Application of Convolutional Neural Networks for Prediction of 1 Disinfection By-Products

Nicolás M. Peleato (nicolas.peleato@ubc.ca)
University of British Columbia

Research Article

Keywords: disinfection by-products, fluorescence spectroscopy, neural networks, machine learning, 22 water quality, natural organic matter

Posted Date: August 12th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-806261/v1

Version of Record: A version of this preprint was published at Scientific Reports on January 12th, 2022.
See the published version at https://doi.org/10.1038/s41598-021-03881-w.

Application of Convolutional Neural Networks for Prediction of Disinfection By-Products

Nicolás M. Peleato[1,*]

[1] University of British Columbia Okanagan, School of Engineering

* Corresponding author: nicolas.peleato@ubc.ca, 1137 Alumni Ave., Kelowna, British Columbia, V1V 1V7

Abstract

Fluorescence spectroscopy can provide high-level chemical characterization and quantification that is suitable for the use in on-line process monitoring and control. However, the high-dimensionality of excitation-emission matrices and superposition of underlying signals is a major challenge to implementation. Herein the use of Convolutional Neural Networks (CNNs) is investigated to interpret fluorescence spectra and predict the formation of disinfection by-products during drinking water treatment. Using deep CNNs, mean absolute prediction error on a test set of data for total trihalomethanes, total haloacetic acids, and the major individual species were all < 6 µg/L and represent a significant difference improved by 39% - 62% compared to dense neural networks. Heat maps that identify spectral areas of importance for prediction showed unique humic-like and protein-like regions for individual disinfection by-product species that can be used to validate models and provide insight into precursor characteristics. The use of fluorescence spectroscopy coupled with deep CNNs shows promise to be used for rapid estimation of DBP formation potentials without the need of extensive data pre-processing or dimensionality reduction. Knowledge of DBP formation potentials in near real-time can enable tighter treatment controls and management efforts to minimize the exposure of the public to DBPs.

Keywords: disinfection by-products; fluorescence spectroscopy; neural networks; machine learning; water quality; natural organic matter

1 Introduction

The use of fluorescence spectra for improved water quality monitoring[1] and as a process analytical technology for bioprocesses, food, and pharmaceutical production, has become increasingly popular.[2,3] Fluorescence signatures are highly dependent on molecular structure, size, and environmental conditions, and therefore can be used to provide insight into chemical composition and properties.[4] The sensitivity and specificity of fluorescence analysis, coupled with the potential real-time monitoring capabilities, fluorescence has applicability to a wide variety of process control applications.[5]

One promising application is the improved prediction and control of disinfection by-product (DBP) formation from drinking water treatment with chlorine. Chlorination is the most common disinfectant used worldwide. However, when chlorine reacts with natural organic matter (NOM), present in all natural water sources, various by-products of health concern are formed.[6] Although many unique DBP species can be formed with varying public health risk, only specific groupings are commonly monitored and regulated in drinking water, including trihalomethanes (THMs) and haloacetic acids (HAAs). The monitoring frequency of regulated DBPs is generally low, with sampling only required once every 3 months for water systems in the United States and Canada.[7]

Significant attention has been paid to developing models that can predict DBP formation in order to improve knowledge of expected concentrations and inform water treatment operations of potential issues on a more frequent basis.[8–12] Since DBPs are formed from the reaction of chlorine and NOM, models must incorporate a measure of NOM. However, NOM is a chemically diverse grouping of organic

molecules whose characteristics are dependent on the surrounding environment. As such, the breadth of potential NOM characteristics and the spatial and temporal variability results in significant challenges in identifying an optimal measure that can capture this complexity and reactivity with chlorine.[13] Fluorescence spectroscopy has considerable potential for the prediction and monitoring of DBP precursor material. Many NOM compounds fluoresce and fluorescence measures can capture some chemical characteristics of NOM.[14] Previously, fluorescence has been used with success to predict or identify correlations with regulated DBPs,[15–17] as well as unregulated or by-products of emerging concern such as chloral hydrate[18] and haloacetonitriles.[19,20]

A common challenge to implementing fluorescence as a monitoring tool is the high-dimensionality and superposition of the resulting emissions. When utilizing fluorescence spectra collected at iterated excitation/emission wavelengths, a dimensionality reduction approach is often used to simplify excitation-emission matrices (EEMs).[14] By identifying a few underlying components that explain most of the variance in the data, the hypothesis is that noise is reduced, and subsequent modelling using a reduced dimensionality improves prediction. A basic simplification or dimensionality reduction approach would be to select peaks or regions in the fluorescence spectra where regional integration or peak fluorescence can be determined. While this type of expert guided approach has been used extensively in the past, discarding the majority of collected data neglects the richness of information contained. For complex systems such as those that include identifying natural organic matter (NOM) in water, organic fluorophores with similar chemical structures are not easily distinguished in the spectra. The use of principal component analysis (PCA) or parallel factors analysis (PARAFAC) has revealed underlying signals resembling fluorophores, which can be tied to spectral regions from which chemical properties can be inferred.[14,21,22] These analysis approaches are often limited to linear dimensionality reduction, so non-linear features such as Rayleigh or Raman scattering need to be removed from the spectra.[23] Furthermore, potential impacts of environmental conditions such as pH or temperature,[4] or possible charge-transfer interactions[24] may invalidate the assumption of a linear relationship between fluorophore concentrations and fluorescence intensity. Inner filter effects are also prevalent, where incident excitation light and emitted fluorescence is quenched by other chromophores present in the sample, result in a non-linear intensity response.[25] Constraints imposed by the method can be helpful when derived from prior knowledge of the system, such as non-negativity of fluorescence emissions, reducing bias and possibly resulting in a more accurate depiction of underlying structures. However, these same constraints may limit the overall accuracy of reconstruction based on the condensed representation.[26]

It may be advantageous to directly use all data collected in fluorescence EEMs to limit potential errors introduced from dimensionality reduction. However, for water quality analysis, there have been limited studies that explore the use of full fluorescence EEMs without dimensionality reduction. Non-linear regression using high-dimensional inputs can be accomplished using neural networks. More recent work with Convolutional Neural Networks (CNNs or ConvNets) has shown this type of network structure is well suited to interpreting images or other tasks datasets with local groups of values that are highly correlated.[27] Instead of training weighted connections between every individual node, CNNs train spatial filters or kernels to identify small recurring features in the input space. The use of filters allows for parameter sharing, where trained weights are used throughout the input space and are not tied to specific input nodes, giving rise to spatial invariance of features.[28] Furthermore, CNNs typically employ pooling layers where outputs in specific locations are merged with nearby outputs, creating invariance to small distortions in the input and reducing the dimensionality of the representation.[27,28] CNNs have been successfully applied in chemometric applications such as interpreting Raman and mid-infrared spectra for identifying *Escherichia coli* and meats,[29] pharmaceuticals in tablets with near infrared spectra,[30] categorizing wines using infrared spectra,[31] and classification of manganese valence.[32] However, there has been no use of CNNs for interpreting 2D fluorescence spectra, and previous implementations have focused on 1D infrared or Raman spectra. Furthermore, the use of CNNs for fluorescence analysis of water

quality has not been explored. It is hypothesized that the strengths of CNNs for processing and interpreting spatially dependent data will be well suited for 2D fluorescence spectra where local groups of values are highly correlated.

