIM showed relatively lower test R^2 values of 0.7627 and 0.7396, respectively, suggesting that these species may be more difficult to predict accurately, potentially due to their narrower formation conditions, lower concentration ranges, or greater sensitivity to variables not captured in the feature set. Additionally, the test RMSE values reflect the concentration ranges and units of each species, with TIM showing a higher RMSE due to its broader distribution in the data set. Overall, while all species were predicted with reasonable accuracy, the results suggest that I-THMs that are more frequently observed and have distinct chemical characteristics, such as TIM, tend to show better model performance.

3.4. Feature Importance Analysis. In machine learning, model interpretation is important for understanding the relationship between the target variable and the features, as well as the underlying mechanisms that drive the properties and behavior of the target.[46,60] This process is especially important for ensuring that model predictions align with fundamental domain knowledge and experimental insights. Moreover, interpreting the model provides valuable information on the key factors that drive I-THM formation, which can help identify critical areas for intervention. Ultimately, this understanding supports the development of targeted strategies to mitigate the formation of these DBPs and manage the associated health risks in drinking water. To explore the impact of domain-knowledge embedding on model interpretation, we applied the SHAP method to interpret both M2 and M6 model configurations after optimizing the hyperparameters. The

https://doi.org/10.1021/acs.est.5c05409
Environ. Sci. Technol. XXXX, XXX, XXX−XXX

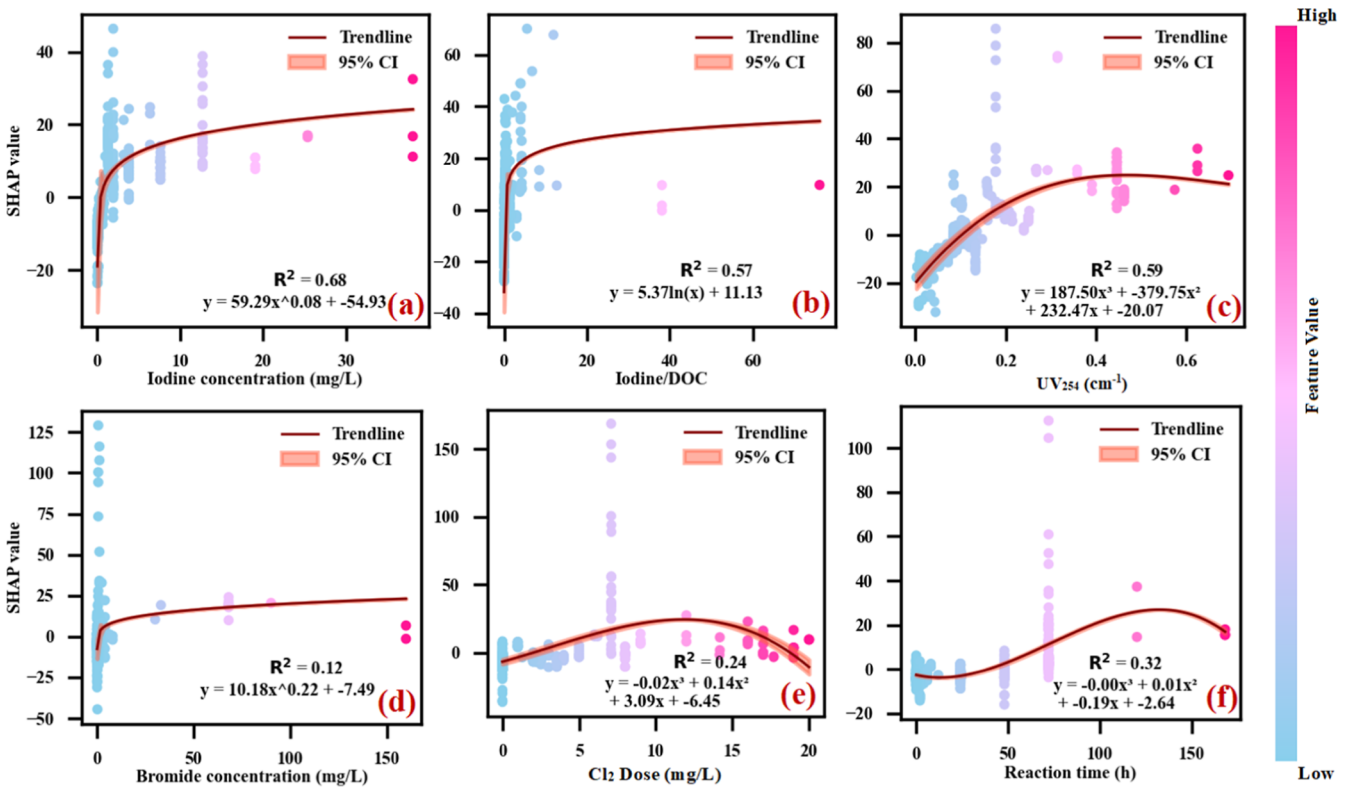


Figure 4. SHAP dependence plots for the top six most important features for the M2 model configuration with optimized hyperparameters (a) iodine concentration (mg/L); (b) iodine/DOC; (c) UV_{254} (cm^{-1}); (d) bromide concentration (mg/L); (e) Cl_2 dose (mg/L); and (f) reaction time. The *y*-axis represents the SHAP value for each feature, and *x* axis represents feature values. The color bar indicates feature values, with pink corresponding to higher values and sky blue to lower values. Trendlines, generated using LOESS (locally weighted regression), illustrate the average marginal effect of each feature. The orange shaded regions denote 95% confidence intervals for the trendlines. The widespread of SHAP values at given feature levels reflects interactions with other variables rather than model error, and even when trendline R^2 values are low, the plots still reveal meaningful directionality and conditional effects.

SHAP method, combined with PDP analysis, enabled a comprehensive understanding of how each input feature influences the model's predictions from multiple perspectives. By leveraging these interpretative tools, we gained deeper insights into the relative importance of features and the mechanisms driving I-THM concentrations, enhancing both the transparency and reliability of the model's decision-making process.

The feature importance ranking, derived from the SHAP plots for the optimized models, is shown in Figure 3. These plots illustrate both the distribution and the direction of each feature's contribution to the model's predictions. The ranking of features, as depicted on the *y*-axis (from top to bottom), indicates their relative importance (Figure 3a,d).[71] In the M2 configuration, the most important features identified in the SHAP plots include Iodine Concentration, Iodine/DOC, UV_{254}, Bromide Concentration, Cl_2 Dose, and Reaction Time (Figure 3a,b). The M6 configuration showed a similar rank in feature importance (Figure 3d,e). Buffer concentration was identified more important than bromide concentration, apart from that the topmost important features remained the same. The elevated importance of buffer concentration observed in the M6 configuration should be interpreted with caution, as it may result from correlations with other omitted features (e.g., pH or iodine source) rather than a direct mechanistic influence. While buffer systems are typically treated as experimental controls, their potential to influence oxidant stability or reaction kinetics suggests that this observation could warrant further targeted experimental investigation.

