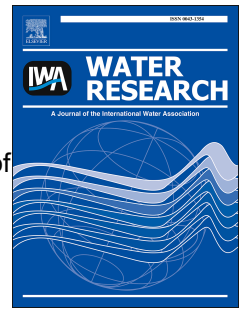
Accepted Manuscript



Neural networks for dimensionality reduction of fluorescence spectra and prediction of drinking water disinfection by-products

Nicolas M. Peleato, Raymond L. Legge, Robert C. Andrews

PII: S0043-1354(18)30159-3

DOI: 10.1016/j.watres.2018.02.052

Reference: WR 13604

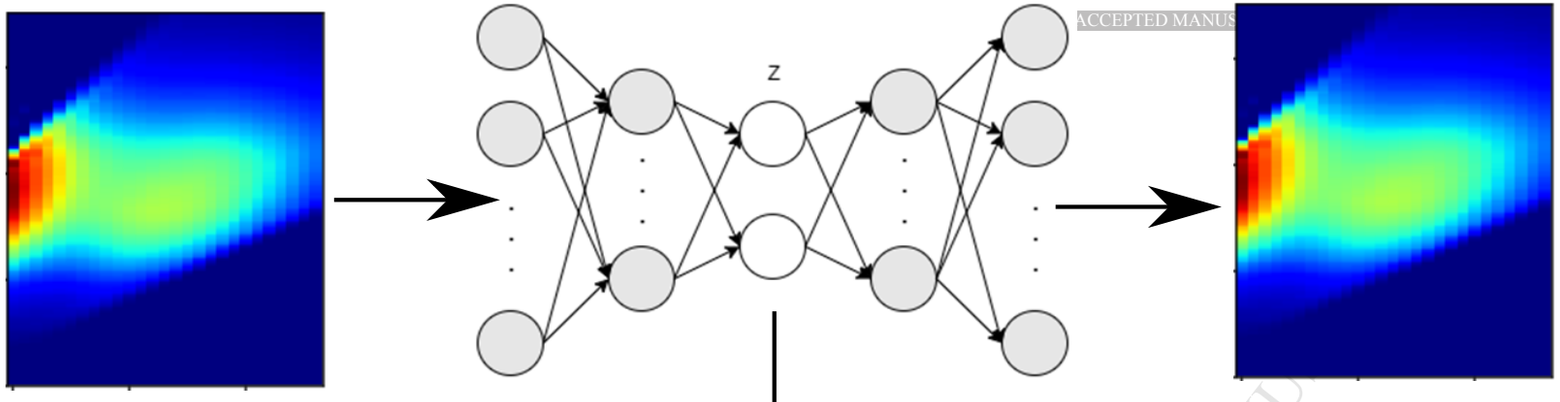
To appear in: *Water Research*

Received Date: 25 October 2017

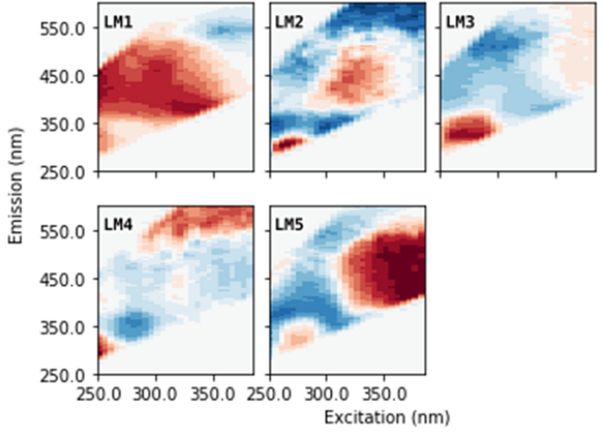
Revised Date: 14 February 2018

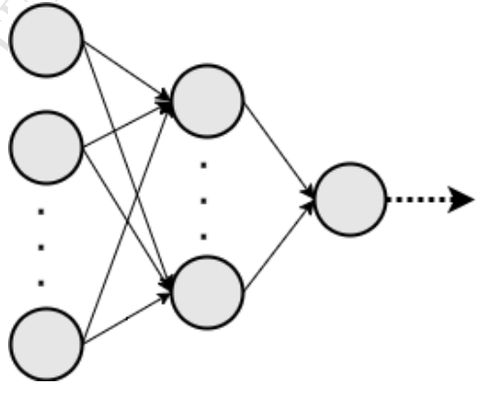
Accepted Date: 20 February 2018

Please cite this article as: 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* (2018), doi: 10.1016/j.watres.2018.02.052.

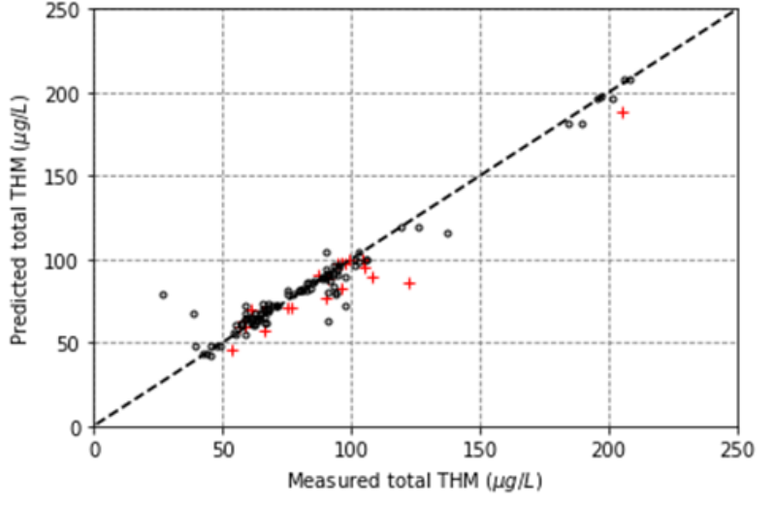


Fluorescence analysis by autoencoder





Improved disinfection by-product formation prediction



Neural Networks for Dimensionality Reduction of Fluorescence Spectra and Prediction of Drinking Water Disinfection By-Products

Nicolas M. Peleato [a*], Raymond L. Legge [b], Robert C. Andrews [a]

[*] Corresponding author: nicolas.peleato@mail.utoronto.ca; +1 416 946 0486

[a] Department of Civil Engineering, University of Toronto, 35 St. George St., Toronto, Ontario, Canada, M5S 1A4

[b] Department of Chemical Engineering, University of Waterloo, 200 University Ave., Waterloo, Ontario, Canada, N2L 3G1

Abstract

The use of fluorescence data coupled with neural networks for improved predictability of drinking water disinfection by-products (DBPs) was investigated. Novel application of autoencoders to process high-dimensional fluorescence data was related to common dimensionality reduction techniques of parallel factors analysis (PARAFAC) and principal component analysis (PCA). The proposed method was assessed based on component interpretability as well as for prediction of organic matter reactivity to formation of DBPs. Optimal prediction accuracies on a validation dataset were observed with an autoencoder-neural network approach or by utilizing the full spectrum without pre-processing. Latent representation by an autoencoder appeared to mitigate overfitting when compared to other methods. Although

DBP prediction error was minimized by other pre-processing techniques, PARAFAC yielded interpretable components which resemble fluorescence expected from individual organic fluorophores. Through analysis of the network weights, fluorescence regions associated with DBP formation can be identified, representing a potential method to distinguish reactivity between fluorophore groupings. However, distinct results due to the applied dimensionality reduction approaches were observed, dictating a need for considering the role of data pre-processing in the interpretability of the results. In comparison to common organic measures currently used for DBP formation prediction, fluorescence was shown to improve prediction accuracies, with improvements to DBP prediction best realized when appropriate pre-processing and regression techniques were applied. The results of this study show promise for the potential application of neural networks to best utilize fluorescence EEM data for prediction of organic matter reactivity.

Keywords: Fluorescence spectroscopy; disinfection by-products; neural networks; autoencoder; dimensionality reduction; water treatment

1 Introduction

Presence of naturally occurring organic matter is of ubiquitous concern for drinking water treatment operations. Organic matter (OM) is known to adversely impact treatment processes such as filtration or adsorption processes and is a major source of disinfectant demand (Fabris et al., 2008). Reactions between OM and oxidants used for disinfection, most commonly chlorine, are known to produce disinfection by-products (DBPs). Regulation of DBPs typically focus on two groupings of organic halides, trihalomethanes (THMs) and haloacetic acids (HAAs) (Hua and Reckhow, 2007). Control and management of OM prior to disinfection is therefore directly tied to DBP formation potential and is essential to protecting treated water quality.

One of the major challenges with OM is the breadth and chemical variability of compounds present in source waters, which is not readily captured by routine organic measures such as dissolved or total organic carbon (DOC/TOC) and absorbance of ultraviolet light at 254 nm (UVA) (Matilainen et al., 2011). These organic estimators are used in models which predict DBP formation potential due to their relative simplicity, allowing for possible continuous or routine monitoring (Chowdhury et al., 2009). In an effort to improve DBP predictability and modelling, fluorescence has been investigated as a sensitive measure of OM

character and reactivity (Hua et al., 2010; Pifer and Fairey, 2012; Roccaro et al., 2009). It is hypothesized that predictability of DBP formation will increase with use of fluorescence data that reflects the chemical composition of organic matter ultimately dictating the reactivity of OM. In contrast to other OM characterization techniques such as liquid chromatography with organic carbon detection (LC-OCD) or high resolution mass spectrometry, fluorescence measurements require little sample preparation or acquisition time, therefore lending to possible online implementation (Shutova et al., 2014).