This paper investigates the use of deep NNs and CNNs to interpret fluorescence spectra for the prediction of DBP formation potential. The two major groups of regulated DBPs are assessed, THMs and HAAs including the individual species that made up these groups in the samples analysed (trichloromethane, bromodichloromethane, trichloroacetic acid, and dichloroacetic acid). A method to interpret the CNN results is also used to identify fluorescence regions that are most likely associated with high DBP formation potentials.

2 Results

A dataset of DBP formation potentials and associated fluorescence EEM measurements were used to assess the capabilities of deep NNs and CNNs for water quality analysis. Water samples analyzed were from a pilot-scale treatment plant receiving river water. Samples were taken throughout a treatment train consisting of several unit processes including coagulation, flocculation, sedimentation, ozonation, advanced oxidation (peroxide and ozone), and filtration through anthracite or activated carbon. As such, the samples analyzed had a wide range of NOM concentrations and characteristics. Dissolved organic carbon varied from 2.6 to 6.3 mg L^{-1}, and specific ultraviolet absorbance varied from 0.75 to 2.53 L mg^{-1} m^{-1} over all samples. DBP formation potentials were determined by maintaining a free chlorine residual of 1.5 mg/L for 24 hours. Although all four chlorinated or brominated THM and nine HAA species could be detected, only trichloromethane (TCM), bromodichloromethane (BDCM), trichloroacetic acid (TCAA), and dichloroacetic acid (DCAA) were consistently identified at concentrations above detection limits.

2.1 Dense Networks

An iterative optimization approach was used to understand the impact of NN structure on overall performance. While many aspects of network structure can be optimized, the focus was on the number of layers (i.e. depth). A dense network was trained with an increasing number of layers to identify the degree to which network depth can improve prediction accuracy on a test set. Figure 1 shows the total THM and HAA predictions results given the number of layers in a dense NN. The error bars in Figure 1 represent the standard deviation of 8 repeated random initializations of the network. A network with 0 hidden layers is simply the input values (dimensions = 6,336) connected to 1 output node. When the number of layers was increased, each layer's nodes were set to half of the previous layer. For example, with two hidden layers, hidden layer 1 would have 3,168 nodes, and layer 2 would have 1,584 nodes.

As observed in Figure 1, increasing the number of hidden layers beyond one in a dense NN did not improve network performance in predicting both total THMs and HAAs. Total THM mean absolute error was at a minimum with one hidden layer (MAE: 12.26 ± 0.34 µg/L) and total HAA error was also at a minimum with one hidden layer (MAE: 10.04 ± 0.33 µg/L). Similar results were observed for individual species (Table 2). While adding additional layers eventually reduced test set error, this came at the cost of increased variability between random network initializations. For example, the coefficient of variation (CV) increased from 3.3% for a one-layer dense network to predict total HAAs to 8.4% for 6 layers. Increased variability in performance could be due to the increased number of learnable parameters and the relatively small sample size used in this study. The structure chosen, where each layer contained half the nodes of the previous layer, resulted in 15,870,977 trainable parameters with one hidden layer. However, it should be noted that a dense network with one hidden layer of 10 nodes (56,361 trainable parameters) resulted in a decrease in prediction accuracy (1 layer, 10 nodes: total THM MAE 13.54 ± 0.47 µg/L; total HAA MAE 12.62 ± 3.00 µg/L). The lack of improvement of dense NNs beyond one hidden layer

is expected given the small data size and demonstrates that deep dense networks are unlikely to provide advantages in modelling small water quality datasets.

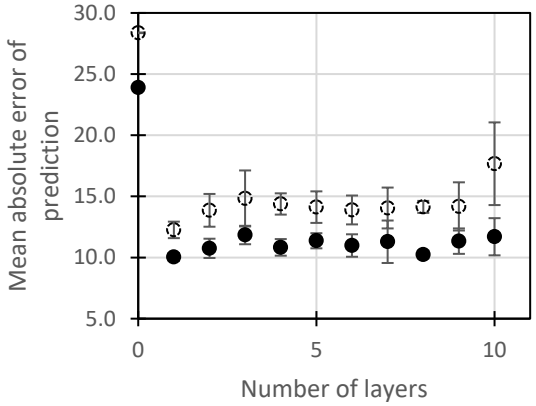

Figure 1 Mean absolute error (MAE) of prediction for a test set (n = 28) of a) THM concentrations and b) HAA concentrations using a dense neural network structure. Error bars represent the standard deviation of MAE over 8 random initializations of the network weights.

2.2 Convolutional Networks

In contrast to dense NNs, prediction accuracy was minimized with increasing CNN network depth (Figure 2). Network depth was investigated by increasing the number of convolutional layers and the number of layer sets (convolution followed by max pooling). Convolutional layers provide learned filters or kernels that identify small features in the spectrum, while max pooling layers decrease dimensionality and pool redundant features.[27,28] Including 5 hidden convolutional layers without pooling layers was found to optimize prediction accuracy (Figure 2a). Increasing the number of layer sets also significantly improved performance (Figure 2b). A further decrease in error of 19.3% for total THMs and 28.9% for total HAAs was observed when increasing both the number of convolutional layers between max pooling layers and increasing the number of max pooling layers (Figure 2c; Table 2).

It was also of interest to investigate the role of the size of receptive fields for each filter and the number of filters included in each convolutional layer. The receptive field size identifies the number of adjacent data points to be considered by each filter. Previous work in chemometrics has shown relatively large receptive fields to work well[31] and could identify features that span over large areas of the spectra, however, expanding the filter size increases the number of trainable parameters. Alternatively, by including convolutional layers in sequence, the receptive field's effective size is expanded, minimizing the number of trainable parameters and including additional layers of non-linearity.[33] As such, the results suggest that larger receptive fields could improve CNN performance. However, increasing the receptive field size beyond (3, 3) for individual layers did not improve performance (Figure 2e), and expanding receptive fields may be best accomplished by stacking convolutional layers in sequence. It is also of note that increasing the number of trained filters improved performance up to 36, after which no further changes were observed (Figure 2d).