A notable finding from this study is that the M6 configuration, which incorporates only half of the original features, successfully captures a pattern similar to the M2 configuration and can serve as a reliable alternative, offering decent model performance (Figure 2f) while significantly reducing the time and efforts required for analyzing nonessential features experimentally. Moreover, one of the newly introduced features, Iodine/DOC, ranks highly in terms of feature importance, confirming our hypothesis that integrating domain knowledge enhances model interpretability. Similarly, the domain knowledge-driven feature, oxidant/DOC, remains relevant, retaining its positions in the 11th and 10th ranks in the M2 and M6 models, respectively. This underscores its contribution to the models' performance, despite the reduction in feature set in the M6 model.

The color bar in the SHAP beeswarm plot (Figure 3a,d) represents feature values, with blue indicating lower feature values and red signifying higher values.[72] Positive or negative SHAP values correspond to whether the features have a positive or negative impact on I-THM concentration.[6] For example, Iodine Concentration, Iodine/DOC, and UV_{254} exhibit high (red) values on the right side and low (blue) values on the left side, suggesting that increasing these features reduces I-THM concentration. This is consistent with the expected behavior of these disinfection parameters in controlling DBP formation.[24,73]

https://doi.org/10.1021/acs.est.5c05409
Environ. Sci. Technol. XXXX, XXX, XXX−XXX

The mean absolute SHAP values for each feature, as presented in Figure 3b,e, were utilized to calculate the categorical contributions of features to I-THM formation. In the M2 configuration (Figure 3c), iodine-related parameters emerged as the largest contributors, accounting for 26.82% of I-THM formation, which aligns with the established role of iodine in I-THM production.[20,74,75] Disinfection-related parameters exhibited a similar contribution of 26.69%, followed by water quality parameters at 24.18%. In contrast, the M6 configuration showed a shift in contributions, with disinfection parameters accounting for the highest share at 35.84%, while iodine-related and water quality parameters each contributed about 28% (Figure 3f). These findings highlight that disinfection parameters, iodine-related parameters, and water quality parameters are the most critical categories influencing I-THM formation. The increased importance of disinfection parameters in the M6 configuration reflects the reduced feature space, where fewer parameters were available, and the model placed greater emphasis on operational variables such as chlorine dose and reaction time to explain I-THM variability. However, both configurations consistently demonstrated the lowest contributions from preoxidation parameters and experimental conditions. This suggests that preoxidation plays a minimal role in controlling I-THM formation in water, a conclusion supported by earlier findings.[30,76] Specifically, configurations M3, M4, M5, and M6 excluded preoxidation parameters entirely yet achieved comparable model performance (Figure 2c,d).

The SHAP analysis and feature importance for individual I-THM species developed using the M2 model configuration is shown in Figure S4. The analysis reveals that iodine concentration, Cl_2 dose, and UV_{254} are consistently key features across all I-THM species (Figure S4a,b), aligning closely with the results from the total I-THM model (Figure 3a,b). Notably, bromide concentration emerged as a critical factor in the formation of brominated I-THMs, including DBIM, BDIM, and BCIM. This observation aligns with the well-established role of bromide in promoting the formation of brominated DBPs.[18,77] The sum of mean absolute SHAP values (Figure S4c) for each model also shows iodine concentration, UV_{254}, Cl_2 dose, bromide concentration, reaction time as key features for individual and total I-THM concentration prediction.

3.5. Interpreting Model Behavior. In addition to the above observation, the relationships between input parameters and total I-THM formation were further examined through SHAP dependence plots and PDPs. These analyzes provided insights into how each feature influenced the target variable. Figure 4 illustrates the SHAP dependence plots for the top six most important features identified by SHAP analysis in the M2 model configurations (Figure 3a). The SHAP dependence plots and PDPs show a similar trend for these key features, indicating consistency in the model's interpretation of feature influence and the robustness of the identified relationships between these features and the target variable. The SHAP dependence plots for the remaining features are provided in Figure S5, while PDPs for all features are presented in Figure S6. Additionally, SHAP dependence plots and partial PDPs for all features in the M6 configuration are presented in Figures S7 and S8. These results align closely with those from the M2 configuration, reinforcing the conclusion that M6 effectively captures the underlying relationships in I-THM formation. The resemblance between M6 and M2 highlights that a model with a reduced feature set can achieve comparable robustness and performance, further validating the generalizability and efficiency of the streamlined M6 configuration.

The dependence plot for iodine concentration reveals a higher I-THM formation potential with increasing iodine levels (Figures 4a & S6a). A significant portion of the samples in our data set with high iodine concentrations contained ICM, such as iopamidol, as the iodine source. ICM have a higher iodine content per molecule, and when their concentrations are converted into mg/L as iodine, they far exceed the levels contributed by iodide alone. Consequently, the elevated I-THM yields observed at high iodine concentrations (e.g., >5 mg/L) can largely be attributed to the presence of ICM. However, at lower iodine concentrations (e.g., <5 mg/L), where samples predominantly consist of iodide as the iodine source, a positive correlation is also evident. This indicates that higher iodide concentrations contribute to increased I-THM formation. During disinfection, iodide is oxidized by oxidants such as Cl_2 or NH_2Cl to HOI. HOI then reacts with electron-rich sites in organic molecules, leading to the formation of I-DBPs (Figure S1). Therefore, higher concentrations of iodide inevitably result in greater I-THM production, which aligns with previous studies.[24,73] The increment at iodine concentration >5 mg/L is sluggish, suggesting that other factors may limit I-THM formation at elevated iodine levels.

The observed increasing trend of the Iodine/DOC ratio (Figures 4b & S6b) aligns with findings reported by Jones et al.[18] The trendline shows an increase up to an Iodine/DOC ratio of 20, beyond which it starts to decline or flatten. When the iodine concentration is insufficient and the DOC concentration in the water is high (lower Iodine/DOC), the limited production of HOI causes DOC in the water to react with available oxidants, such as chlorine or chloramines. This reaction primarily leads to the formation of noniodinated DBPs, including chloro- and bromo-substituted THMs and HAAs. Conversely, in the presence of sufficient iodine, more HOI will be produced, facilitating its reaction with DOC to form I-DBPs. However, in the presence of excess iodine source, and low concentration of DOC (iodine/DOC > 20), the formation of I-DBPs can be limited due to insufficient DOC available for reaction.

The dependence plots for UV_{254} revealed a positive trend up to 0.4 cm^{-1} (Figures 4c and S6c), after which the trendline becomes flat. UV_{254}, a measure of the absorbance of UV light at 254 nm, serves as an indicator of aromatic contents in water.[78] These compounds, characterized by their electron-rich aromatic structures, are highly reactive during chlorination or iodination, providing reactive sites for halogenation that lead to I-THM formation. However, beyond a UV_{254} value of 0.4 cm^{-1}, the system may experience reduction in I-THM formation. At higher UV_{254} levels, the dominance of aromatic DOM may shift the reaction pathways toward the formation of other DBPs, such as brominated or chlorinated DBPs. These DBPs are favored because chlorine and bromine are more reactive in electrophilic aromatic substitution reactions than iodine.[79−81]