Fluorescence data, collected as a high dimensional excitation-emission matrix (EEM), present an analysis challenge, making inclusion in traditional modelling approaches difficult, such as linear regression. Reduction of EEM dimensionality is typically practiced, either through manual selection of peaks or regions or multiway dimensionality reduction techniques such as parallel factors analysis (PARAFAC) or principle component analysis (PCA) (Bridgeman et al., 2011). In particular, PARAFAC analysis has been proven effective for identifying underlying components which most resemble the expected excitation/emission characteristics of organic fluorophores (Kathleen R Murphy et al., 2014). In comparison to PCA, it can be argued that PARAFAC is a more appropriate model to account for the three-dimensional

nature of EEMs (Bridgeman et al., 2011). PCA is a two-way method, which requires data to be unfolded prior to analysis therefore discarding information regarding the three-way structure of the data (Bro, 1997). Furthermore, PCA results in components with rotational freedom that makes direct relations to real fluorescence profiles difficult (Stedmon et al., 2003). However, PARAFAC can be shown to be a constrained PCA model and as such, PCA will represent a greater degree of variance within the dataset (Bro, 1997). The components produced from PCA are strictly orthogonal and independent under the assumption of multivariate normality (Murphy, 2012), which may be advantageous for subsequent statistical modeling using PCA results.

A neural network (NN) approach may allow for dimensionality reduction of fluorescence spectra without explicit constraints (Bieroza et al., 2011). For example, when compared to PCA on several test sets, Hinton and Salakhutdinov (Hinton and Salakhutdinov, 2006) demonstrated improved performance of autoencoder NNs; a network where the output is a reconstruction of input after passing through a constrained bottleneck. Few applications of NNs for fluorescence spectroscopy have been reported. Wolf et al. (Wolf et al., 2007) showed improvement to utilizing a NN for prediction of membrane bioreactor performance

by applying PCA prior to training the NN. Bieroza et al. (Bieroza et al., 2011) applied a self-organized map (SOM), a type of NN with a competitive learning approach, and PARAFAC to a set of fluorescence data for raw and treated drinking water samples from 16 surface water plants in the UK. These two methods were used for reducing fluorescence data dimensionality prior to being used as input to a NN, as well as a multilinear model, that predicted TOC removal due to treatment. Rhee et al. (Rhee et al., 2005) employed SOMs for non-linear dimensionality reduction of fluorescence EEMs for monitoring fermentation processes.

Previous studies which have investigated fluorescence as a surrogate for predicting DBP formation employed a range of dimensionality simplification reduction techniques. Roccaro et al. (Roccaro et al., 2009) reported strong correlations between changes in the ratio of fluorescence intensities (at 500 and 450 nm) before and after chlorination to THM and HAA formation. Hua et al. (Hua et al., 2010) utilized PARAFAC to identify two components which were likely THM precursors and better surrogates than SUVA. Similarly, Pifer and Fairey (Pifer and Fairey, 2012) reported that one PARAFAC component was highly correlated with chloroform concentrations and represented a marked improvement compared to SUVA. Bergman et al. (2016) demonstrated success with utilizing fluorescence PARAFAC data

for DBP prediction through a binary classification tree approach to determine adherence with regulations and predicting bromide incorporation factors. Interpretation of the fluorescence EEMs by other means have also been investigated for DBP prediction. Trueman et al. (2016) applied several novel approaches including lasso regression, boosted regression trees, and supervised principle component regression. Through cross-validation the authors demonstrated improved accuracy of alternative approaches compared to linear or unsupervised PCA-based models with both full and bench-scale samples. To facilitate the on-line application of fluorescence sensors, Li et al. (2016) developed a novel UV fluorescence sensor using a single UV LED 280 nm light-emitting diodes. These were used to determine protein and humic-like fluorescence, which improved overall correlations with THM formation across 16 drinking water sources.

This work describes novel use of autoencoder neural networks to interpret high-dimensional fluorescence spectra. This proposed dimensionality reduction method is tested on the basis of predicting disinfection by-product (DBP) formation during drinking water treatment. Dimensionality reduction of fluorescence EEMs is typically practiced for several reasons including identifying underlying interpretable components which resemble organic fluorophores, or simplifying the dataset to

eliminate noise and improve subsequent modelling. Methods such as PCA or PARAFAC achieve these goals to differing degrees and selection of dimensionality reduction techniques should depend on the study objectives. To utilize pre-processed fluorescence EEMs, use of neural networks for improvements to regression and DBP prediction was investigated through comparison to commonly applied linear regression. Efforts have been taken to report an accurate assessment of DBP formation predictability and error rates using validation datasets rather than the more commonly reported overall correlations over entire datasets.

2 Methods

2.1 WATER QUALITY AND DISINFECTION BY-PRODUCTS

Water samples used in this study were obtained from a pilot-scale treatment system which continuously receives Otonabee River water (Peterborough, Ontario, Canada). Several parallel treatment trains were used to collect samples with distinct organic concentrations and character. Treatment steps for each train included conventional treatment (coagulation/flocculation/sedimentation) followed by ozonation or H_2O_2 + O_3 with varying dose levels. The pre-treated water was then passed selectively to six parallel filtration columns, described

181 further in Peleato et al. (2017) with varying media types (anthracite
182 or activated carbon) as well as biological activity levels. In total 2
183 sampling days in each of the months of May, September, and
184 October resulted in analysis of 120 samples. Each day included
185 duplicate samples from raw water, post pre-treatment (3 types:
186 conventional or oxidation), and post filtration (6 distinct filters).
187 This resulted in a dataset with a large degree of variance in organic
188 concentrations and characteristics that were all derived from
189 common source water.

190 Dissolved organic carbon was quantified by the persulfate
191 wet oxidation method described in Standard Method 5310 D
192 (APHA/AWWA/WEF, 2012) with an O-I Corporation Model 1010
193 TOC Analyzer (College Station, Texas, USA). Ultraviolet
194 absorbance was measured at 254 nm with a CE 3055 model
195 spectrophotometer (Cecil Instruments, Cambridge, England)
196 following Standard Method 5910 B (APHA/AWWA/WEF, 2012).
197 Across sample types, DOC ranged from 2.6 to 6.3 mg L^{-1}, UVA
198 and SUVA varied from 0.024 to 0.125 cm^{-1} and 0.75 to 2.53 L mg^{-1}
199 m^{-1}, respectively. Water temperature ranged between 13.7 and
200 25.4 ^{o}C.

201 With respect to DBP formation, samples were collected
202 were dosed with sodium hypochlorite to result in a free chlorine
203 residual of 1.5 ± 0.5 mg L^{-1} after 24 hours based on the Standard

Method 4500-CI G (APHA/AWWA/WEF, 2012). To achieve this residual, chlorine doses were between 5 and 7 mg L^{-1} Cl_2. Following incubation for 24 hours at 20^{o}C, chlorine residual was measured and free chlorine was quenched using excess ascorbic acid. Both THMs and HAAs were quantified using liquid-liquid extraction and gas chromatography. A Hewlett Packard 5890 Series II Plus gas chromatograph equipped with a DB 5.625 capillary column and electron capture detector was used (Agilent, Mississauga, ON). Standard Method 6232 B was followed for quantification of the four THM species; with Standard Method 6251 B for nine HAA species (APHA/AWWA/WEF, 2012).

2.2 FLUORESCENCE

Fluorescence spectra were collected using an Agilent Cary Eclipse fluorescence spectrophotometer (Mississauga, Canada). Optimal instrument settings were determined based on previous studies and in-house testing (Peiris et al., 2009). Excitation and emission wavelength ranges were 250 – 380 nm (5 nm increments), and 250 – 600 nm (2 nm increments), respectively. A fluorescence spectrum of Milli-Q® water was subtracted from each sample to account for the solvent background. This spectrum was also used to apply Raman corrections at an excitation wavelength of 350 nm and bandwidth of 5 nm in order to report fluorescence intensities in Raman Units (RU) (Lawaetz and

Stedmon, 2009). Absorbance spectra between 250 and 600 nm (1 nm increments) for each sample were recorded using an Agilent 8453 UV-Vis spectrophotometer (Mississauga, Canada) to be used to correct for any potential inner filter effects (Kothawala et al., 2013). Corrected and Raman normalized spectra were used for all subsequent dimensionality reduction and analysis.

2.2.1 PARAFAC

Fluorescence EEMs were analyzed using parallel factors analysis (PARAFAC). A methodology as described by Murphy et al. (Murphy et al., 2013) was followed using the drEEM toolbox for MATLAB. Rayleigh and Raman scatter regions were removed for conformity to the linear assumptions required for PARAFAC. Several samples were identified as outliers through observation of sample leverages on the model and were removed. A total of 12 samples were removed to create a stable and valid model. The validity of the PARAFAC model, or determining the correct number of components, was established through several means. Spectral loadings of the components were observed to conform to general guidelines regarding how organic fluorophores signals appear (e.g. only one emission peak, no abrupt changes in loadings). Split-half validation was also carried out based on a randomized split of the dataset, forming 3 unique comparisons of dataset halves. For each unique half an independent PARAFAC

model was developed; components were matched to all other combinations as well as the complete model. Finally, calculated model residuals were observed to be random with few minor peaks. Model results were reported as F_{max} values in RU. The model was applied to all outliers removed in creating the model, so no samples were excluded from DBP regressions.