An example of the learned filters are shown in Figure 3. CNNs create hierarchical representations of data showing how specific irrelevant spectral features are discarded and specific areas of the spectra needed to predict DBP concentrations are magnified.[29] The initial filter layer identifies large and smooth and broad features in the spectra. After pooling, feature maps become more coarse and more distinct patterns between filters can be discerned, highlighting specific areas of the spectra. In the last layer of feature maps, many filters highlight one constant emission level over several excitation bands (left to right).

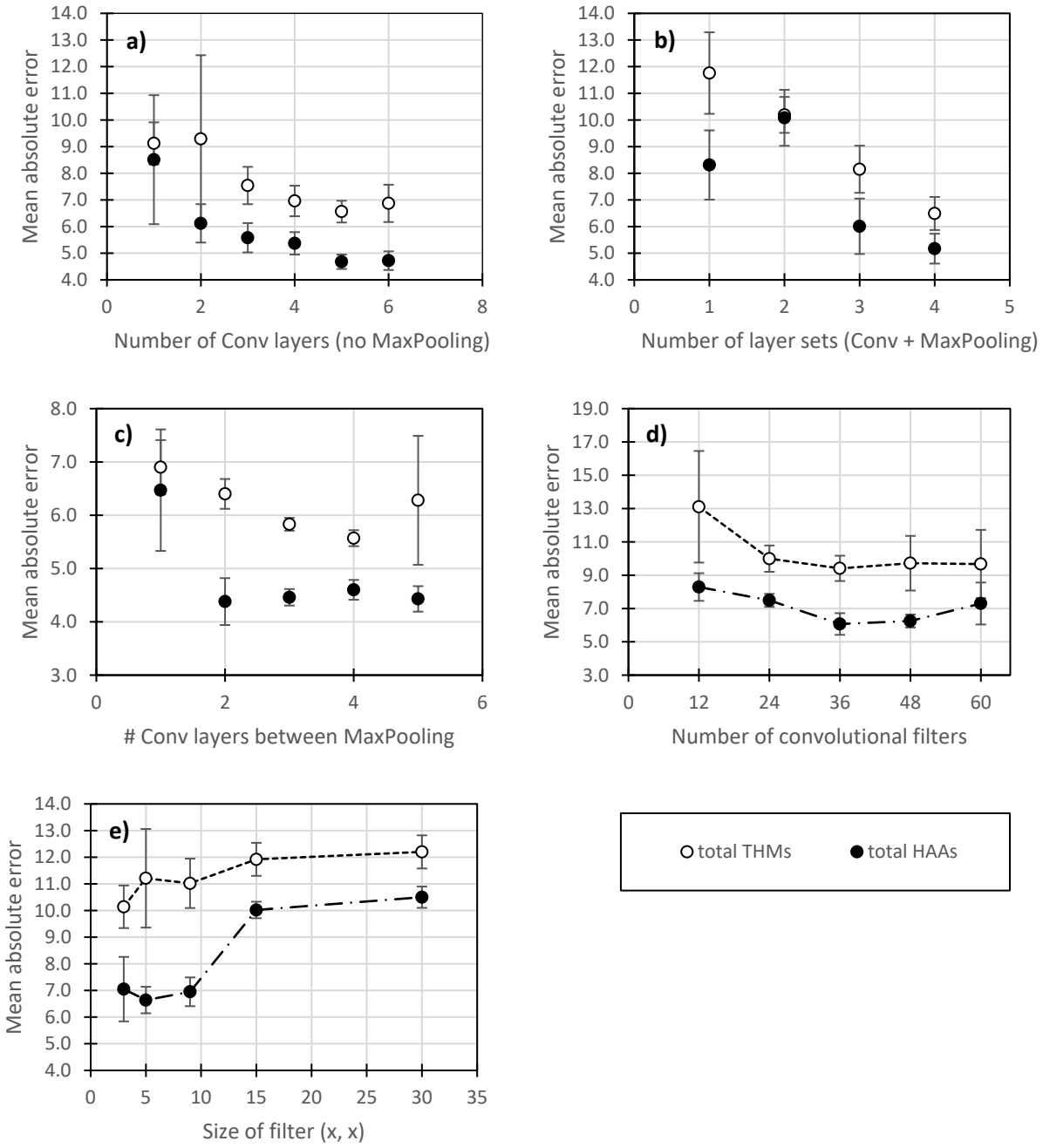

Figure 2 Impact of CNN structure and depth on MAE of test set predictions. a) varies the number of convolutional layers without any max pooling layers, b) varies the number of layer sets with 1 convolutional layer followed by max pooling, c) varies the number of convolutional layers between max pooling layers (4 max pooling layers in total), d) varies the number of convolutional filters for 1 convolutional layer without max pooling, e) varies the size of the receptive field for 1 convolutional layer without max pooling.

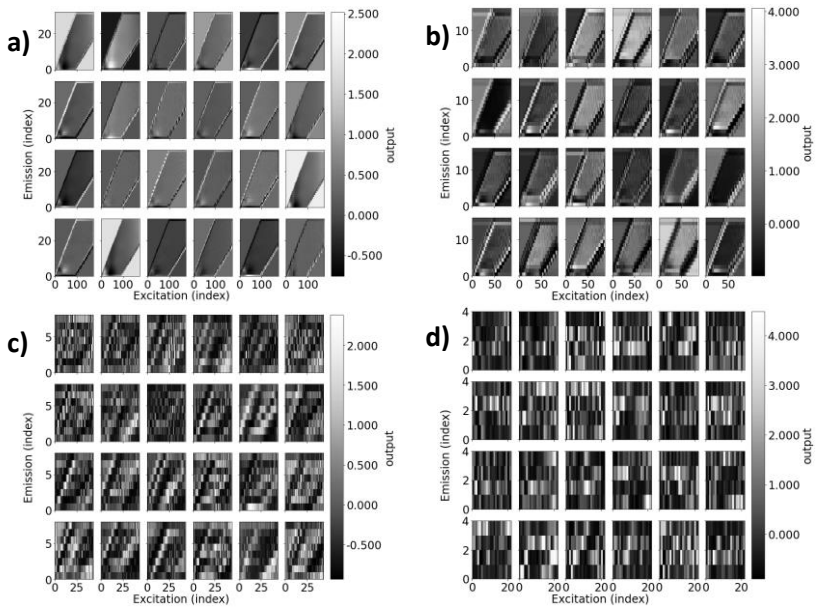

Figure 3 Feature maps of convolutional filters (24) from 4 convolutional layers chosen between max pooling layers. a) first convolutional layer, b) after the first max pooling, c) after the second max pooling, d) after the third max pooling. All max pooling was carried out over a (2,2) window.