The bromide concentration dependence plots also exhibited an increasing trend up to approximately 10 mg/L (Figures 4d & S6d). Beyond this threshold, the curve started to decline or flatten. This indicates that bromide concentration promotes I-THM formation only up to a certain level, after which further increases in bromide concentration do not enhance I-THM production. This phenomenon may be attributed to excessive

https://doi.org/10.1021/acs.est.5c05409
Environ. Sci. Technol. XXXX, XXX, XXX−XXX

bromide concentrations competing with other reactants (e.g., iodine and DOC) for available oxidants to form brominated DBPs, such as brominated THMs. Previous studies showed that bromide concentrations appear to suppress I-DBP formation during chlorination or chloramination in the presence of DOM.[10,25,82−84] Moreover, formation of less reactive bromamines (bromochloramine) has been reported during the chloramination of bromide containing waters, which may further reduce the production of brominated DBPs due to their lower reactivity with DOM.[85,86]

The dependence plots for Cl_2 dose reveals a positive relationship with I-THM formation potential up to a threshold of 10 mg/L (Figures 4e & S6e). Chlorine, commonly introduced during water treatment as chlorine gas or sodium hypochlorite, reacts to form free chlorine, which exists as hypochlorous acid (HOCl, pK_a 7.5) and hypochlorite ion (OCl^-) depending on the pH.[6,87] These forms of free chlorine are highly reactive and serve as potent disinfectants, effectively eliminating pathogens in water. In the presence of iodide (I^-), chlorine reacts to form HOI, which promotes I-THM formation. ICM, such as iopamidol, can be oxidized by chlorine and release iodide into the water. The released iodide is further oxidized to HOI, which drives the formation of I-DBPs. Since iopamidol typically contains multiple iodine atoms, its degradation releases significant amounts of iodide, potentially elevating HOI levels. However, an excessive chlorine dose can shift the reaction toward the formation of iodate (IO_3^-), thereby reducing I-THM formation.[25,88]

I-THM formation trend observed in the dependence plots for reaction time demonstrates a rapid initial increase until 3−5 days, followed by a gradual slowdown (Figures 4f & S6f). This indicates that the I-THM formation occurs rapidly up to the first a few days. Previous studies have emphasized the significant impact of reaction time on I-DBP formation during both chlorination and chloramination. For example, Zhang et al. reported the highest I-DBP formation within the first 12 h of chlorinating natural surface water containing iodide.[25] Similarly, Wang et al. found that I-DBP formation rates from iodide and humic acids increased rapidly at the beginning of the reaction but gradually slowed and plateaued after 5 days.[74] Guo et al. observed differences in iodoform formation rates between natural organic matter and surface water, though both systems exhibited a rapid initial formation phase followed by a slower, gradual increase.[89] Additionally, Liu et al. reported a similar trend, with higher iodoform yields occurring early in the chlorination process, reaching a plateau after 2 days.[82] These findings align well with the findings from our result, reinforcing the trend of rapid I-DBP formation during the early stages of the reaction, followed by stabilization at later stages. One potential explanation for this trend is the depletion of precursors. As the reaction progresses, the concentrations of iodine and reactive organic precursors, such as humic substances, diminish as they are consumed. Another reason can be the rapid oxidation of iodide by chlorine or chloramines during the initial stages of the reaction, forming highly reactive iodine species (e.g., HOI and molecular iodine). These active iodine species quickly react with organic precursors to produce I-THMs.[28,90] Additionally, as the reaction progresses, HOI can be further oxidized in the presence of free chlorine to form iodate, an inert species that does not contribute to I-DBP formation.[24] This shift in iodine speciation, along with the gradual depletion of precursors, explains the observed stabilization of I-THM levels.

Among the other features, SUVA exhibited a positive correlation with I-THM formation up to a value of 4 L/(mg·m) (Figures S5a and S6g). Similarly, DOC displayed a positive trend, with I-THM formation increasing up to 7.5 mg/L (Figures S5g and S6m). ICM, particularly iopamidol, demonstrated a higher I-THM formation potential compared to iodide in water (Figures S5f and S6l). This finding demonstrates that the presence of ICM compounds has a more pronounced influence on I-THM formation than iodide, aligning with observations from previous studies.[20,21,74] The analysis of pH revealed a positive trend, indicating that lowering pH can help reduce I-THM formation (Figures S5h and S6n). This suggests that pH management can play a critical role in controlling I-THM formation by influencing the reactivity of precursors. Additionally, NH_2Cl dose also showed a positive correlation with I-THM formation potential, indicating that the use of chloramines as an alternative disinfectant may promote I-THM formation (Figures S5i and S6o).

Although preoxidation parameters ranked low in overall SHAP importance, this does not imply that preoxidation is mechanistically unimportant. Rather, it suggests that their influence is likely highly context-specific, depending on factors such as the type of oxidant used, the iodine source (iodide vs ICM), and the composition of the organic matrix. The SHAP dependence plots for preoxidation parameters offer additional insight into their nuanced role in I-THM formation. For example, the dependence plot for preoxidant dose (Figure S5c) reveals a nonlinear relationship with a weak upward trend at moderate doses, suggesting that higher doses may under certain conditions slightly increase I-THM formation. This effect may occur when preoxidation leads to partial oxidation of ICM, insufficient conversion of iodide to iodate due to short contact time or suboptimal pH, or when specific oxidants alter NOM reactivity in ways that do not suppress iodine incorporation. Similarly, the plot for preoxidation time (Figure S5k) shows a shallow hump, indicating that longer contact times may initially promote I-THM formation before declining potentially due to dynamic shifts in iodine speciation. The plot for preoxidant type (Figure S5n) also reveals slight variability in SHAP values, reinforcing that different oxidant (e.g., ozone, permanganate, chlorine dioxide) vary in their capacity to transform iodide to iodate or influence precursor reactivity. These observations support the broader conclusion that the effectiveness of preoxidation is both oxidant-specific and condition-dependent. While preoxidation can be a valuable strategy for I-THM control, particularly when iodide is the dominant iodine source, its standalone impact may be limited. Therefore, preoxidation should be implemented as part of an integrated treatment strategy, alongside enhanced DOC removal and optimized disinfection practices, to ensure robust I-THM mitigation across diverse water quality scenarios. In our data set, the diverse application of preoxidants across studies may have diluted their observable effect in the model. Additionally, preoxidation may exert indirect influences by modifying the reactivity of precursors, effects that could be captured through other features such as UV_{254} or iodine/DOC. Future research should explore targeted preoxidation strategies, ideally in conjunction with iodine speciation analysis, to better define the conditions under which preoxidants can effectively reduce I-THM formation.

https://doi.org/10.1021/acs.est.5c05409
Environ. Sci. Technol. XXXX, XXX, XXX−XXX