2.2.2 Principal component analysis

PCA was carried out in R (V 3.2.5). The dataset used was identical to the one for PARAFAC (including outlier omission). Prior to analysis, excitation/emission pairs were mean centered and scaled to unit variance in order to remove bias towards compounds and spectral regions with higher variability.

2.2.3 Neural networks

In this work neural networks were used both for dimensionality reduction and regression. While the general premise is similar in both applications, the network structures and objectives are distinct. Neural networks were constructed and trained using Google's TensorFlow™, an open source library for machine learning in Python (Abadi et al., 2015). The networks were trained using the Adam optimization algorithm (Kingma and Adam, 2015). Network structure and parameters were chosen based on sequential iterations with the goal of minimizing prediction or reconstruction error and comparability to other

dimensionality reduction techniques. For instance, to allow for the comparison to PARAFAC results, the number of nodes in the latent layer of the autoencoder was set to 5. Two hidden layers of 128 and 64 nodes were used for all trained networks, since this was found to be a suitable compromise between minimizing prediction or reconstruction error without overcomplicating the network structure and making learning good weights difficult.

For networks trained for prediction of DBPs, the cost function used for network training utilized either mean squared error (J_{MSE}) or Huber loss (J_H). Typically, the threshold (δ) for Huber loss is set to 1 and provides a loss function which is more robust and less sensitive to outliers.

$$J_{MSE}(W) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{2}(\hat{y}_i - y_i)^2$$

$$J_H(W) = \frac{1}{n}\sum_{i=1}^{n}\begin{cases}\frac{1}{2}(\hat{y}_i - y_i)^2, & for(\hat{y}_i - y_i) \le \delta \\ \delta\left|(\hat{y}_i - y_i) - \frac{1}{2}\delta^2\right|, & for(\hat{y}_i - y_i) > \delta\end{cases}$$

Where, W is the set of weights in the network

n is the number of samples in the training set

$\hat{y}_i$ is the estimated target value i

y_i is the measured target value i

δ is the threshold separating linear and squared loss

In addition to the error involved in reconstructing x to $\hat{x}$, L1 regularization of the network weights was also applied. On an

intuitive level, L1 regularization penalizes large weights; for every weight in the network, *w*, a term of $\lambda|w|$ is added to the cost function, where λ defines the strength of regularization. This encourages the network to not heavily focus on a few inputs, therefore mitigating overfitting.

All network units, or nodes, contained a rectified linear activation function, which have shown to be both a better model of biological neurons with improved performance and sparsity. In combination with the L1 regularization using rectified linear units (ReLU) further encourages sparsity in the network, which has several computational and representational advantages (Glorot et al., 2011). Since non-zero weights are penalized, the trained network is encouraged to only consider inputs which improve regression accuracy.

$$f_{ReLU}(a) = \begin{cases} a & when\ a > 0 \\ 0 & when\ a \leq 0 \end{cases}$$

Where, a is the node activation value

2.2.4 Autoencoder

The basic premise of an autoencoder is to define a neural network that can recreate a given input through a defined lower dimensional bottleneck. This unsupervised feature learning method allows for limiting information loss while still encoding features in a lower dimensional space. An autoencoder comprises two halves: the encoder and decoder. The encoder approximates a

function to convert an input vector (x) into a lower dimensional representation taken as the output of the latent layer (z) (i.e. $z = f(x)$). The decoder function receives the encoded vector as input and outputs the reconstructed input ($\hat{x}$) (i.e. $\hat{x} = g(z)$) (Figure 1). Through imposing a constrained dimensionality to z, the autoencoder is forced to compress data and cannot simply learn to copy the input perfectly (Goodfellow et al., 2016).

The objective or cost function comprised of reconstruction error, as determined by mean squared error (J_{MSE}), along with L1 weight regularization to prevent overfitting and encourage sparsity.

$$J_{AE}(W) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{2}(\hat{x_i} - x_i)^2 + \lambda\sum_{p=1}^{k}|w_p|$$

Where, w_p is a weight in the network

k is total number of weights across all layers

λ is a set parameter controlling the strength of regularization

n is the number of samples

The autoencoder was developed using the same training set used for PARAFAC and PCA (including outlier omission). Visualization of the latent layer can be achieved by analysis of the weights connected to the nodes in z. This allows for a visual representation of the features being maximally activated by the

latent units. The latent maps or features represented by each latent unit are like loading values produced by PCA or PARAFAC.

$$x_j = \frac{W_{ij}}{\sqrt{\sum_j^d (W_{ij})^2}}$$

Where, *i* is the hidden latent unit in the bottleneck *z*

j is a position in the input vector, i.e. an excitation/emission pair

d is the dimensionality of the input

W_{ij} is the set of weights in the network connected between hidden unit *i* and position *j* of the flattened input vector of dimensionality *d*

3 Results and Discussion

3.1 DIMENSIONALITY REDUCTION

A 5 component PARAFAC model was validated based on methodology described by Murphy et al. (Murphy et al., 2013). One protein-like and four humic-like components were identified (Figure 2). To provide context to the fluorohpores identified by PARAFAC, the components were cross-checked with the OpenFluor database (Kathleen R. Murphy et al., 2014). Characteristics of components 1 – 3 conformed well to terrestrial humic-like substances abundant in surface waters (Kowalczuk et al., 2009; Kathleen R Murphy et al., 2014; Shutova et al., 2014;

354 Stedmon et al., 2003). Evident from the lower fluorescence
355 emissions, C4 likely represents humic-like material arising from
356 biological processes (Murphy et al., 2011; Osburn et al., 2011).
357 The excitation/emission of C5 is typical for tryptophan and
358 therefore representative of protein-like material (Murphy et al.,
359 2011).

360 Using the same dataset, PCA was also applied. As a basis
361 for comparison to PARAFAC and other dimensionality reduction
362 approaches, the number of PCs was constrained to 5. These
363 explained 99.73% of the variance in the dataset, comparable to the
364 99.64% by the 5 component PARAFAC model. Compared to
365 PARAFAC components, those produced by PC were less
366 interpretable in terms of individual fluorophores, evident from the
367 loading plots in Figure 3. Protein-like peaks both in the range of
368 tryptophan and tyrosine were observed in PC4 and PC5. Humic-
369 like fluorophores were not separated by PCA and general
370 representation of humic-like fluorescence in each PC was
371 observed. While physical interpretation is limited when using
372 PCA, it may still provide a lower dimensional representation
373 relevant to predicting formation of DBPs.

374 Latent representations by the autoencoder were more
375 comparable to PCA, where multiple fluorophores are represented
376 in one component and do not necessary conform well to typical

characteristics of organic fluorophores (Figure 4). For instance, LM5 shows the highest representation of peaks in the humic-like regions, with a secondary peak similar to tryptophan (ex/em 280/340 nm). However, the latent maps from the autoencoder show distinction between humic-like peaks (e.g. LM2 and LM5), similar to PARAFAC components. It should be noted that humic-like peaks identified by autoencoder do not conform to PARAFAC components, and this approach has yielded an alternative set of lower dimensional components. Both PCA and the autoencoder emphasized differences in low excitation/emission regions where protein-like fluorescence is expected. In particular, the AE approach identifies fluorescence signals which conform to tryptophan-like characteristics (ex/em 280/340 nm) as well as possibly tyrosine-like fluorescence (ex/em 280/300) in LM2 and a cut-off peak (ex/em 250/300) in LM4. This is contrary to PARAFAC which yields differentiation of humic-like components and only one protein-like component similar to tryptophan.

3.2 PREDICTING DBP FORMATION

Fluorescence data can be used to potentially provide an improved representation of organic composition and reactivity to form disinfection by-products. This hypothesis stems from the increased representation of chemical characteristics in fluorescence EEMs. The excitation-emission maxima and other characteristics

400 are dependent on the fluorophore observed, including its molecular
401 structure, molecular weight, functional groups of compounds, and
402 environment (Baghoth et al., 2011). Better representation of the
403 chemical properties of the OM should therefore improve prediction
404 of the OM reactivity for DBP formation; a process also heavily
405 dependent on the molecular properties and functional groups
406 present, such as aromatic moieties which are implicated as the
407 primary DBP precursors (Hua et al., 2015). Previous work has
408 reported increased correlations between trihalomethanes (THMs)
409 and haloacetic acids (HAAs) with fluorescence measures including
410 PARAFAC components (Hua et al., 2010; Pifer and Fairey, 2012),
411 peak intensities or ratios (Hao et al., 2012; Roccaro et al., 2009),
412 and PCA (Peleato and Andrews, 2015). However, results
413 presented to-date have often been limited by linear correlation
414 strength on all samples (i.e. no separation of a test dataset) and
415 utilizing samples with similar organic characteristics. The reduced
416 accuracy in DBP prediction shown when using validation sites (i.e.
417 sites which were not included in the model training) have been
418 observed when applying binary classification trees, exemplifying
419 the importance of considering a validation set (Bergman et al.,
420 2016). We address these limitations by using a dataset that
421 includes water treated by coagulation, ozone, H_2O_2 + O_3, and
422 biofiltration. Pre-oxidation by ozone or H_2O_2 + O_3 impacts

organic character or structure significantly, although the overall DOC or mass of organics is not expected to change to a large extent. Furthermore, to ensure a more accurate assessment of predictive power of the organic measures and modelling approach, separation of a validation (20%, n = 24) and training (80%, n = 96) datasets was carried out by random selection. The validation set was not used in dimensionality reduction analysis or modelling of DBP formation. A 10-fold cross-validation on the training dataset approach was used to determine optimal model parameters such as learning rate or the number of nodes in each layer. All input variables were normalized to the range of 0 to 1.