The comparison of dense and CNN structures shows a marked performance improvement using a convolutional approach (Figure 2 and Table 2). Compared to the optimized dense systems, a network with one single convolutional layer improved prediction accuracy by 54.6% for total THMs (MAE reduced by 6.69 µg/L) and 55.9% for total HAAs (MAE reduced by 5.61 µg/L) (Table 2). Improvements were also significant for individual species (39.1% - 61.5% reduction in MAE). In general, the most significant decrease in error was observed using CNNs, followed by adding several pooling layers. While adding multiple convolutional layers between pooling layers did further reduce error, the gains were more minor (MAE difference 0.25 – 1.48 µg/L) but statistically significant (paired t-test $p < 0.02$). However, adding stacked convolutional layers resulted in a slight increase in error for total HAA prediction (-0.23 µg/L). As such, the pooling of features and reduction of dimensions is likely key to best interpreting EEMs. It is hypothesized that the success of using pooling layers is due to building in dimensionality reduction and reducing the number of nodes in the final fully-connected or dense layer.

Table 1 Mean absolute error (MAE) of predictions on tests set for several model types. Range or error is the standard deviation from 8 repeated initializations of the same model (where applicable). Error (±) is calculated as the standard deviation over 8 random initialization of the network weights.

	Model			
Disinfection By-Product Species	Dense (1 layer)	CNN (1 layer)	CNN (4 pooling layers, 1 convolutional layer)	CNN (4 pooling layers, 4-5 convolutional layers)
Total THMs	12.26 ± 0.34	6.62 ± 0.13	6.06 ± 0.22	5.57 ± 0.15
Trichloromethane (TCM)	8.80 ± 0.32	7.00 ± 0.84	4.87 ± 0.62	3.39 ± 0.10
Bromodichloromethane (BDCM)	6.34 ± 0.51	4.69 ± 0.33	4.11 ± 0.22	3.86 ± 0.06
Total HAAs	10.04 ± 0.33	4.54 ± 0.24	4.20 ± 0.47	4.43 ± 0.13
Dichloroacetic acid (DCAA)	7.82 ± 0.43	6.07 ± 1.01	4.84 ± 0.26	4.19 ± 0.08
Trichloroacetic acid (TCAA)	8.36 ± 0.33	5.85 ± 0.90	4.58 ± 0.16	4.22 ± 0.10

2.3 Model explanations

The objective of identifying model explanations was to confirm that that model predicts high or low concentrations of DBPs based on fluorescence features that are known or possibly associated with DBP precursors. There are scattering signals (i.e. not from organic material) or other potential artifacts from the sample analysis process that would bias the model to "know" concentrations of DBPs for incorrect reasons. The second objective was to identify fluorescence regions most highly associated with specific DBP formation potentials. This information could be used to further understanding of the characteristics of DBP precursors and potentially optimize treatment processes that preferentially remove compounds with those characteristics.

An occlusion method was applied to identify spectral areas that most significantly influence prediction accuracy. The occlusion method identifies spectral regions most relevant to a prediction by randomly occluding or setting a segment of all inputs in a specified region to 0. The error incurred due to this occlusion indicates how relatively important that specific area is to accurate predictions. The error was calculated as the difference between non-occluded and occluded predictions, and the direction or sign of the error was preserved. As such, positive values indicate that the model underestimated DBP formation with a specific patch occluded, and negative values indicate overestimated DBP formation. A total of 20,000 iterations of random patches per model were chosen to build the heat maps. A random approach to selecting the patch was taken to reduce any bias from neighbouring values since the variables included in each patch would change between iterations. Figure 4 shows the average heat maps identified from training deep CNNs on total DBPs and individual species. Likewise, Figure 5 shows heat maps based on dense NNs.

From occlusion heat maps of variable or spectra area importance (Figure 4), it was observed that fluorescence in the area of approximately ex: 225 – 260 nm and em: 370 – 500 nm was most impactful of prediction accuracy for all DBPs (both total and individual species). A second common area of importance at ex > 300 nm and em > 400 nm was also observed. Fluorescence in these two regions is generally considered to be humic-like and fulvic-like material.[34] Several heat maps also show areas of importance in protein-like fluorescence regions (excitation: 230 – 250 nm, emission: 300 – 360 nm) associated with tryptophan-like or tyrosine-like fluorescence.[34] Spectral importance in these regions conforms well to expectations of DBP precursor type material that can fluoresce, generally thought to be aromatic humic-like or fulvic-like material.[34] The heatmaps provide evidence that the NNs are utilizing signals from regions that are reasonable for DBP prediction. Previous DBP prediction methods based on fluorescence data have utilized the same spectral regions.[15,16,35]

Compared to dense NNs, CNN heat maps show broader areas of importance with more gradual changes (Figure 4 and Figure 5). Gradual changes conform with the expectation of fluorescence signals from fluorophores, and sharp changes are not typically associated with fluorescence from natural organic matter.[14] Furthermore, CNN heat maps emphasize higher excitation bands. Particularly for prediction of trihalomethanes, peaks at excitation > 300 nm were important for positive predictions, while dense NN heat maps placed less importance on these areas. Several CNN heat maps show signals that have several excitation peaks, with limited changes in emission. For example, the BDCM heat map shows peaks at approximately 250 nm and 330 nm, with emissions constant at 450 nm. Multiple excitation peaks at one singular emission conform well with expectations of fluorescence from individual fluorophores, where multiple wavelengths can cause excitation, however, the emission is always from the lowest singlet state and therefore only at one wavelength.[36] The identification of areas of importance at several emission bands suggests several distinct fluorophores contributing to DBP formation potential rather than individual components.