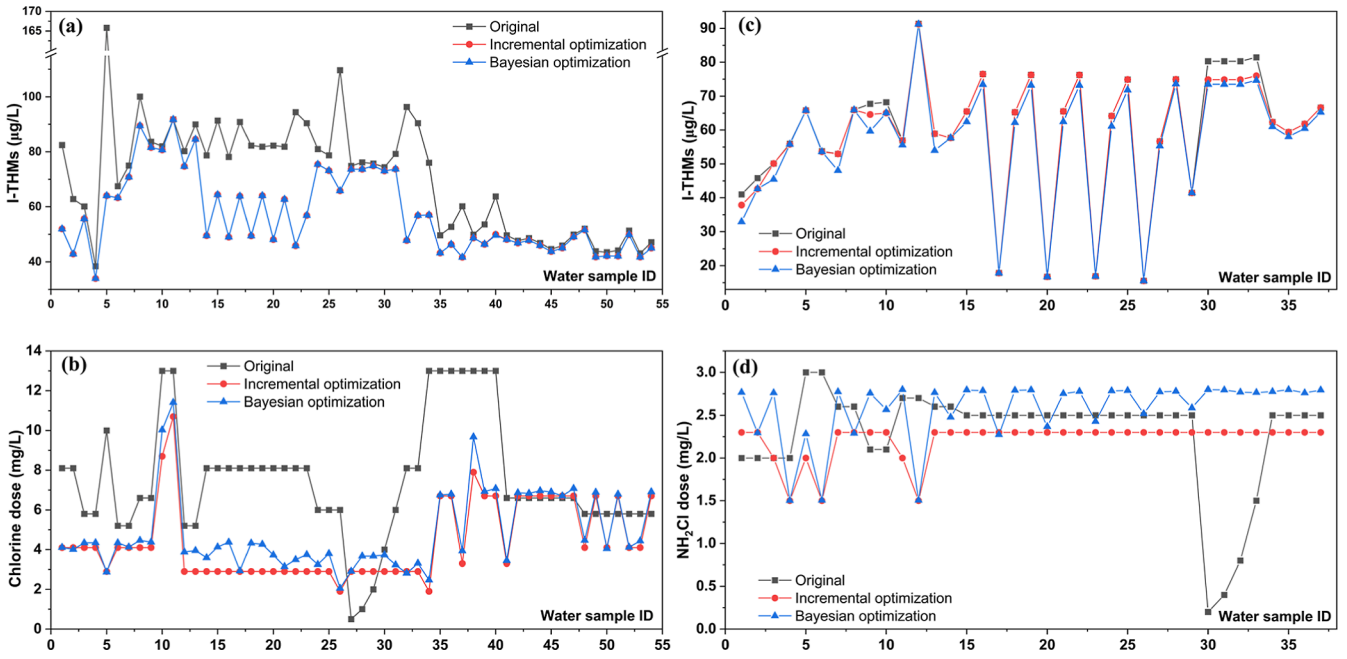


Figure 5. Optimization of chlorine and NH_2Cl doses (mg/L as Cl_2) for minimizing I-THMs by using the M2 model with iopamidol as the iodine source at 63 μg/L as iodine. (a) I-THM formation from 54 water samples. Original represents I-THM concentrations predicted using the M2 model with parameters from original experiments. Incremental optimization represents I-THM concentrations predicted using the M2 model with an optimized chlorine dose determined by incremental optimization (other parameters remain unchanged from Original). Bayesian optimization represents I-THM concentrations predicted using the M2 model with an optimized chlorine dose determined by Bayesian optimization (other parameters remain unchanged from Original). (b) Chlorine doses corresponding to the 54 water samples in (a). (c) and (d) represent predictions with similar methods to (a,b), except NH_2Cl was used instead of chlorine. For some samples (e.g., #35−40 in a and b), minimal reduction in I-THM concentration was observed despite lower chlorine doses, indicating that other factors such as high iodine or organic precursor levels are the primary drivers in those cases.

4. ENVIRONMENTAL IMPLICATIONS

From our ML findings, we recommend several strategies to minimize I-THM formation under typical drinking water treatment conditions. Reducing iodine concentration is crucial, as elevated iodine levels in source water enhance I-THM formation (Figure 3). Preoxidation processes, such as ozonation or permanganate addition, can convert iodide into iodate, a nontoxic iodine species. However, while preoxidation can mitigate iodide's direct contribution to I-THM formation by converting it to iodate,[31,91] it does not entirely eliminate the risk of I-THM formation, particularly in the presence of ICM.[30,76] This suggests that relying on preoxidation alone may not be a good solution in controlling I-THM formation and highlights the need for an integrated approach that combines preoxidation with other treatment strategies, such as enhanced DOC removal or optimized disinfection processes. Addressing the contribution of ICM (e.g., iopamidol), which exhibit a higher I-THM formation potential than iodide, is essential (Figures S5f and S6l). Implementing stricter wastewater treatment protocols and source control measures can help reduce ICM input into water systems, thereby lowering iodine concentration. Removing aromatic organic compounds, indicated by UV_{254} absorbance (Figures 4c & S6c) and SUVA (Figures S5a and S6g), will likely reduce I-THM precursors. Similarly, enhancing the removal of DOC during pretreatment can reduce I-THM formation (Figures S5g and S6m).[92,93] Additionally, optimizing the Iodine/DOC ratio to below 20 can further mitigate I-THM formation (Figures 4b and S6b). Controlling bromide concentrations in source water can also play an important role in reducing I-THMs, particularly brominated I-THMs (Figure S4). While elevated bromide levels promote competition between bromide and iodine for oxidants, maintaining a low level of bromide may help minimize I-THM formation (Figures 4d and S6d). These strategies collectively emphasize a comprehensive approach to minimizing I-THMs while ensuring the effectiveness of disinfection processes.

Chlorine dose, a readily adjustable parameter in water treatment, was identified as a significant factor influencing I-THM formation in our study (Figure 3a). To explore whether the established model (M2) could optimize chlorine dose to minimize I-THM formation, we conducted a series of experiments using M2 to predict I-THMs in 54 chlorinated water samples (Text S5). Through both incremental and Bayesian optimization, we successfully identified the optimal chlorine dose for each sample that minimized I-THM formation (Figures 5a,b and S9−S15). These results demonstrate that optimizing chlorine dose can serve as an effective strategy for I-THM mitigation, and a robust machine learning model (e.g., M2) can rapidly determine this optimal strategy, enabling efficient and data-driven decision-making in water treatment processes. In contrast, adjusting chloramine dose did not reduce I-THMs (Figures 5c,d and S9−S15). Avoiding the use of chloramines is likely to reduce I-THMs. However, if chloramines are necessary, employing combined chlorine and chloramine processes may minimize I-THM yields by converting iodine sources to iodate during prechlorination.[12,94,95] These findings underscore the potential of machine learning for predicting and understanding I-THM formation, as well as guiding drinking water treatment for I-THM control.

https://doi.org/10.1021/acs.est.5c05409
Environ. Sci. Technol. XXXX, XXX, XXX−XXX

To further evaluate the generalizability of the developed model beyond the original training and test data sets, we compiled an external validation set consisting of 26 independent water samples from previously published studies, including our own prior work (external validation data set in Supporting Information). These studies, detailed in Table S1, encompass a wide range of treatment conditions, such as peroxidation, chlorine, NH_2Cl, and various iodine sources and concentrations. Importantly, the studies used in this external set are completely distinct from those used in model development, ensuring a robust test of model transferability. Using the optimized M2 model configuration, we applied the trained model to this independent data set. The model achieved an R^2 of 0.67 and RMSE of 68.80 (Figure S16), indicating strong predictive performance even on unseen data. Cosine similarity measures directional agreement between predicted and actual values, and a value of 0.97 suggests that the model effectively captures the overall trend and variability of I-THM formation. These results highlight the model's potential to support chlorine-based disinfection decision-making, particularly in optimizing dosing strategies under diverse water quality conditions.