3.2.1 Prediction with data pre-treatment

The possible role of dimensionality reduction in improving DBP formation prediction was investigated. Separate neural networks were trained using four versions of fluorescence information: 1) baseline no dimensionality reduction (full spectrum), 2) PARAFAC component scores, 3) PCA component scores, and 4) output of the 5 latent autoencoder nodes. The accuracy with varying data pre-treatments both from cross-validation and on the validation dataset are shown in Table 1 and Figure 5. Further to testing data pre-treatments, comparison of using the Huber loss or squared error cost function was examined. Consistently Huber loss had superior performance on the

validation set with lower accuracy in cross-validation. This observation demonstrated the advantage of using a robust error function and prevented some degree of overfitting. The exception was improved performance of squared error when using the full EEM spectrum for predicting both THMs and HAAs.

For prediction of THMs, optimum validation performance (MAE: 7.46 µg L^{-1}) was observed using spectral data pre-processed by an autoencoder with comparable performance (MAE: 7.97 µg L^{-1}) using the full EEM. Dimensionality reduction with PARAFAC resulted in the poorest performance (MAE: 20.24 µg L^{-1}), resulting in loss of accuracy compared to the unprocessed full spectrum. Based on variance of predictions between all CV-folds on the validation data, all MAE differences were significant as determined by t-tests ($p < 0.024$). This observation suggests loss of information related to THM precursors through the application of PARAFAC and constraints of interpretable components. Pre-treatment with an autoencoder was observed to result in the most robust regression, with the lowest discrepancy between CV and validation set error rates (CV MAE: 4.87 µg L^{-1}, validation MAE: 7.46 µg L^{-1}).

Predictability of total HAA formation was consistently lower compared to THMs. Prediction accuracy on the validation set varied less across all pre-processing approaches (10.75 to 14.22

469 µg L^{-1} MAE). For HAA prediction, pre-processing was not found
470 to improve regression accuracy and utilizing the full spectrum
471 resulted in the greatest CV and validation MAE. It should be
472 considered that while pre-processing and organic surrogates are
473 being compared in this analysis, other factors influence DBP
474 formation, such as pH, have not been included in the models.
475 The uniqueness of the separated validation dataset should
476 be considered when assessing the model performance. It should be
477 noted that when considering the variance between CV folds (29.6
478 to 44.6% coefficient of variation), comparisons of pre-treatment
479 methods were not found to be significant ($p > 0.05$). However, the
480 validation dataset was separated initially and not utilized for
481 developing the dimensionality reduction models. As such, we
482 believe along with a larger test size (validation n = 24; CV test n =
483 9-10), the emphasis should be on comparison of validation dataset
484 error. With each CV fold, prediction on the validation data was
485 also carried out. Considering the variability imparted by data used
486 for training, all comparisons of the validation MAE were found to
487 be significant ($p < 0.05$) for both THMs and HAAs.
488 The role of NN regression was determined through
489 comparison with a conventional multi linear regression (MLR)
490 method. The fluorescence results derived from dimensionality
491 reduction were used as the multi-variate inputs to a multi linear

492 regression model. Accuracy of the AE, PCA, and PARAFAC
493 derived scores in multi linear regression models are reported in
494 Table 2. Validation accuracy using MLR was comparable for each
495 data pre-treatment. A consistent trend of data pre-treatment
496 performance on the validation dataset from best to worse was AE >
497 PCA > PARAFAC. This relationship was less pronounced for CV
498 error rates, particularly for HAA prediction. Improvement of
499 validation accuracy with AE-NN regression vs MLR for THM
500 prediction (7.46 µg L^{-1} vs 9.64 µg L^{-1}) was contrasted to a decrease
501 in prediction accuracy for HAAs (11.93 µg L^{-1} vs. 9.64 µg L^{-1}).
502 However, for all cases the MAE from cross-validation was greater
503 using MLR (13.52 to 20.92 µg L^{-1}) compared to NN regression
504 (3.08 to 6.33 µg L^{-1}). This suggests on average, between all folds
505 during cross-validation, NN regression may have advantages
506 despite the comparable performance on the validation dataset.
507 Trueman et al. (2016) used a comparable cross-validation approach
508 and bench-scale samples subjected to advanced oxidation, with
509 reported CV MAE ≥ 9.5 µg L^{-1}.

510 ### 3.2.2 Comparison to conventional organic measures

511 The performance of the fluorescence/neural network
512 approach was compared to baseline models which utilize
513 conventional organic measures of DOC, UVA (at 254 nm), and
514 SUVA. Overall linear model strength between DOC and UVA

515 with THM concentrations were moderate (R^2: 0.65 and 0.56,
516 respectively). The model strength or correlations between DOC or
517 UVA with THMs were lower compared to those reported by Li et
518 al. (2016) (DOC R^2: 0.89; UVA R^2: 0.79), which included 16
519 drinking water sources as well as coagulation and anion exchange
520 treatments. This supports our expectation that the advanced
521 oxidation treatments resulted in significant changes to organic
522 character, while not altering overall measures such as DOC. Using
523 a linear model, validation error was minimized using DOC (MAE:
524 15.15 µg L^{-1}) however it was over 2 times greater when compared
525 to the autoencoder/fluorescence. As shown in Figure 6, UVA
526 resulted in groupings of THM predictions and indicate that this
527 measure did not capture organic properties which result in THM
528 formation. To establish that the difference in performance was not
529 due to a linear model vs. neural network regression, a neural
530 network with DOC and UVA as inputs was trained. Validation
531 error was comparable to the linear model, however increased CV
532 performance was observed.

533 Correlations with total HAA formation were found to be
534 low (R^2 0.09 to 0.48) although validation set error rates were
535 comparable to fluorescence results using both NN regression and
536 MLR. This is possibly due to the decreased range in HAA
537 formation, 28.1 to 139.5 µg L^{-1} HAAs vs. 26.5 to 208.2 µg L^{-1}

538 THMs. The comparable accuracy between organic surrogates and
539 regression approach may also suggest that HAA formation is more
540 significantly dependent on other factors that have not been
541 included in the model such as pH.

542 ### 3.3 FLUORESCENCE REGIONAL IMPORTANCE FOR
543 DBP FORMATION

544 Through the established weights in the models, it is of
545 interest to understand the relative contributions of each input to the
546 predictability of DBPs. The process of determining variable
547 importance was carried out using the Connection Weight Approach
548 described by Olden and Jackson (Olden and Jackson, 2002) and
549 Olden et al. (Olden et al., 2004). For each input, the product of
550 connected weights between the network layers is calculated. This
551 was performed 20 times with different random weight
552 initializations for every constructed network. Normalization of the
553 calculated variable importance was conducted to diminish
554 variability based on the absolute value of the initial network
555 weights. The relative input variable importance using varying data
556 pre-processing methods are shown in Figure 7. Ranking of
557 PARAFAC variables by connection weights shows predominant
558 positive association between humic-like fluorophores with THM
559 and HAA formation. C4 was observed to have the highest positive
560 connection weights for THM prediction, indicating this terrestrial

561 humic-like fluorophore with one excitation band is likely a major
562 THM precursor. Based on the HAA model, increased importance
563 of C4 ($p < 0.01$) and increased negative association with C3 ($p <$
564 0.01) were noted. This suggests stronger association between
565 humic-like substance from possible microbial origins and HAA
566 formation. Negative associations with humic-like C3 and protein-
567 like C5 were observed. C3 in particular is unique in the high
568 emission characteristics > 450 nm. Through comparison to
569 characterization by ultra-high resolution mass spectrometry, it has
570 been suggested that fluorophores emitting above 450 nm likely
571 have greater average carbon oxidation states (≥ 0) and higher
572 double bond equivalency per carbon (Lavonen et al., 2015).
573 Presence of oxidized organic material is expected based on the
574 dataset containing samples which have been treated with ozone or
575 an advanced oxidation process. The method used here illustrates
576 sensitivity to identifying fluorescence signal regions associated
577 with decreased DBP formation potential from the application of
578 strong pre-oxidants. A visualization of the fluorescence regions
579 associated with DBP formation is shown as Figure 8, which were
580 calculated through weighted reconstruction of EEMs using the
581 loading values and relative variable importance. Based on
582 PARAFAC, positive correlations with humic-like regions in the
583 ex/em region of 250-340/375-450 nm and THM/HAA formation

584 can be seen. The negative association between protein-like
585 fluorescence and DBP formation is also illustrated.

586 Variable importance using the latent maps from the AE
587 (Figure 7) is less interpretable due to the ambiguity of fluorophore
588 representation in each latent variable. The visualization of
589 fluorescence regions weighted by the autoencoder-neural network
590 aided in determining variable importance (Figure 8). Generally,
591 there is negative association between fluorescence < ex/em
592 260/310 nm and THM/HAA formation, however positive
593 connection weights are seen with tryptophan-like fluorescence at
594 ex/em 280/340 nm. This observation is contrary to the results from
595 PARAFAC, in particular, increased importance of tryptophan-like
596 fluorescence for HAA formation prediction was observed when
597 using the AE, full spectrum, and PCA approaches. Furthermore,
598 autoencoder-neural network regression placed high positive
599 weights to high emission regions > 550 nm.