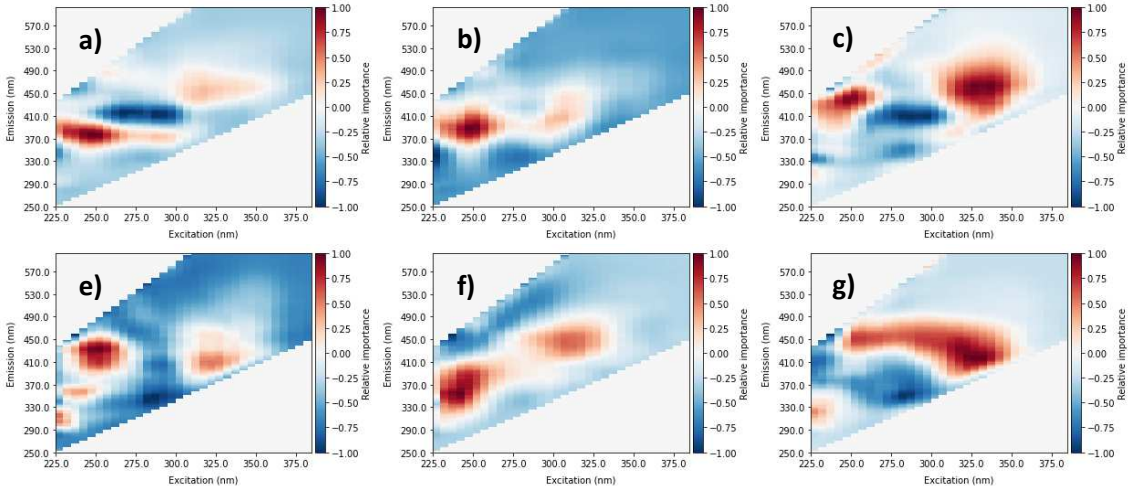

Figure 4 Heat maps from random occlusion of variable importance for CNN prediction of a) total THMs, b) trichloromethane, c) bromodichloromethane, d) total HAAs, e) trichloroacetic acid, f) dichloroacetic acid.

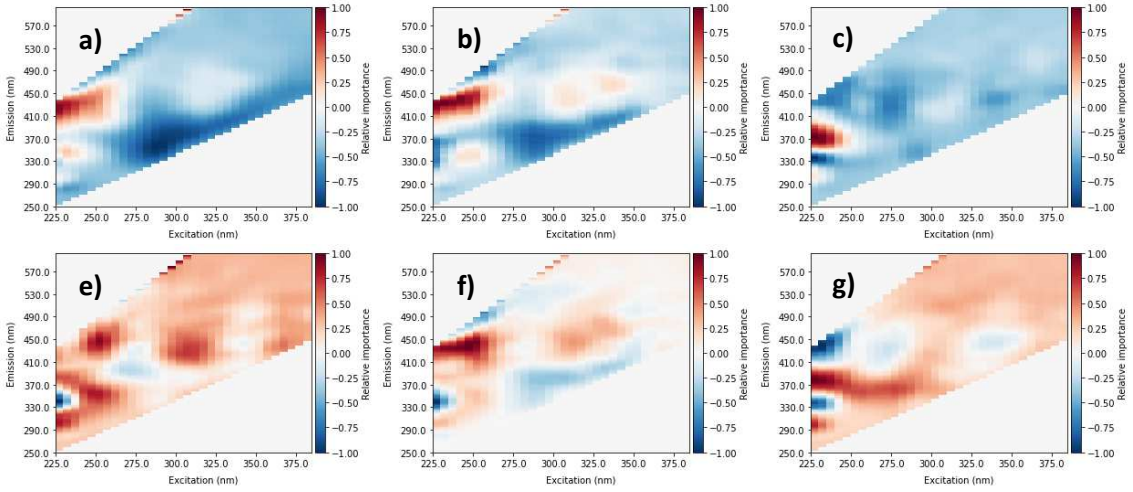

Figure 5 Heat maps from random occlusion of variable importance for dense network prediction of a) total THMs, b) trichloromethane, c) bromodichloromethane, d) total HAAs, e) trichloroacetic acid, f) dichloroacetic acid.

It can also be observed that there is greater continuity of areas of importance with individual DBP species and the total levels. Since the individual species should sum to total concentrations, total THM and total HAA heat maps would be expected to show similar characteristics to the individual species. Ex/em 350 nm/380 nm is observed in CNN heat maps for total THMs (Figure 3a) and TCM (Figure 3b). The secondary peak for total THMs at approximately ex/em 325 nm/450 nm is mirrored in the BDCM heat map (Figure 3c). Similar conformance was not observed with dense NN heat maps, for example, the areas of highest importance for BDCM (Figure 4c) was not present in the total THM heat map (Figure 4a). However, while some overlap is present between CNN heat maps of species and the total DBP levels, it should be noted that not all peaks are mirrored (e.g. BDCM peak at ex/em 250 nm/450 nm not seen in total THM map).

Individual species of DBPs showed differences between CNN heat maps. BDCM areas of importance were shifted to higher excitation and emission areas compared to TCM. A similar pattern can be seen between DCAA and TCAA. Identified differences in spectral areas between individual species were expected since preferential yields of specific by-products from pure model compounds have pointed to certain molecular structures resulting in the preferential formation of individual DBP species.[37,38] A shift to the greater importance of fluorophores at emissions > 450 nm could indicate BDCM and DCAA formation resulting from humic-like material with greater oxygen/carbon ratios and lower hydrogen/carbon ratios, implying an oxidation state ≥ 0.[39]

A second notable difference is the increased importance of protein-like material (ex/em 230 – 250 nm/300 – 350 nm) for HAA predictions. This peak location is typically associated with aromatic amino acids such as tryptophan and tyrosine.[34] Previous studies show that aromatic amino acids[40] and protein-like fluorescence signals strongly correlate with HAA formation potentials.[19,41,42] In particular, the protein-like peak was observed to be most prominent for the prediction of TCAA. This observation conforms well to previous results that show higher TCAA formation than DCAA from aromatic amino acids that would contribute to the observed fluorescence signal.[40] From the dense NN heat map of TCAA, regions surrounding the expected aromatic amino acid peak are positive. However, there is a strong negative relationship in the specific location of tryptophan fluorescence (ex/em 230 nm/340 nm).

3 Discussion

This study investigated the use of deep CNNs to interpret fluorescence spectra and predict the formation of regulated chlorination DBPs from a drinking water treatment plant. The observed results indicate that deep CNNs are well suited to the task of interpreting fluorescence excitation-emission matrices and prediction of DBPs for several reasons: 1) overall prediction accuracy for all DBP groups and species were significantly reduced compared to dense NNs and previous modelling approaches using dimensionality reduction, 2) results from random initializations were less variable using deep CNNs compared to dense and shallow CNNs, 3) deep CNN heat maps show trained networks utilize data from spectral regions that are well known to be associated with DBP formation potentials, and 4) compared to dense NNs, CNNs show heat maps with characteristics more conformant with expectations of fluorescence from organic precursor material.