A notable limitation of this study is the lack of detailed DOM characterization beyond surrogate indicators like UV_{254} and SUVA. Future studies integrating fluorescence spectroscopy, molecular weight distribution, or high-resolution mass spectrometry could improve mechanistic interpretation by linking specific DOM fractions to I-THM precursors. Developing standardized data sets that include both I-THM species and DOM quality would greatly enhance model richness and environmental applicability. Future studies should also aim to experimentally validate the optimal chlorine and chloramine doses predicted by machine learning models through controlled bench-scale disinfection experiments. Such validation would strengthen confidence in the model's predictive capability and support its adoption for operational decision-making in real-world treatment scenarios.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.est.5c05409.

More details about data collection, data preprocessing, hyperparameter optimization, model interpretation, model application, additional results and discussion, and Supporting Information (PDF)

Data sets used in this study (XLSX)

■ AUTHOR INFORMATION

Corresponding Author

Tao Ye − *Department of Civil and Environmental Engineering, South Dakota School of Mines and Technology, Rapid City, South Dakota 57701, United States;* orcid.org/0000-0003-4768-1811; Email: Tao.Ye@sdsmt.edu

Authors

Md. Mahjib Hossain − *Department of Civil and Environmental Engineering, South Dakota School of Mines and Technology, Rapid City, South Dakota 57701, United States*

Rabbi Sikder − *Department of Civil and Environmental Engineering, South Dakota School of Mines and Technology, Rapid City, South Dakota 57701, United States;* orcid.org/0000-0001-8870-4958

Guanghui Hua − *Department of Civil and Environmental Engineering, South Dakota State University, Brookings, South Dakota 57007, United States*

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.est.5c05409

Author Contributions

#M.M.H. and R.S. equal contribution.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

The authors acknowledge helpful comments received from the anonymous, peer reviewers. This study is supported by the U.S. National Science Foundation grants (CBET-2308711 & 2308712). The authors also would like to thank the startup fund from the CEE department and the South Dakota School of Mines and Technology for supporting this study.

■ REFERENCES

(1) Richardson, S. D.; Postigo, C. Drinking Water Disinfection By-products. In *Emerging Organic Contaminants and Human Health*; Barceló, D., Ed.; Springer Berlin Heidelberg, 2012; pp 93−137.

(2) Hossain, M. M.; Ahmed, A.; Ashraf Ali, M. An assessment of fecal contamination of groundwater in Bangladesh. *AIP Conf. Proc.* **2023**, *2713* (1), 060005.

(3) Sikder, R.; Zhang, T.; Ye, T. Predicting THM Formation and Revealing Its Contributors in Drinking Water Treatment Using Machine Learning. *ACS EST Water* **2024**, *4* (3), 899−912.

(4) Richardson, S. D.; Plewa, M. J.; Wagner, E. D.; Schoeny, R.; DeMarini, D. M. Occurrence, genotoxicity, and carcinogenicity of regulated and emerging disinfection by-products in drinking water: A review and roadmap for research. *Mutat. Res., Rev. Mutat. Res.* **2007**, *636* (1), 178−242.

(5) Ye, T.; Xu, B.; Lin, Y.-L.; Hu, C.-Y.; Lin, L.; Zhang, T.-Y.; Gao, N.-Y. Formation of iodinated disinfection by-products during oxidation of iodide-containing waters with chlorine dioxide. *Water Res.* **2013**, *47* (9), 3006−3014.

(6) Yu, Y.; Hossain, M. M.; Sikder, R.; Qi, Z.; Huo, L.; Chen, R.; Dou, W.; Shi, B.; Ye, T. Exploring the potential of machine learning to understand the occurrence and health risks of haloacetic acids in a drinking water distribution system. *Sci. Total Environ.* **2024**, *951*, 175573.

(7) Council Directive. On the quality of water intended for human consumption. *Off. J. Eur. Communities: Inf. Not.* **1998**, *330*, 32−54.

(8) Environmental Protection Agency. National primary drinking water regulations: Stage 2 disinfectants and disinfection byproducts rule. *Fed. Regist.* **2006**, *71*, 388−493.

(9) Seidel, C. J.; McGuire, M. J.; Summers, R. S.; Via, S. Have utilities switched to chloramines? *J. AWWA* **2005**, *97* (10), 87−97.

(10) Ackerson, N. O. B.; Liberatore, H. K.; Richardson, S. D.; Plewa, M. J.; Ternes, T. A.; Duirk, S. E. Chloramination of iopamidol- and bromide-spiked waters containing natural organic matter. *Water Supply* **2021**, *21* (2), 886−898.

(11) Hong, Y.; Liu, S.; Song, H.; Karanfil, T. HAA formation during chloramination—significance of monochloramine's direct reaction with DOM. *J. AWWA* **2007**, *99* (8), 57−69.

(12) Jones, D. B.; Saglam, A.; Triger, A.; Song, H.; Karanfil, T. I-THM Formation and Speciation: Preformed Monochloramine versus Prechlorination Followed by Ammonia Addition. *Environ. Sci. Technol.* **2011**, *45* (24), 10429−10437.

(13) Allen, J. M.; Plewa, M. J.; Wagner, E. D.; Wei, X.; Bokenkamp, K.; Hur, K.; Jia, A.; Liberatore, H. K.; Lee, C.-F. T.; Shirkhani, R.; Krasner, S. W.; Richardson, S. D. Drivers of Disinfection Byproduct