600 Representation of the full EEM weighted connections
601 yielded a noisier but more nuanced image of fluorescence regions
602 associated with DBP formation (Figure 8). Similar to PARAFAC
603 and PCA but contrary to the autoencoder, humic-like peaks with
604 emissions ~450 nm had positive weightings. Specific low ex/em
605 peaks in the protein-like region were also identified to have
606 positive weights. Pronounced high relative weights at

607 approximately ex/em 280/310 nm and 380/436 nm correspond well
608 to expected Raman peaks from water. While the EEMs were first
609 pre-processed to remove influence of Rayleigh and Raman regions,
610 artifacts may have remained which were identified by the model to
611 be positively correlated with DBP formation. Comparatively to the
612 autoencoder regions, fluorescence at high emissions > 550 nm
613 were also positively associated with both THM and HAA
614 formation.

615 Evident from the contradicting regions associated with
616 DBP formation regression is the influence of the pre-processing
617 method. Regions identified by PARAFAC conform to
618 expectations of types of organic material likely to result in
619 formation of DBPs and are most interpretable. However, increased
620 performance of the autoencoder or using the full EEM when
621 predicting THMs and HAA formation on the validation dataset
622 using both NN regression and MLR gives credence that these
623 approaches were better able to include fluorescence regions
624 associated with DBP formation. Our interpretation of the non-
625 conformance of these results is that significant consideration of the
626 pre-processing method should be taken when interpreting reduced-
627 dimensionality EEM results. We hypothesize that due to the
628 apparent influence of data pre-processing, utilizing the full
629 spectrum with weight normalization to encourage relevant input

selection may result in a more accurate representation of fluorescence regions associated with NOM reactivity to form DBPs.

4 Conclusions

A NN approach to both dimensionality reductions, utilizing an AE as well as for DBP formation regression was shown to be advantageous. Results on a randomly separated validation data set indicate that, while PARAFAC produces components which resemble organic fluorophores, the constrained dimensionality approach likely results in information loss that improves prediction of both total THMs and total HAAs. Compared to common organic measures an AE-NN regression provides greater training and validation set prediction accuracies for THMs and similar performance for HAAs. AE dimensionality reduction appears to potentially mitigate overfitting based on minor differences between CV training error and validation errors. Comparison of MLR to NN yields similar accuracy on validation data, indicating that pre-treatment methods should be emphasized, and the regression approach may not be as important. Through analysis of the connection weights, variable importance can be quantified allowing for greater understanding regarding how the trained NN model functions. Particularly through the more interpretable

PARAFAC components, differing positive and negative correlations between components and DBP formation was observed. While humic-like fluorophores or fluorescence regions were generally observed to be associated with DBP formation, a PARAFAC component likely representing organic material transformed by an oxidation process was negatively associated with formation potentials.

Results presented in this study suggest the novel applicability of autoencoders for interpretation of fluorescence results. Compared to PARAFAC analysis, autoencoders produced components with more limited in interpretability, however resulted in increased representation of the data as evidenced from improved DBP formation prediction. While autoencoders optimized prediction of THMs, utilizing the full spectrum without any prior dimensionality reduction was observed to result in the greatest performance for HAAs in this study. Furthermore, improved DBP formation prediction using a NN approach was observed compared to linear regression typically practiced. The approach taken in this work is well suited for handling large and high-dimensional datasets, which are increasingly common. Furthermore, the possible use of fluorescence as a continuous monitoring device will require flexible, robust, and scalable analysis methods.

5 Acknowledgments

This work was funded in part by the Canadian Water Network and the Natural Sciences and Engineering Research Council of Canada (NSERC) Chair in Drinking Water Research at the University of Toronto. We would like to thank the personnel at the Peterborough Utilities Commission for their continuing support of pilot studies.

6 References

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G., Isard, M., Jia, Y., Kaiser, L., Kudlur, M., Levenberg, J., Man, D., Monga, R., Moore, S., Murray, D., Shlens, J., Steiner, B., Sutskever, I., Tucker, P., Vanhoucke, V., Vasudevan, V., Vinyals, O., Warden, P., Wicke, M., Yu, Y., Zheng, X., 2015. TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems. None 1, 19. doi:10.1038/nn.3331

APHA/AWWA/WEF, 2012. Standard Methods for the Examination of Water and Wastewater. Stand. Methods 541. doi:ISBN 9780875532356

Baghoth, S.A., Sharma, S.K., Amy, G.L., 2011. Tracking natural

696 organic matter (NOM) in a drinking water treatment plant
697 using fluorescence excitation-emission matrices and
698 PARAFAC. Water Res. 45, 797–809.
699 doi:10.1016/j.watres.2010.09.005
700 Bergman, L.E., Wilson, J.M., Small, M.J., VanBriesen, J.M., 2016.
701 Application of Classification Trees for Predicting Disinfection
702 By-Product Formation Targets from Source Water
703 Characteristics. Environ. Eng. Sci. 33, 455–470.
704 doi:10.1089/ees.2016.0044
705 Bieroza, M., Baker, A., Bridgeman, J., 2011. Classification and
706 calibration of organic matter fluorescence data with multiway
707 analysis methods and artificial neural networks: An
708 operational tool for improved drinking water treatment.
709 Environmetrics 22, 256–270. doi:10.1002/env.1045
710 Bridgeman, J., Bieroza, M., Baker, A., 2011. The application of
711 fluorescence spectroscopy to organic matter characterisation
712 in drinking water treatment. Rev. Environ. Sci.
713 Bio/Technology 10, 277–290. doi:10.1007/s11157-011-9243-
714 x
715 Bro, R., 1997. PARAFAC. Tutorial and applications, in:
716 Chemometrics and Intelligent Laboratory Systems. pp. 149–
717 171. doi:10.1016/S0169-7439(97)00032-4
718 Chowdhury, S., Champagne, P., McLellan, P.J., 2009. Models for

719 predicting disinfection byproduct (DBP) formation in
720 drinking waters: A chronological review. Sci. Total Environ.
721 doi:10.1016/j.scitotenv.2009.04.006
722 Fabris, R., Chow, C.W.K., Drikas, M., Eikebrokk, B., 2008.
723 Comparison of NOM character in selected Australian and
724 Norwegian drinking waters. Water Res. 42, 4188–4196.
725 doi:10.1016/j.watres.2008.06.023
726 Glorot, X., Bordes, A., Bengio, Y., 2011. Deep sparse rectifier
727 neural networks. AISTATS ’11 Proc. 14th Int. Conf. Artif.
728 Intell. Stat. 15, 315–323. doi:10.1.1.208.6449
729 Goodfellow, I., Bengio, Y., Courville, A., 2016. Deep Learning
730 [WWW Document]. MIT Press. URL
731 http://www.deeplearningbook.org/
732 Hao, R., Ren, H., Li, J., Ma, Z., Wan, H., Zheng, X., Cheng, S.,
733 2012. Use of three-dimensional excitation and emission
734 matrix fluorescence spectroscopy for predicting the
735 disinfection by-product formation potential of reclaimed
736 water. Water Res. 46, 5765–5776.
737 doi:10.1016/j.watres.2012.08.007
738 Hinton, G.E., Salakhutdinov, R.R., 2006. Reducing the
739 Dimensionality of Data with Neural Networks. Science (80-.
740). 313, 504–507. doi:10.1126/science.1127647
741 Hua, B., Veum, K., Yang, J., Jones, J., Deng, B., 2010. Parallel

742 factor analysis of fluorescence EEM spectra to identify THM
743 precursors in lake waters. Environ. Monit. Assess. 161, 71–
744 81. doi:10.1007/s10661-008-0728-1
745 Hua, G., Reckhow, D.A., 2007. Characterization of Disinfection
746 Byproduct Precursors Based on Hydrophobicity and
747 Molecular Size. doi:10.1021/ES062178C
748 Hua, G., Reckhow, D.A., Abusallout, I., 2015. Correlation between
749 SUVA and DBP formation during chlorination and
750 chloramination of NOM fractions from different sources.
751 Chemosphere 130, 82–89.
752 doi:10.1016/j.chemosphere.2015.03.039
753 Kingma, D.P., Adam, J.B., 2015. A method for stochastic
754 optimization, in: International Conference on Learning
755 Representation. doi:10.1109/ICCCBDA.2017.7951902
756 Kothawala, D.N., Murphy, K.R., Stedmon, C.A., Weyhenmeyer,
757 G.A., Tranvik, L.J., 2013. Inner filter correction of dissolved
758 organic matter fluorescence. Limnol. Oceanogr. Methods 11,
759 616–630. doi:10.4319/lom.2013.11.616
760 Kowalczuk, P., Durako, M.J., Young, H., Kahn, A.E., Cooper,
761 W.J., Gonsior, M., 2009. Characterization of dissolved
762 organic matter fluorescence in the South Atlantic Bight with
763 use of PARAFAC model: Interannual variability. Mar. Chem.
764 113, 182–196. doi:10.1016/j.marchem.2009.01.015