Compared to previous work that utilized dimensionality reduction prior to regression, the use of CNNs significantly improved the accuracy of prediction. Using the same dataset and training/test data, optimal results of total THM MAE 7.46 µg/L and total HAA MAE 10.75 µg/L were previously reported.[17] Significant reduction of prediction accuracy was achieved using CNN architectures, particularly for HAA prediction (this study: total THM MAE 5.53 µg/L and total HAA MAE 4.37 µg/L). It is not straightforward to compare results to other studies given the variation in number of samples, methods for formation potential determination, range of concentrations in the training/test sets, and performance metrics. However, results found in this study also represent improvements over HAA and THM formation predictions previously reported using similar performance metrics (e.g. total THM MAE 13.5 µg/L and total HAA MAE 7.7 µg/L).[15] Furthermore, previous approaches to utilize fluorescence data for DBP predictions have relied heavily on dimensionality reduction to identify relevant fluorescence features, which adds complexity and may neglect to capture features of importance. In contrast, deep CNNs present an opportunity to utilize full fluorescence spectra without the need for manual or highly supervised feature selection through peak-picking, regional integrations, or PARAFAC analysis. It is thought that deep CNNs provide an opportunity for complex behaviours to be represented by several simpler representations. Observation of feature maps produced by convolutional layers show a hierarchy of feature representations, with general and smooth representations in high layers and progressively coarser and more specific highlighted spectral areas as increasing numbers of convolution and pooling layers are applied.

Neural networks are often discussed as black-box type algorithms, where the internal reasoning is unknown or difficult to illustrate. However, it is imperative that the logic of prediction algorithms used in applied tasks, such as prediction of potentially toxic disinfection by-products, is discernible. There is also an opportunity to use these powerful data-driven approaches to help identify important variables or characteristics of the system. The use of heat maps generated from an occlusion approach to identifying spectral areas that highly influence predictions gives insight into the decision-making process and helps confirm that trained networks are relying on data from spectral regions associated with DBP precursors. Furthermore, heat maps can help direct future more detailed studies investigating the characteristics of precursor material.

As such, the use of fluorescence spectroscopy coupled with machine learning techniques, such as deep CNNs, show promise to be used for rapid estimation of DBP formation potentials. In the context of typical regulatory thresholds for water treatment (total THMs < 80 µg/L, total HAAs < 60 µg/L) the presented methodology produced error levels (MAE 3.39 – 5.53 µg/L) that would be appropriate for rapidly informing operations and management regarding conformance with regulatory thresholds. Knowledge of DBP formation potentials in near real-time can enable tighter treatment controls and management efforts to minimize the public's exposure to DBPs.

4 Methods

4.1 Water samples

Water samples were obtained from parallel pilot treatment trains that were fed Otonabee River water (Peterborough, Ontario, Canada). Samples were obtained throughout the treatment train for fluorescence analysis and for determining DBP formation potentials. Processes applied included coagulation, flocculation, sedimentation, ozonation, advanced oxidation (peroxide and ozone), and filtration through anthracite or activated carbon. Further information on the pilot-scale set-up and water samples can be found in Peleato et al. (2017).[43]

4.2 Fluorescence

Fluorescence spectra were collected using an Agilent Cary Eclipse fluorescence spectrophotometer (Mississauga, Canada). Excitation and emission wavelength ranges were 225 – 380 nm (5 nm increments), and 250 – 600 nm (2 nm increments), respectively. The fluorescence spectra were blank subtracted using Milli-Q® water. The spectrum for Milli-Q® water was also used to apply Raman corrections at an excitation wavelength of 350 nm and bandwidth of 5 nm to allow fluorescence intensities to be reported in Raman Units (RU).[44] Absorbance values collected over the excitation-emission range were used to correct for inner-filter effects. Rayleigh scattering lines were removed by setting all values above 2nd order Rayleigh or below 1st order Rayleigh to 0. The absorbance corrected spectra were then scaled between 0 and 1 for each excitation/emission pair.

4.3 Neural Networks

All NNs were trained in Python 3.6 using the Keras library (v2.3.1; TensorFlow v1.15.0 backend). Hardware used was a Intel® Xeon® E2286G CPU and a NVIDIA GeForce RTX 2080. Training of all models took less than 20 minutes.

Two general types of NNs were investigated: dense networks where there is a weighted connection between every node in subsequent layers, and CNNs. The number of nodes in each hidden layer of a dense network was defined as half of the previous layer. For example, with two hidden layers, hidden layer 1 would have 3,168 nodes and layer 2 would have 1,584.

A general schematic of the CNN structure is shown in Figure 6. Convolutional layers involved training a set of 2D filters or kernels, which are weighting functions multiplied with input values in a specific spatial window. Filters are smaller than the input dimensions and are slid across the entire input to produce feature mapping of the input. Since one filter or weighting function is used for the whole input space, fewer trainable parameters are needed than dense networks. It also gives rise to feature invariance since the trained filter can identify a feature in any position of the input space. Max pooling layers look for the maximum value within a spatial window and then uses that maximum value to represent the output over that spatial window, effectively reducing dimensionality. For more details on the mathematics of CNNs and the training process, see LeCun et al. (1998)[45] and Goodfellow et al., (2016).[28] CNN layers were considered as a set of convolutional layers followed by max pooling. Following convolutional layers, the structure was flattened, where pooled feature maps are vectorized into a dense hidden layer followed by a single output node. A varied number of structures were investigated to identify changes in performance based on depth (number of layers), size of convolutional filters (spatial window), number of max pooling layers, and number of convolutional filters. A summary of these structures is presented in Table 2.

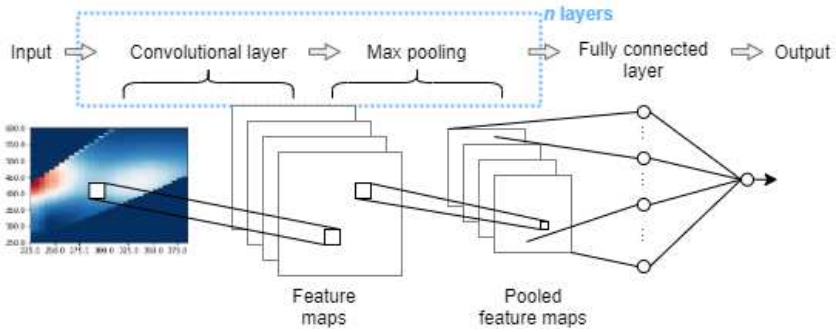

Figure 6 General schematic of convolutional neural network structure. The number of convolutional layers as well as the number of layers (convolution + max pooling) can be varied.

Table 2 Descriptions of the general model types used.