Cytotoxicity in U.S. Drinking Water: Should Other DBPs Be Considered for Regulation? *Environ. Sci. Technol.* **2022**, *56* (1), 392−402.
(14) Dong, H.; Qiang, Z.; Richardson, S. D. Formation of Iodinated Disinfection Byproducts (I-DBPs) in Drinking Water: Emerging Concerns and Current Issues. *Acc. Chem. Res.* **2019**, *52* (4), 896−905.
(15) Postigo, C.; Zonja, B. Iodinated disinfection byproducts: Formation and concerns. *Current Opinion in Environmental Science & Health* **2019**, *7*, 19−25.
(16) Wei, X.; Wang, S.; Zheng, W.; Wang, X.; Liu, X.; Jiang, S.; Pi, J.; Zheng, Y.; He, G.; Qu, W. Drinking Water Disinfection Byproduct Iodoacetic Acid Induces Tumorigenic Transformation of NIH3T3 Cells. *Environ. Sci. Technol.* **2013**, *47* (11), 5913−5920.
(17) Bloodgood, M. A.; Chowdary, S. A.; Daiber, E. J.; Shi, H.; Granger, C. O.; Richardson, S. D. A balancing act: Optimizing free chlorine contact time to minimize iodo-DBPs, NDMA, and regulated DBPs in chloraminated drinking water. *J. Environ. Sci.* **2022**, *117*, 315−325.
(18) Jones, D. B.; Saglam, A.; Song, H.; Karanfil, T. The impact of bromide/iodide concentration and ratio on iodinated trihalomethane formation and speciation. *Water Res.* **2012**, *46* (1), 11−20.
(19) Li, C.; Lin, Q.; Dong, F.; Li, Y.; Luo, F.; Zhang, K. Formation of iodinated trihalomethanes during chlorination of amino acid in waters. *Chemosphere* **2019**, *217*, 355−363.
(20) Ye, T.; Xu, B.; Wang, Z.; Zhang, T.-Y.; Hu, C.-Y.; Lin, L.; Xia, S.-J.; Gao, N.-Y. Comparison of iodinated trihalomethanes formation during aqueous chlor(am)ination of different iodinated X-ray contrast media compounds in the presence of natural organic matter. *Water Res.* **2014**, *66*, 390−398.
(21) Duirk, S. E.; Lindell, C.; Cornelison, C. C.; Kormos, J.; Ternes, T. A.; Attene-Ramos, M.; Osiol, J.; Wagner, E. D.; Plewa, M. J.; Richardson, S. D. Formation of Toxic Iodinated Disinfection By-Products from Compounds Used in Medical Imaging. *Environ. Sci. Technol.* **2011**, *45* (16), 6845−6854.
(22) Christiansen, C. X-ray contrast media—an overview. *Toxicology* **2005**, *209* (2), 185−187.
(23) Steger-Hartmann, T.; Länge, R.; Schweinfurth, H. Environmental Risk Assessment for the Widely Used Iodinated X-Ray Contrast Agent Iopromide (Ultravist). *Ecotoxicol. Environ. Saf.* **1999**, *42* (3), 274−281.
(24) Pantelaki, I.; Voutsa, D. Formation of iodinated THMs during chlorination of water and wastewater in the presence of different iodine sources. *Sci. Total Environ.* **2018**, *613−614*, 389−397.
(25) Zhang, T.-Y.; Xu, B.; Hu, C.-Y.; Lin, Y.-L.; Lin, L.; Ye, T.; Tian, F.-X. A comparison of iodinated trihalomethane formation from chlorine, chlorine dioxide and potassium permanganate oxidation processes. *Water Res.* **2015**, *68*, 394−403.
(26) Zhu, Y.; Ling, Y.; Peng, Z.; Zhang, N. Formation of emerging iodinated disinfection by-products during ballast water treatment based on ozonation processes. *Sci. Total Environ.* **2020**, *743*, 140805.
(27) Zhang, H.; Dong, H.; Adams, C.; Qiang, Z.; Luan, G.; Wang, L. Formation and speciation of disinfection byproducts during chlor-(am)ination of aquarium seawater. *J. Environ. Sci.* **2015**, *33*, 116−124.
(28) Ye, T.; Xu, B.; Lin, Y.-L.; Hu, C.-Y.; Xia, S.-J.; Lin, L.; Mwakagenda, S. A.; Gao, N.-Y. Formation of iodinated disinfection by-products during oxidation of iodide-containing water with potassium permanganate. *J. Hazard. Mater.* **2012**, *241−242*, 348−354.
(29) Jones, D. B.; Song, H.; Karanfil, T. The effects of selected preoxidation strategies on I-THM formation and speciation. *Water Res.* **2012**, *46* (17), 5491−5498.
(30) Dong, H.; Qiang, Z.; Liu, S.; Li, J.; Yu, J.; Qu, J. Oxidation of iopamidol with ferrate (Fe(VI)): Kinetics and formation of toxic iodinated disinfection by-products. *Water Res.* **2018**, *130*, 200−207.
(31) Zhang, M.-S.; Xu, B.; Wang, Z.; Zhang, T.-Y.; Gao, N.-Y. Formation of iodinated trihalomethanes after ferrate pre-oxidation during chlorination and chloramination of iodide-containing water. *J. Taiwan Inst. Chem. Eng.* **2016**, *60*, 453−459.
(32) Rook, J. J. Formation of haloforms during chlorination of natural waters. *J. Soc. Water Treat. Exam.* **1974**, *23*, 234−243.
(33) Bellar, T. A.; Lichtenberg, J. J.; Kroner, R. C. The occurrence of organohalides in chlorinated drinking waters. *J. - Am. Water Works Assoc.* **1974**, *66* (12), 703−706.
(34) Wawryk, N. J. P.; Craven, C. B.; Blackstock, L. K. J.; Li, X.-F. New methods for identification of disinfection byproducts of toxicological relevance: Progress and future directions. *J. Environ. Sci.* **2021**, *99*, 151−159.
(35) Richardson, S. D.; Fasano, F.; Ellington, J. J.; Crumley, F. G.; Buettner, K. M.; Evans, J. J.; Blount, B. C.; Silva, L. K.; Waite, T. J.; Luther, G. W.; McKague, A. B.; Miltner, R. J.; Wagner, E. D.; Plewa, M. J. Occurrence and Mammalian Cell Toxicity of Iodinated Disinfection Byproducts in Drinking Water. *Environ. Sci. Technol.* **2008**, *42* (22), 8330−8338.
(36) Plewa, M. J.; Wagner, E. D.; Richardson, S. D.; Thruston, A. D.; Woo, Y.-T.; McKague, A. B. Chemical and Biological Characterization of Newly Discovered Iodoacid Drinking Water Disinfection Byproducts. *Environ. Sci. Technol.* **2004**, *38* (18), 4713−4722.
(37) Pan, Y.; Zhang, X.; Li, Y. Identification, toxicity and control of iodinated disinfection byproducts in cooking with simulated chlor-(am)inated tap water and iodized table salt. *Water Res.* **2016**, *88*, 60−68.
(38) Ersan, G.; Ersan, M. S.; Karanfil, T. Statistical modeling for iodinated trihalomethanes: Preformed chloramination versus pre-chlorination followed by ammonia addition. *Chemosphere* **2024**, *363*, 142876.
(39) Xu, S.; Hu, S.; Zhu, L.; Wang, W. Haloquinone Chloroimides as Toxic Disinfection Byproducts Identified in Drinking Water. *Environ. Sci. Technol.* **2021**, *55* (24), 16347−16357.
(40) Zhang, D.; Bond, T.; Pan, Y.; Li, M.; Luo, J.; Xiao, R.; Chu, W. Identification, Occurrence, and Cytotoxicity of Haloanilines: A New Class of Aromatic Nitrogenous Disinfection Byproducts in Chloraminated and Chlorinated Drinking Water. *Environ. Sci. Technol.* **2022**, *56* (7), 4132−4141.
(41) Ersan, G.; Ersan, M. S.; Kanan, A.; Karanfil, T. Predictive modeling of haloacetonitriles under uniform formation conditions. *Water Res.* **2021**, *201*, 117322.
(42) Peng, F.; Lu, Y.; Wang, Y.; Yang, L.; Yang, Z.; Li, H. Predicting the formation of disinfection by-products using multiple linear and machine learning regression. *J. Environ. Chem. Eng.* **2023**, *11* (5), 110612.
(43) Haggerty, R.; Sun, J.; Yu, H.; Li, Y. Application of machine learning in groundwater quality modeling - A comprehensive review. *Water Res.* **2023**, *233*, 119745.
(44) Huang, R.; Ma, C.; Ma, J.; Huangfu, X.; He, Q. Machine learning in natural and engineered water systems. *Water Res.* **2021**, *205*, 117666.
(45) Li, L.; Rong, S.; Wang, R.; Yu, S. Recent advances in artificial intelligence and machine learning for nonlinear relationship analysis and process control in drinking water treatment: A review. *Chem. Eng. J.* **2021**, *405*, 126673.
(46) Zhong, S.; Zhang, K.; Bagheri, M.; Burken, J. G.; Gu, A.; Li, B.; Ma, X.; Marrone, B. L.; Ren, Z. J.; Schrier, J.; Shi, W.; Tan, H.; Wang, T.; Wang, X.; Wong, B. M.; Xiao, X.; Yu, X.; Zhu, J.-J.; Zhang, H. Machine Learning: New Ideas and Tools in Environmental Science and Engineering. *Environ. Sci. Technol.* **2021**, *55* (19), 12741−12754.
(47) Rahman, M. H.-U.; Sikder, R.; Tonmoy, T. A.; Hossain, M. M.; Ye, T.; Aich, N.; Gadhamshetty, V. Transforming PFAS management: A critical review of machine learning applications for enhanced monitoring and treatment. *J. Water Process Eng.* **2025**, *70*, 106941.
(48) Zhong, S.; Hu, J.; Yu, X.; Zhang, H. Molecular image-convolutional neural network (CNN) assisted QSAR models for predicting contaminant reactivity toward OH radicals: Transfer learning, data augmentation and model interpretation. *Chem. Eng. J.* **2021**, *408*, 127998.
(49) Zhang, K.; Zhang, H. Machine Learning Modeling of Environmentally Relevant Chemical Reactions for Organic Compounds. *ACS EST Water* **2024**, *4* (3), 773−783.
(50) Kazi, M.-K.; Varghese, S.; Sarker, N.; Aich, N.; Gadhamshetty, V. Advancing PFAS Remediation through Physics-Based Modeling of