765 Lavonen, E.E., Kothawala, D.N., Tranvik, L.J., Gonsior, M.,
766 Schmitt-Kopplin, P., Köhler, S.J., 2015. Tracking changes in
767 the optical properties and molecular composition of dissolved
768 organic matter during drinking water production. Water Res.
769 85, 286–294. doi:10.1016/j.watres.2015.08.024
770 Lawaetz, A.J., Stedmon, C.A., 2009. Fluorescence Intensity
771 Calibration Using the Raman Scatter Peak of Water. Appl.
772 Spectrosc. 63, 936–940. doi:10.1366/000370209788964548
773 Li, W.T., Jin, J., Li, Q., Wu, C.F., Lu, H., Zhou, Q., Li, A.M.,
774 2016. Developing LED UV fluorescence sensors for online
775 monitoring DOM and predicting DBPs formation potential
776 during water treatment. Water Res. 93, 1–9.
777 doi:10.1016/j.watres.2016.01.005
778 Matilainen, A., Gjessing, E.T., Lahtinen, T., Hed, L., Bhatnagar,
779 A., Sillanpää, M., 2011. An overview of the methods used in
780 the characterisation of natural organic matter (NOM) in
781 relation to drinking water treatment. Chemosphere.
782 doi:10.1016/j.chemosphere.2011.01.018
783 Murphy, K.P., 2012. Machine Learning: A Probabilistic
784 Perspective, MIT Press. doi:10.1007/978-3-642-21004-4_10
785 Murphy, K.R., Hambly, A., Singh, S., Henderson, R.K., Baker, A.,
786 Stuetz, R., Khan, S.J., 2011. Organic Matter Fluorescence in
787 Municipal Water Recycling Schemes: Toward a Unified

788 PARAFAC Model. Environ. Sci. Technol. 45, 2909–2916.
789 doi:10.1021/es103015e
790 Murphy, K.R., Stedmon, C.A., Bro, R., 2014. Chemometric
791 analysis of organic matter fluorescence, in: Aquatic Organic
792 Matter Fluorescence. pp. 339–375.
793 doi:10.13140/2.1.2595.8080
794 Murphy, K.R., Stedmon, C.A., Graeber, D., Bro, R., 2013.
795 Fluorescence spectroscopy and multi-way techniques.
796 PARAFAC. Anal. Methods 5, 6557–6566.
797 doi:10.1039/c3ay41160e
798 Murphy, K.R., Stedmon, C.A., Wenig, P., Bro, R., 2014.
799 OpenFluor- an online spectral library of auto-fluorescence by
800 organic compounds in the environment. Anal. Methods 6,
801 658–661. doi:10.1039/c3ay41935e
802 Olden, J.D., Jackson, D.A., 2002. Illuminating the “black box”: A
803 randomization approach for understanding variable
804 contributions in artificial neural networks. Ecol. Modell. 154,
805 135–150. doi:10.1016/S0304-3800(02)00064-9
806 Olden, J.D., Joy, M.K., Death, R.G., 2004. An accurate
807 comparison of methods for quantifying variable importance in
808 artificial neural networks using simulated data. Ecol. Modell.
809 178, 389–397. doi:10.1016/j.ecolmodel.2004.03.013
810 Osburn, C.L., Wigdahl, C.R., Fritz, S.C., Saros, J.E., 2011.

811 Dissolved organic matter composition and photoreactivity in
812 prairie lakes of the U.S. Great Plains. Limnol. Oceanogr. 56,
813 2371–2390. doi:10.4319/lo.2011.56.6.2371
814 Peiris, B.R.H., Budman, H., Moresoli, C., Legge, R.L., 2009.
815 Acquiring reproducible fluorescence spectra of dissolved
816 organic matter at very low concentrations. Water Sci.
817 Technol. 60.
818 Peleato, N.M., Andrews, R.C., 2015. Comparison of three-
819 dimensional fluorescence analysis methods for predicting
820 formation of trihalomethanes and haloacetic acids. J. Environ.
821 Sci. 27, 159–167. doi:10.1016/j.jes.2014.04.014
822 Peleato, N.M., Sidhu, B.S., Legge, R.L., Andrews, R.C., 2017.
823 Investigation of ozone and peroxone impacts on natural
824 organic matter character and biofiltration performance using
825 fluorescence spectroscopy. Chemosphere 172, 225–233.
826 doi:10.1016/j.chemosphere.2016.12.118
827 Pifer, A.D., Fairey, J.L., 2012. Improving on SUVA 254 using
828 fluorescence-PARAFAC analysis and asymmetric flow-field
829 flow fractionation for assessing disinfection byproduct
830 formation and control. Water Res. 46, 2927–2936.
831 doi:10.1016/j.watres.2012.03.002
832 Rhee, J. Il, Lee, K.-I., Kim, C.-K., Yim, Y.-S., Chung, S.-W., Wei,
833 J., Bellgardt, K.-H., 2005. Classification of two-dimensional

fluorescence spectra using self-organizing maps. Biochem. Eng. J. 22, 135–144. doi:10.1016/j.bej.2004.09.008

Roccaro, P., Vagliasindi, F.G.A., Korshin, G. V., 2009. Changes in NOM fluorescence caused by chlorination and their associations with disinfection by-products formation. Environ. Sci. Technol. 43, 724–729. doi:10.1021/es801939f

Shutova, Y., Baker, A., Bridgeman, J., Henderson, R.K., 2014. Spectroscopic characterisation of dissolved organic matter changes in drinking water treatment: From PARAFAC analysis to online monitoring wavelengths. Water Res. 54, 159–169. doi:10.1016/j.watres.2014.01.053

Stedmon, C.C. a, Markager, S., Bro, R., 2003. Tracing dissolved organic matter in aquatic environments using a new approach to fluorescence spectroscopy. Mar. Chem. 82, 239–254. doi:10.1016/S0304-4203(03)00072-0

Trueman, B.F., MacIsaac, S.A., Stoddart, A.K., Gagnon, G.A., 2016. Prediction of disinfection by-product formation in drinking water via fluorescence spectroscopy. Environ. Sci. Water Res. Technol. 2, 383–389. doi:10.1039/C5EW00285K

Wolf, G., Almeida, J.S., Crespo, J.G., Reis, M.A.M., 2007. An improved method for two-dimensional fluorescence monitoring of complex bioreactors. J. Biotechnol. 128, 801–812. doi:10.1016/j.jbiotec.2006.12.029

857

Table 1 Cross–validation (CV) and validation results for neural networks with varying data pre-treatments and cost function. MSE: mean squared error, MAE: mean absolute error, AE: autoencoder, PCA: principle component analysis, PARAFAC: parallel factors analysis, HL: Huber-loss, SE: squared error.

Data pre-treatment	**CV MSE (µg/L)^2**		**CV MAE (µg/L)**		**Validation MSE (µg/L)^2**		**Validation MAE (µg/L)**	
	HL	SE	HL	SE	HL	SE	HL	SE
THMs								
Full spectrum	66.91	36.03	3.70	3.29	334.85	127.09	9.82	7.97
AE	77.48	64.41	4.87	4.96	120.03	198.07	7.46	11.93
PCA	82.57	61.98	4.80	5.43	268.92	245.76	13.39	12.32
PARAFAC	167.76	96.70	6.33	6.51	753.01	435.07	20.24	16.39
HAAs								
Full spectrum	25.45	17.10	3.08	2.74	173.95	159.44	10.75	10.28
AE	49.11	32.05	4.97	4.17	195.53	329.66	11.93	15.23
PCA	47.63	25.08	5.05	3.71	177.67	249.56	11.85	12.53
PARAFAC	68.00	36.39	4.74	4.45	363.81	348.93	14.22	18.81

Table 2 Cross-validation and validation results (MAE) for multi linear regression using fluorescence data pre-processed by a dimensionality reduction method.

Data pre-treatment	**CV MAE ($\mu g\ L^{-1}$)**	**Validation MAE ($\mu g\ L^{-1}$)**
THMs		
AE	18.34	9.65
PCA	20.65	13.19
PARAFAC	20.92	20.39
HAAs		
AE	13.52	9.64
PCA	14.49	11.92
PARAFAC	13.63	14.00

Table 2 Cross–validation (CV) and validation results for linear models with conventional organic measures. MSE: mean squared error, MAE: mean absolute error.

Organic measure	**CV MSE (µg/L)^2**	**CV MAE (µg/L)**	**Validation MSE (µg/L)^2**	**Validation MAE (µg/L)**	**Full dataset R^2**
THMs					
DOC	492.39	16.13	303.26	15.15	0.65
UVA	525.69	17.57	524.82	17.59	0.56
SUVA	859.87	21.85	864.25	21.13	0.29
DOC + UVA, neural network	227.33	10.23	365.33	16.33	-
HAAs					
DOC	227.22	11.97	303.53	12.03	0.48
UVA	267.81	13.33	396.97	14.46	0.30
SUVA	312.80	14.57	466.70	15.79	0.09
DOC + UVA, neural network	84.12	6.94	197.83	10.18	-

Figure 1 Schematic of an example autoencoder structure with one hidden layer and latent layer (z) with two nodes.

Figure 2 Loading plots for the 5 identified PARAFAC components.