Model description	**Num. hidden layers**	**Hidden layer**	**Output layers**
Dense	n	Dense ($nodes = 0.5 \cdot previous\ layer$) -> Batch Normalization -> Activation (elu)	Output
CNN 1 layer, no max pooling	1	Convolution 2D ($number\ of\ filters, size\ of\ filter\ (x,x)$) -> Batch Normalization -> Activation (elu)	Flatten -> Output
CNN *n* layers	n	Convolution 2D ($number\ of\ filters, size\ of\ filter\ (x,x)$) -> Batch Normalization -> Activation (elu) -> Max Pooling 2D (2,2)	Flatten -> Output
CNN *n* convolutions, 4 layers	4	$\boldsymbol{n} \times \{$ Convolution 2D ($number\ of\ filters, size\ of\ filter\ (x,x)$) -> Batch Normalization -> Activation (elu) **} ->** Max Pooling 2D (2,2)	Flatten -> Output

Common between all structures and types of NNs was the use of batch normalization to speed up training,[46] followed by activation using an exponential linear unit (elu) activation function (equation 1).

$$ELU: f(x) = \begin{cases} x, & x > 0 \\ e^x - 1, & x \leq 0 \end{cases} \quad \text{(equation 1)}$$

All networks were trained with a mean squared error loss function coupled with $L2$ regularization to prevent overfitting (equation 2). The Adam optimization algorithm was used for all training.

$$Loss = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y_i})^2 + \lambda\|w^2\| \quad \text{(equation 2)}$$

Where, y_i is the network output for sample i (prediction)
$\hat{y_i}$ is the true value for sample i
N is the total number of samples
λ is a hyperparameter to control $L2$ regularization (set to 0.01)
w are all the network weights

Prediction accuracy was determined on a test set (20% of all data, n=28) that was not used for training the network. The metric used to assess predictive performance was mean absolute error (MAE), primarily since it provides a metric in the same units used in analysis and is more easily interpreted.

4.4 Occlusion method

An occlusion approach was used to generate heat maps of spectral areas that most influence prediction accuracy. After training a network, test data was modified by iteratively setting a spectral area or patch equal to 0. These occluded or corrupted training samples are then fed through the network to produce a prediction of DBP concentration (Table 3).

Table 3 Description of the occlusion method used to identify spectral heat maps or areas of importance.

Occlusion method	
1:	Train a network using original training data (X_{train})
2:	Predict outputs (y_{test}) using original test data (X_{test})
3:	**For** *t = 1* to *20,000* **do**
4:	Randomly select EEM patch from X_{test}
5:	Set patch = 0 to create corrupted test set, $X_{occluded}$
6:	Predict outputs ($y_{occluded}$) using $X_{occluded}$
7:	Average error calculated over all test data. $Error = \frac{1}{N_{test}}\sum_{i=1}^{N_{test}}(y_{test} - y_{occluded})$, where N_{test} is the number of samples in the test set.
8:	Check all excitation/emission were included in random selection
9:	Calculate average error over all iterations for each excitation/emission pair

The difference between initial predictions and occluded predictions provided an estimate of the importance of the occluded patch. If initial predictions and occluded predictions are identical or close, the trained network is not relying on that spectral area to estimate DBPs. On the other hand, if the error is high, the occluded region is influential on the accurate prediction of DBP levels.

References

1. Storey, M. V., van der Gaag, B. & Burns, B. P. Advances in on-line drinking water quality monitoring and early warning systems. *Water Research* **45**, 741–747 (2011).
2. Faassen, S. & Hitzmann, B. Fluorescence Spectroscopy and Chemometric Modeling for Bioprocess Monitoring. *Sensors* **15**, 10271–10291 (2015).
3. Beutel, S. & Henkel, S. In situ sensor techniques in modern bioprocess monitoring. *Applied Microbiology and Biotechnology* vol. 91 1493–1505 (2011).
4. Bridgeman, J., Bieroza, M. & Baker, A. The application of fluorescence spectroscopy to organic matter characterisation in drinking water treatment. *Reviews in Environmental Science and Biotechnology* vol. 10 277–290 (2011).
5. Murphy, K. R., Stedmon, C. A. & Bro, R. Chemometric analysis of organic matter fluorescence. in *Aquatic organic matter fluorescence* 339–375 (2014). doi:10.13140/2.1.2595.8080.
6. Wagner, E. D. & Plewa, M. J. CHO cell cytotoxicity and genotoxicity analyses of disinfection by-products: An updated review. *Journal of Environmental Sciences* **58**, 64–76 (2017).
7. Guilherme, S., Dorea, C. C. & Rodriguez, M. J. Decision-making scheme for disinfection by-product monitoring intended for small drinking water systems. *Environ. Sci.: Water Res. Technol.* **3**, 366–376 (2017).
8. Chen, B. & Westerhoff, P. Predicting disinfection by-product formation potential in water. *Water Research* **44**, 3755–3762 (2010).
9. Kulkarni, P. & Chellam, S. Disinfection by-product formation following chlorination of drinking water: Artificial neural network models and changes in speciation with treatment. *Science of The Total Environment* **408**, 4202–4210 (2010).
10. Lin, H. *et al.* Radial basis function artificial neural network able to accurately predict disinfection by-product levels in tap water: Taking haloacetic acids as a case study. *Chemosphere* **248**, 125999 (2020).
11. Singh, K. P. & Gupta, S. Artificial intelligence based modeling for predicting the disinfection by-products in water. *Chemometrics and Intelligent Laboratory Systems* **114**, 122–131 (2012).
12. Sadiq, R. & Rodriguez, M. J. Disinfection by-products (DBPs) in drinking water and predictive models for their occurrence: a review. *Science of The Total Environment* **321**, 21–46 (2004).
13. Matilainen, A. *et al.* An overview of the methods used in the characterisation of natural organic matter (NOM) in relation to drinking water treatment. *Chemosphere* **83**, 1431–1442 (2011).
14. Murphy, K. R., Bro, R. & Stedmon, C. A. Chemometric Analysis of Organic Matter Fluorescence. in *Aquatic Organic Matter Fluorescence* (eds. Coble, P., Lead, J., Baker, A., Reynolds, D. M. & Spencer, R. G. M.) 339–375 (Cambridge University Press, 2014). doi:10.1017/CBO9781139045452.016.
15. Trueman, B. F., MacIsaac, S. A., Stoddart, A. K. & Gagnon, G. A. Prediction of disinfection by-product formation in drinking water via fluorescence spectroscopy. *Environ. Sci.: Water Res. Technol.* **2**, 383–389 (2016).
16. Pifer, A. D. & Fairey, J. L. Improving on SUVA254 using fluorescence-PARAFAC analysis and asymmetric flow-field flow fractionation for assessing disinfection byproduct formation and control. *Water Research* **46**, 2927–2936 (2012).
17. Peleato, N. M., Legge, R. L. & Andrews, R. C. Neural networks for dimensionality reduction of fluorescence spectra and prediction of drinking water disinfection by-products. *Water Research* **136**, 84–94 (2018).
18. Xu, X. *et al.* EEM–PARAFAC characterization of dissolved organic matter and its relationship with disinfection by-products formation potential in drinking water sources of northeastern China. *Science of The Total Environment* **774**, 145297 (2021).