https://doi.org/10.1021/acs.est.5c05409
Environ. Sci. Technol. XXXX, XXX, XXX−XXX

2D Materials: Recent Progress, Challenges, and Opportunities. *Ind. Eng. Chem. Res.* **2025**, *64* (4), 1894−1906.
(51) Peleato, N. M. Application of convolutional neural networks for prediction of disinfection by-products. *Sci. Rep.* **2022**, *12* (1), 612.
(52) Luo, Z.-N.; He, H.; Zhang, T.-Y.; Wei, X.-L.; Dong, Z.-Y.; Xu, M.-Y.; Zhao, H.-X.; Zheng, Z.-X.; Pan, R.-J.; Hu, C.-Y.; Zeng, C.; El-Din, M. G.; Xu, B. Enhanced iodinated disinfection byproducts formation in iodide/iodate-containing water undergoing UV-chloramine sequential disinfection: Machine learning-aided identification of reaction mechanisms. *Water Res.* **2025**, *272*, 122975.
(53) Lundberg, S. M.; Erion, G.; Chen, H.; DeGrave, A.; Prutkin, J. M.; Nair, B.; Katz, R.; Himmelfarb, J.; Bansal, N.; Lee, S.-I. From local explanations to global understanding with explainable AI for trees. *Nat. Mach. Intell.* **2020**, *2* (1), 56−67.
(54) Ozsahin, D. U.; Mustapha, M. T.; Mubarak, A. S.; Ameen, Z. S.; Uzun, B. Impact of feature scaling on machine learning models for the diagnosis of diabetes. In *2022 International Conference on Artificial Intelligence in Everything (AIE), 2−4 Aug. 2022*, 2022, pp 87−94.
(55) Mahmud Sujon, K.; Binti Hassan, R.; Tusnia Towshi, Z.; Othman, M. A.; Abdus Samad, M.; Choi, K. When to Use Standardization and Normalization: Empirical Evidence From Machine Learning Models and XAI. *IEEE Access* **2024**, *12*, 135300−135314.
(56) Wang, H.; Zeng, J.; Dai, R.; Wang, Z. Understanding Rejection Mechanisms of Trace Organic Contaminants by Polyamide Membranes via Data-Knowledge Codriven Machine Learning. *Environ. Sci. Technol.* **2024**, *58* (13), 5878−5888.
(57) Yang, L.; Meng, X.; Karniadakis, G. E. B-PINNs: Bayesian physics-informed neural networks for forward and inverse PDE problems with noisy data. *J. Comput. Phys.* **2021**, *425*, 109913.
(58) Raissi, M.; Perdikaris, P.; Karniadakis, G. E. Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *J. Comput. Phys.* **2019**, *378*, 686−707.
(59) Guyon, I.; Elisseeff, A. An introduction to variable and feature selection. *J. Mach. Learn. Res.* **2003**, *3* (Mar), 1157−1182.
(60) Zhu, J.-J.; Yang, M.; Ren, Z. J. Machine Learning in Environmental Research: Common Pitfalls and Best Practices. *Environ. Sci. Technol.* **2023**, *57* (46), 17671−17689.
(61) Luo, M.; Wang, Y.; Xie, Y.; Zhou, L.; Qiao, J.; Qiu, S.; Sun, Y. Combination of Feature Selection and CatBoost for Prediction: The First Application to the Estimation of Aboveground Biomass. *Forests* **2021**, *12*, 216.
(62) Hadianto, A.; Utomo, W. H. CatBoost Optimization Using Recursive Feature Elimination. *J. Online Inf.* **2024**, *9* (2), 169−178.
(63) Khalid, N. H. M.; Ismail, A. R.; Aziz, N. A.; Hussin, A. A. A. Performance Comparison of Feature Selection Methods for Prediction in Medical Data. In *Soft Computing in Data Science*, Singapore, 2023, 2023; Yusoff, M., Hai, T., Kassim, M., Mohamed, A., Kita, E., Eds.; Springer Nature Singapore: pp 92−106.
(64) Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., Gulin, A. CatBoost: unbiased boosting with categorical features. *Advances in neural information processing systems* **2018**, *31*, 6638−6648
(65) Hancock, J. T.; Khoshgoftaar, T. M. CatBoost for big data: an interdisciplinary review. *J. Big Data* **2020**, *7* (1), 94.
(66) Dorogush, A. V.; Ershov, V.; Gulin, A. CatBoost: gradient boosting with categorical features support. *arXiv* **2018**, 1810.11363.
(67) Dewancker, I.; McCourt, M.; Clark, S. Bayesian optimization for machine learning: A practical guidebook. *arXiv* **2016**, 1612.04858.
(68) Frazier, P. I. A tutorial on Bayesian optimization. *arXiv* **2018**, 1807.02811.
(69) Bergstra, J.; Bengio, Y. Random search for hyper-parameter optimization. *J. Mach. Learn. Res.* **2012**, *13* (1), 281−305.
(70) Hutter, F.; Kotthoff, L.; Vanschoren, J. *Automated Machine Learning: Methods, Systems, Challenges*; Springer Nature, 2019.
(71) Marcilio, W. E.; Eler, D. M. From explanations to feature selection: assessing SHAP values as feature selection mechanism. In *2020 33rd SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI)*, 7−10 Nov. 2020, 2020; pp 340−347, .
(72) Li, Z.-q.; Liu, X.; Ning, Z.-l. Dynamic spectrum access based on deep reinforcement learning for multiple access in cognitive radio. *Phys. Commun.* **2022**, *54*, 101845.
(73) Postigo, C.; Richardson, S. D.; Barceló, D. Formation of iodo-trihalomethanes, iodo-haloacetic acids, and haloacetaldehydes during chlorination and chloramination of iodine containing waters in laboratory controlled reactions. *J. Environ. Sci.* **2017**, *58*, 127−134.
(74) Wang, Z.; Xu, B.; Lin, Y.-L.; Hu, C.-Y.; Tian, F.-X.; Zhang, T.-Y.; Gao, N.-Y. A comparison of iodinated trihalomethane formation from iodide and iopamidol in the presence of organic precursors during monochloramination. *Chem. Eng. J.* **2014**, *257*, 292−298.
(75) Bichsel, Y.; von Gunten, U. Oxidation of Iodide and Hypoiodous Acid in the Disinfection of Natural Waters. *Environ. Sci. Technol.* **1999**, *33* (22), 4040−4045.
(76) Li, M.; Zhang, T.-Y.; Xu, B.; Hu, C.-Y.; Dong, Z.-Y.; Wang, Z.; Tang, Y.-L.; Yu, S.-L.; Pan, Y.; Xian, Q. Iodinated trihalomethanes formation in iopamidol-contained water during ferrate/chlor(am)-ination treatment. *Chemosphere* **2021**, *272*, 129568.
(77) Liu, R.; Tian, C.; Hu, C.; Qi, Z.; Liu, H.; Qu, J. Effects of bromide on the formation and transformation of disinfection by-products during chlorination and chloramination. *Sci. Total Environ.* **2018**, *625*, 252−261.
(78) Weishaar, J. L.; Aiken, G. R.; Bergamaschi, B. A.; Fram, M. S.; Fujii, R.; Mopper, K. Evaluation of Specific Ultraviolet Absorbance as an Indicator of the Chemical Composition and Reactivity of Dissolved Organic Carbon. *Environ. Sci. Technol.* **2003**, *37* (20), 4702−4708.
(79) Criquet, J.; Allard, S. Influence of bromide and iodide on the formation of disinfection by-products in drinking water treatment. In *Comprehensive Analytical Chemistry*, Manasfi, T., Boudenne, J.-L., Eds.; Vol. *92*; Elsevier, 2021; pp 117−138, .
(80) Zhang, C.; Maness, J. C.; Cuthbertson, A. A.; Kimura, S. Y.; Liberatore, H. K.; Richardson, S. D.; Stanford, B. D.; Sun, M.; Knappe, D. R. U. Treating water containing elevated bromide and iodide levels with granular activated carbon and free chlorine: impacts on disinfection byproduct formation and calculated toxicity. *Environ. Sci.:Water Res. Technol.* **2020**, *6* (12), 3460−3475.
(81) Gao, Z.-C.; Lin, Y.-L.; Xu, B.; Xia, Y.; Hu, C.-Y.; Zhang, T.-Y.; Qian, H.; Cao, T.-C.; Gao, N.-Y. Effect of bromide and iodide on halogenated by-product formation from different organic precursors during UV/chlorine processes. *Water Res.* **2020**, *182*, 116035.
(82) Liu, S.; Li, Z.; Dong, H.; Goodman, B. A.; Qiang, Z. Formation of iodo-trihalomethanes, iodo-acetic acids, and iodo-acetamides during chloramination of iodide-containing waters: Factors influencing formation and reaction pathways. *J. Hazard. Mater.* **2017**, *321*, 28−36.
(83) Zhang, J.; Chen, D.-D.; Li, L.; Li, W.-W.; Mu, Y.; Yu, H.-Q. Role of NOM molecular size on iodo-trihalomethane formation during chlorination and chloramination. *Water Res.* **2016**, *102*, 533−541.
(84) Ackerson, N. O. B.; Liberatore, H. K.; Plewa, M. J.; Richardson, S. D.; Ternes, T. A.; Duirk, S. E. Disinfection byproducts and halogen-specific total organic halogen speciation in chlorinated source waters − The impact of iopamidol and bromide. *J. Environ. Sci.* **2020**, *89*, 90−101.
(85) Ersan, M. S.; Liu, C.; Amy, G.; Plewa, M. J.; Wagner, E. D.; Karanfil, T. Chloramination of iodide-containing waters: Formation of iodinated disinfection byproducts and toxicity correlation with total organic halides of treated waters. *Sci. Total Environ.* **2019**, *697*, 134142.
(86) Gazda, M.; Dejarme, L. E.; Choudhury, T. K.; Cooks, R. G.; Margerum, D. W. Mass-spectrometric evidence for the formation of bromochloramine and N-bromo-N-chloromethylamine in aqueous solution. *Environ. Sci. Technol.* **1993**, *27* (3), 557−561.
(87) Lau, S. S.; Abraham, S. M.; Roberts, A. L. Chlorination Revisited: Does Cl− Serve as a Catalyst in the Chlorination of Phenols? *Environ. Sci. Technol.* **2016**, *50* (24), 13291−13298.
(88) Criquet, J.; Allard, S.; Salhi, E.; Joll, C. A.; Heitz, A.; von Gunten, U. Iodate and Iodo-Trihalomethane Formation during