Figure 3 Loading plots from PCA

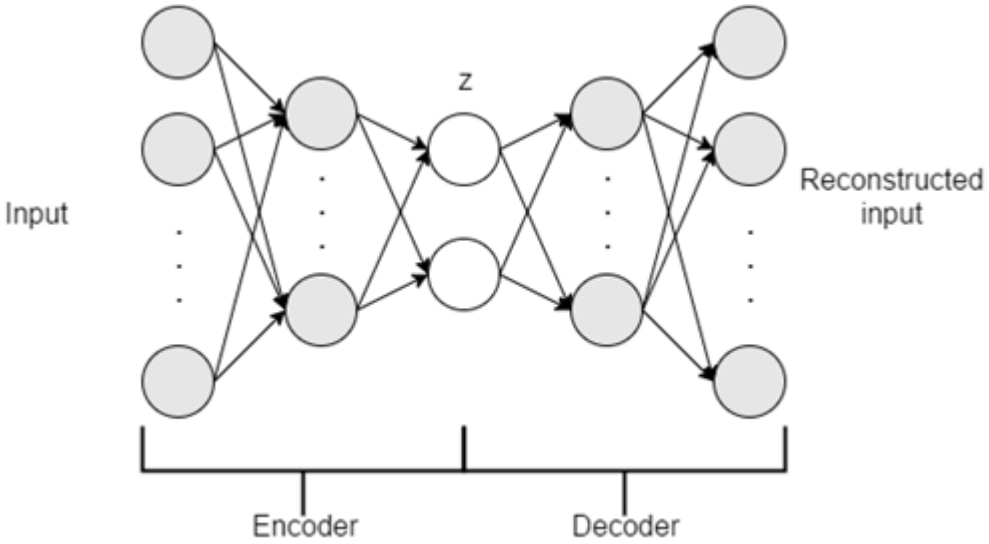
Figure 4 Latent maps from the constrained layer of the autoencoder

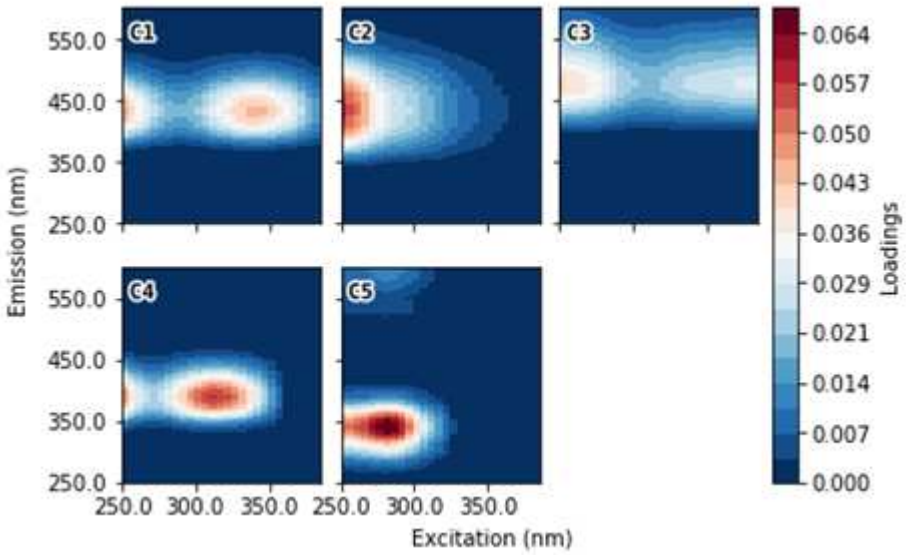
Figure 5 Measured vs. predicted THMs for example models using varying data pre-treatments. Circles represent samples in the training dataset; + represent samples from the validation dataset.

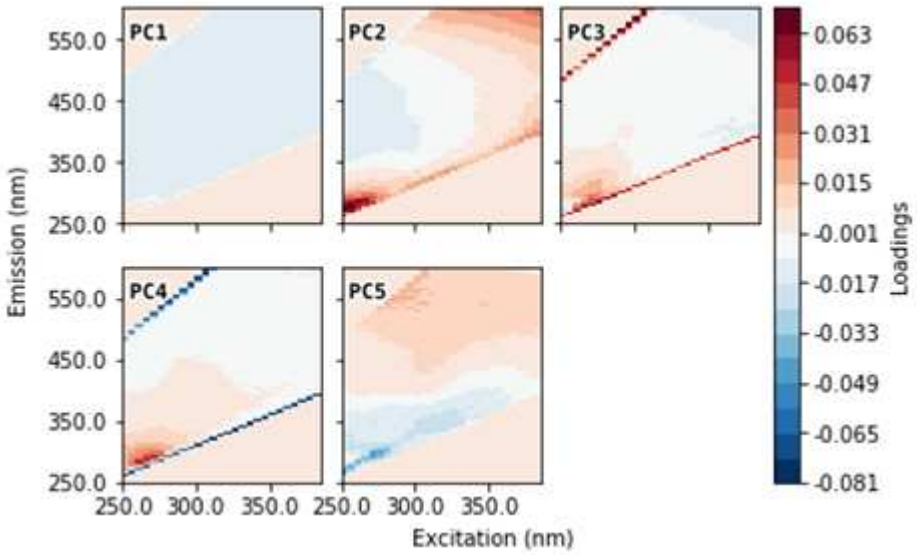
Figure 6 Measured vs. predicted THMs using conventional organic measures. Circles represent samples in the training dataset; + represent samples from the validation dataset.

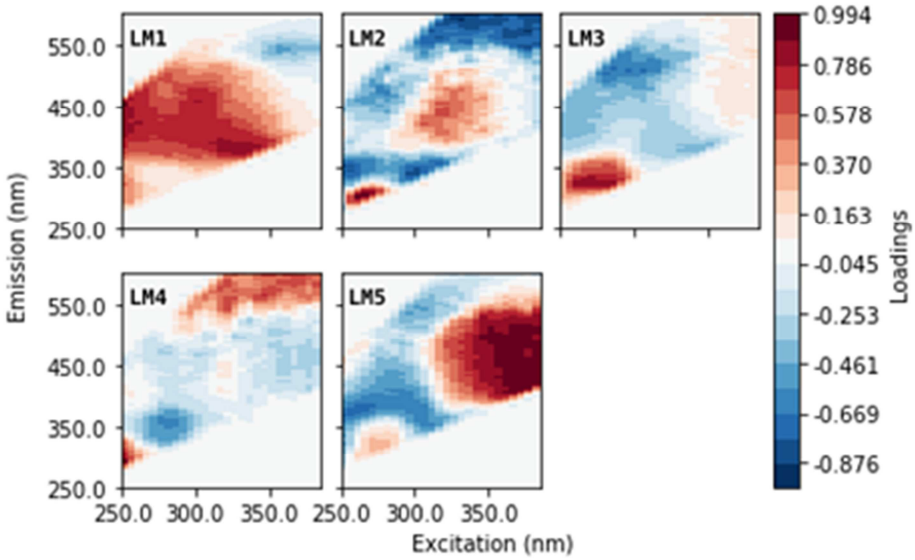
Figure 7 Relative importance of input variables calculated based on connection weights. Vertical bars represent one standard deviation from the 20 random initializations.

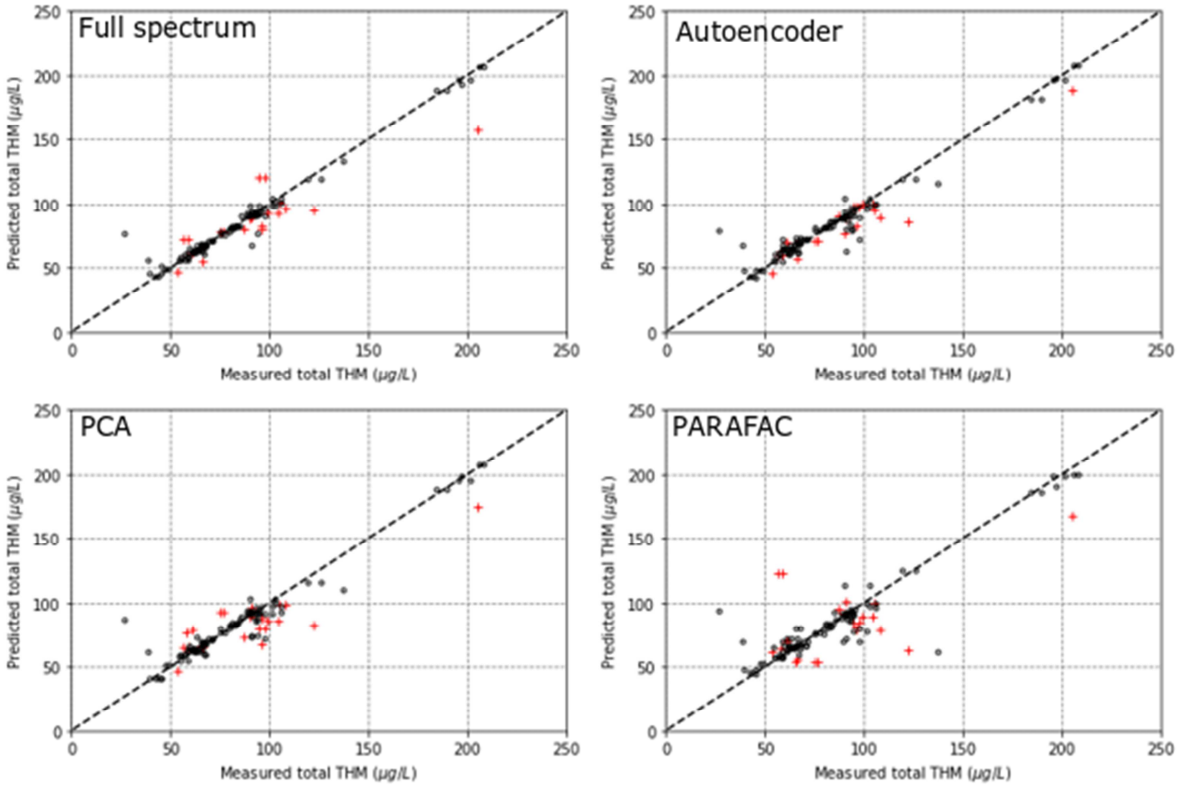
Figure 8 Mappings of fluorescence regions of relative importance for the prediction of THMs and HAAs. a) autoencoder, THMs; b) autoencoder, HAAs; c) full EEM, THMs; d) full EEM, HAAs, e) PARAFAC, THMs; f) PARAFAC, HAAs; g) PCA, THMs; h) PCA, HAAs.