19. Ma, C., Xu, H., Zhang, L., Pei, H. & Jin, Y. Use of fluorescence excitation–emission matrices coupled with parallel factor analysis to monitor C- and N-DBPs formation in drinking water recovered from cyanobacteria-laden sludge dewatering. *Science of The Total Environment* **640–641**, 609–618 (2018).
20. Yang, X., Shang, C., Lee, W., Westerhoff, P. & Fan, C. Correlations between organic matter properties and DBP formation during chloramination. *Water Research* **42**, 2329–2339 (2008).
21. Peiris, R. H. *et al.* Identifying fouling events in a membrane-based drinking water treatment process using principal component analysis of fluorescence excitation-emission matrices. *Water Research* **44**, 185–194 (2010).
22. Shutova, Y., Baker, A., Bridgeman, J. & Henderson, R. K. Spectroscopic characterisation of dissolved organic matter changes in drinking water treatment: From PARAFAC analysis to online monitoring wavelengths. *Water Research* **54**, 159–169 (2014).
23. Murphy, K. R., Stedmon, C. A., Graeber, D. & Bro, R. Fluorescence spectroscopy and multi-way techniques. PARAFAC. *Analytical Methods* **5**, 6557–6566 (2013).
24. Sharpless, C. M. & Blough, N. V. The importance of charge-transfer interactions in determining chromophoric dissolved organic matter (CDOM) optical and photochemical properties. *Environ. Sci.: Processes Impacts* **16**, 654–671 (2014).
25. Kothawala, D. N., Murphy, K. R., Stedmon, C. A., Weyhenmeyer, G. A. & Tranvik, L. J. Inner filter correction of dissolved organic matter fluorescence. *Limnology and Oceanography: Methods* **11**, 616–630 (2013).
26. Bro, R. PARAFAC. Tutorial and applications. in *Chemometrics and Intelligent Laboratory Systems* vol. 38 149–171 (1997).
27. LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. *Nature* **521**, 436–444 (2015).
28. Goodfellow, I., Bengio, Y. & Courville, A. *Deep Learning*. (MIT Press, 2016).
29. Zhang, X. *et al.* Understanding the learning mechanism of convolutional neural networks in spectral analysis. *Analytica Chimica Acta* **1119**, 41–51 (2020).
30. Bjerrum, E. J., Glahder, M. & Skov, T. Data Augmentation of Spectral Data for Convolutional Neural Network (CNN) Based Deep Chemometrics. *arXiv:1710.01927 [cs]* (2017).
31. Malek, S., Melgani, F. & Bazi, Y. One-dimensional convolutional neural networks for spectroscopic signal regression. *Journal of Chemometrics* **32**, e2977 (2018).
32. Chatzidakis, M. & Botton, G. A. Towards calibration-invariant spectroscopy using deep learning. *Sci Rep* **9**, 2126 (2019).
33. Simonyan, K. & Zisserman, A. Very Deep Convolutional Networks for Large-Scale Image Recognition. *arXiv:1409.1556 [cs]* (2015).
34. Chen, W., Westerhoff, P., Leenheer, J. A. & Booksh, K. Fluorescence Excitation−Emission Matrix Regional Integration to Quantify Spectra for Dissolved Organic Matter. *Environ. Sci. Technol.* **37**, 5701–5710 (2003).
35. Roccaro, P., Vagliasindi, F. G. A. & Korshin, G. V. Changes in NOM Fluorescence Caused by Chlorination and their Associations with Disinfection by-Products Formation. *Environ. Sci. Technol.* **43**, 724–729 (2009).
36. Lakowicz, J. R. *Principles of Fluorescence Spectroscopy*. (Springer Science & Business Media, 2013).
37. Dickenson, E. R. V., Summers, R. S., Croué, J.-P. & Gallard, H. Haloacetic acid and Trihalomethane Formation from the Chlorination and Bromination of Aliphatic β-Dicarbonyl Acid Model Compounds. *Environ. Sci. Technol.* **42**, 3226–3233 (2008).
38. Zeng, T. & Arnold, W. A. Clustering Chlorine Reactivity of Haloacetic Acid Precursors in Inland Lakes. *Environ. Sci. Technol.* **48**, 139–148 (2014).
39. Lavonen, E. E. *et al.* Tracking changes in the optical properties and molecular composition of dissolved organic matter during drinking water production. *Water Research* **85**, 286–294 (2015).

40. Hong, H. C., Wong, M. H. & Liang, Y. Amino Acids as Precursors of Trihalomethane and Haloacetic Acid Formation During Chlorination. *Arch Environ Contam Toxicol* **56**, 638–645 (2009).
41. Hua, L.-C., Lin, J.-L., Chen, P.-C. & Huang, C. Chemical structures of extra- and intra-cellular algogenic organic matters as precursors to the formation of carbonaceous disinfection byproducts. *Chemical Engineering Journal* **328**, 1022–1030 (2017).
42. Nemani, V. A., Taylor-Edmonds, L., Peleato, N. M. & Andrews, R. C. Impact of operational parameters on biofiltration performance: organic carbon removal and effluent turbidity. *Water Science & Technology: Water Supply* **16**, 1683–1692 (2016).
43. Peleato, N. M., Sidhu, B. S., Legge, R. L. & Andrews, R. C. Investigation of ozone and peroxone impacts on natural organic matter character and biofiltration performance using fluorescence spectroscopy. *Chemosphere* **172**, 225–233 (2017).
44. Lawaetz, A. J. & Stedmon, C. A. Fluorescence Intensity Calibration Using the Raman Scatter Peak of Water. *Applied Spectroscopy* **63**, 936–940 (2009).
45. Lecun, Y., Bottou, L., Bengio, Y. & Haffner, P. Gradient-based learning applied to document recognition. *Proceedings of the IEEE* **86**, 2278–2324 (1998).
46. Ioffe, S. & Szegedy, C. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. in *International Conference on Machine Learning* 448–456 (PMLR, 2015).

Acknowledgements

This work was funded in part by the Canadian Water Network and the Natural Sciences and Engineering Research Council of Canada (NSERC).

Author Contributions

NP designed the work, wrote the scripts, conducted the analysis, and wrote the manuscript