https://doi.org/10.1021/acs.est.5c05409
Environ. Sci. Technol. XXXX, XXX, XXX−XXX

Chlorination of Iodide-Containing Waters: Role of Bromide. *Environ. Sci. Technol.* **2012**, *46* (13), 7350−7357.
(89) Guo, W.; Shan, Y.; Yang, X. Factors affecting the formation of iodo-trihalomethanes during oxidation with chlorine dioxide. *J. Hazard. Mater.* **2014**, *264*, 91−97.
(90) Bichsel, Y.; von Gunten, U. Hypoiodous acid: kinetics of the buffer-catalyzed disproportionation. *Water Res.* **2000**, *34* (12), 3197−3203.
(91) Shin, J.; von Gunten, U.; Reckhow, D. A.; Allard, S.; Lee, Y. Reactions of Ferrate(VI) with Iodide and Hypoiodous Acid: Kinetics, Pathways, and Implications for the Fate of Iodine during Water Treatment. *Environ. Sci. Technol.* **2018**, *52* (13), 7458−7467.
(92) Golfinopoulos, S. K.; Nikolaou, A. D.; Alexakis, D. E. Innovative Approaches for Minimizing Disinfection Byproducts (DBPs) in Water Treatment: Challenges and Trends. *Appl. Sci.* **2024**, *14* (18), 8153.
(93) Pan, R.; Lin, Y.-L.; Zhang, T.-Y.; Wei, X.-L.; Dong, Z.-Y.; Hu, C.-Y.; Tang, Y.-L.; Xu, B. Sequential combination of pre-chlorination and powdered activated carbon adsorption on iodine removal and I-THMs control in drinking water. *Chemosphere* **2023**, *313*, 137529.
(94) Liu, C.; Ersan, M. S.; Plewa, M. J.; Amy, G.; Karanfil, T. Formation of iodinated trihalomethanes and noniodinated disinfection byproducts during chloramination of algal organic matter extracted from Microcystis aeruginosa. *Water Res.* **2019**, *162*, 115−126.
(95) Allard, S.; Tan, J.; Joll, C. A.; von Gunten, U. Mechanistic Study on the Formation of Cl-/Br-/I-Trihalomethanes during Chlorination/Chloramination Combined with a Theoretical Cytotoxicity Evaluation. *Environ. Sci. Technol.* **2015**, *49* (18), 11105−11114.

https://doi.org/10.1021/acs.est.5c05409
Environ. Sci. Technol. XXXX, XXX, XXX−XXX