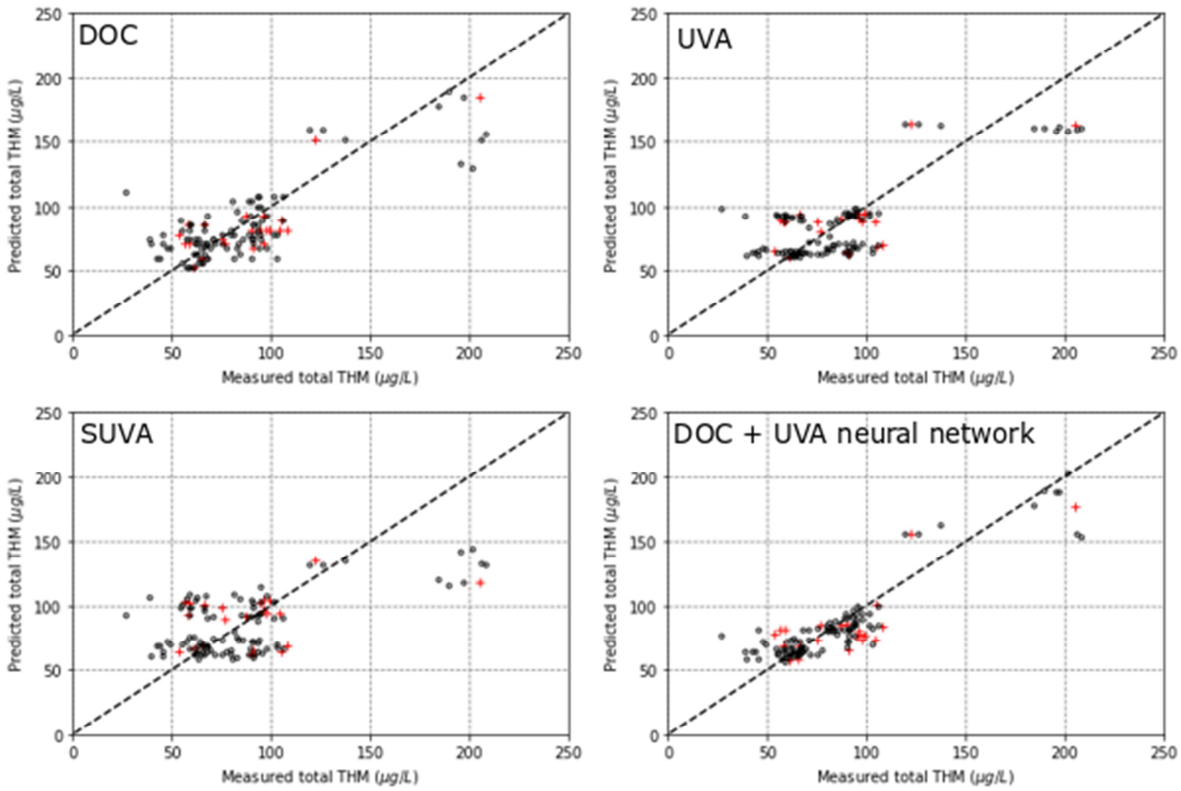


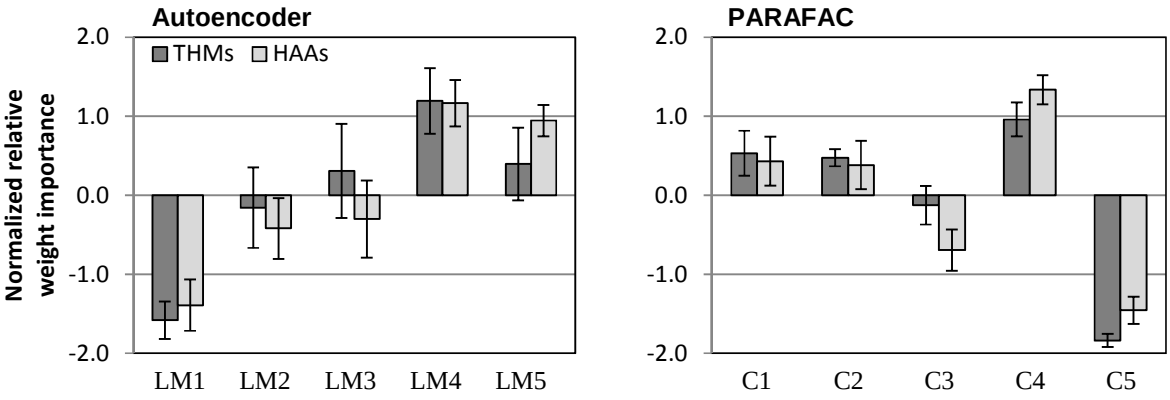


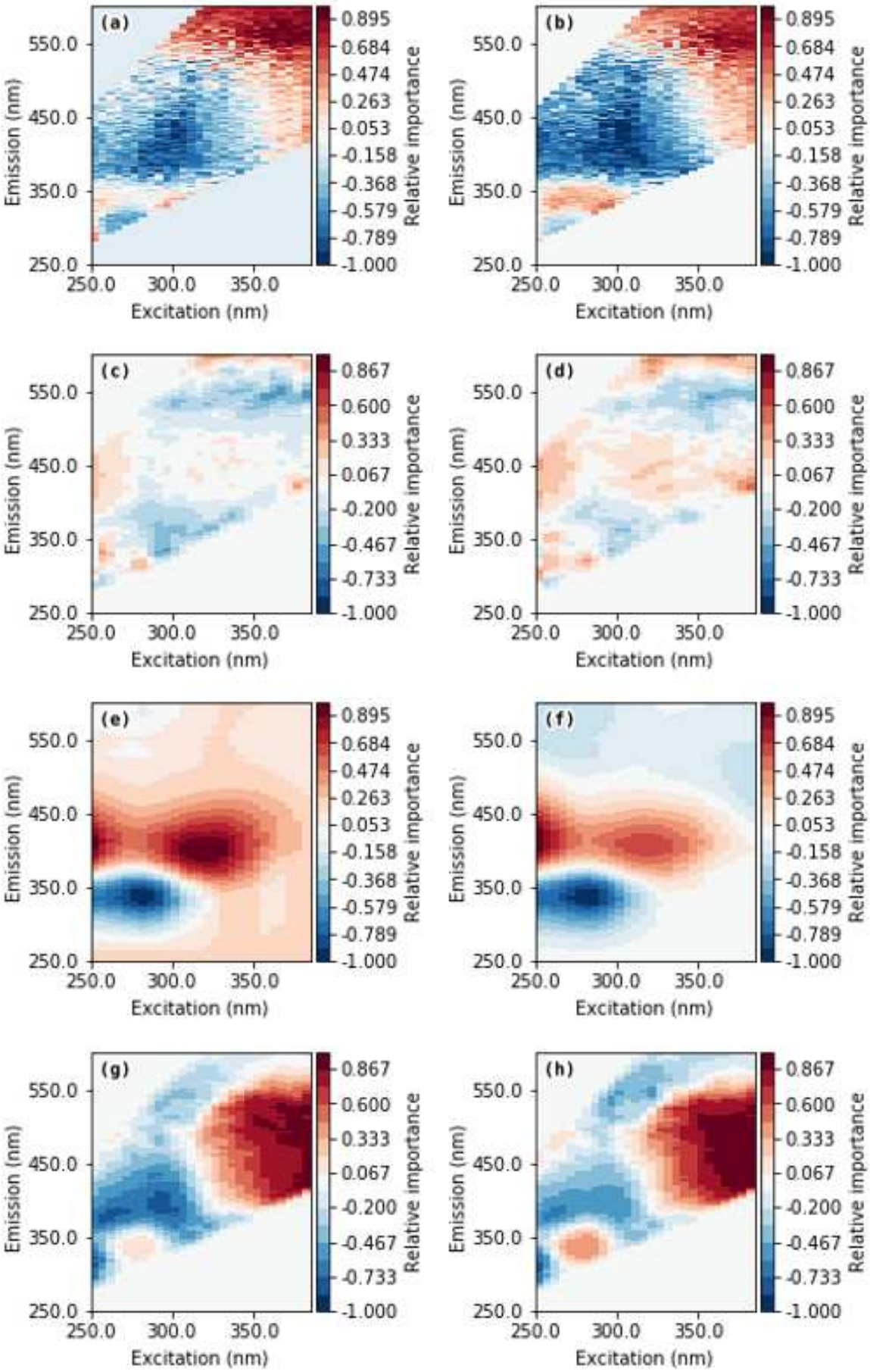











(a)
(b)
(c)
(d)
(e)
(f)
(g)
(h)
Emission (nm)
Excitation (nm)
Relative importance
250.0
300.0
350.0
450.0
550.0
0.895
0.684
0.474
0.263
0.053
-0.158
-0.368
-0.579
-0.789
-1.000
0.867
0.600
0.333
0.067
-0.200
-0.467
-0.733
-1.000

Highlights

- Autoencoder applied for dimensionality reduction of fluorescence spectra
- Improved DBP formation prediction using autoencoder components or full spectrum
- PARAFAC produced interpretable components, however poor reactivity prediction
- Improved cross-validation accuracy using neural networks for regression
- Neural network weights identify fluorescence regions associated with DBP formation