Machine Learning Assisted Classification of Ca-Organic Matter Complexation Using X-ray Spectroscopies

Call Topic: Computing, Analytics and Modeling

Specific Aims: The subsurface (soils and sediments) hosts the largest dynamic store of organic carbon that can be released to the atmosphere upon mineralization. Ca has the potential to play a key role in preventing mineralization of this organic carbon by enhancing organic sorption to mineral phases (termed mineral protection), as it has long been recognized that Ca forms cation bridges that link together negatively charged functional groups from organics and mineral surfaces. A recent meta-study [1] found that exchangeable Ca was a good predictor of organic matter (OM) content for certain soils. Moreover, Ca bridging has been linked to decreased OM mineralization [2]. However, there is little experimental insight into the molecular-scale mechanisms and under what conditions Ca bridging occurs. This is in part due to the complexity of OM, which is a macromolecular assembly containing numerous functional groups capable of binding metals. Molecular insight into Ca-OM binding is essential to provide predictive knowledge on the conditions under which Ca stabilizes organic matter within soils and sediments.

In this work, we propose to shed light on Ca binding to OM, and thus the role that it plays in OM sequestration by utilizing recently developed machine learning (ML) tools combined with X-ray spectroscopies. Specifically, we will use this approach to identify sensitive spectral fingerprints that can be used to distinguish which OM functional groups participate in Ca bridging. The structural and electronic structure information gained will be used to develop a comprehensive chemical and structural classification of Ca-organic complexation, which will provide insight into Ca-OM binding. We believe that this state-of-the-art approach can be used to aid the elucidation of the following hypotheses: **a)** Ca forms inner-sphere complexes with OM functional groups; and **b)** Ca is bound predominantly by carboxylate and catechol functional groups. We will utilize computational and experimental capabilities at EMSL and SSRL to achieve this goal.

Mission Relevance: This work is aligned with the BER and EMSL mission to develop a mechanistic understanding of molecular scale processes controlling ecosystem function, namely, Ca-organic matter interactions and their ability to stabilize subsurface carbon pools, thereby mitigating soil carbon efflux. This work is in line with the overarching EMSL mission area on Environmental Transformations and Interactions.

Background: X-ray spectroscopy, an important chemical speciation technique, has seen impressive recent developments using ML [3,4]. Briefly, X-ray absorption spectroscopy (XAS) encompasses both X-ray absorption near edge structure (XANES) and extended X-ray absorption fine structure (EXAFS) and involves interrogating the unoccupied electronic states by an excited electron from a core level and scattering of the photoelectron from nearby neighboring atoms, respectively. On the other hand, X-ray emission spectroscopy (XES) interrogates the occupied electronic density of states via a de-excited electron back to a core level. Both XAS and XES are manifestly element-specific, as either the excitation or the de-excitation energy, respectively, selects the species of interest. By combining these spectroscopies one can gain a fundamental understanding of the local electronic and atomic structure, elucidating properties of the selected species such as oxidation state, bond lengths, ligand identity, and coordination symmetry and numbers. The development of reliable experimental techniques at DOE funded X-ray facilities has

facilitated the accessibility to high-quality XAS and XES measurements. In addition, theoretical advances in electronic structure theories, that can reliably predict XAS and XES spectra, have also been developed over the last few years. XANES and XES have been recently combined – specifically valence-to-core (VtC) XES – with ML [5] to chemically classify a very wide range of molecular sulfur compounds based on their rich bonding environments. An open access ML toolkit [6] has been developed and run on top of standard open access packages (scikit-learn [7], Keras [8], Tensorflow [9]) that implement various ML classifications on spectra.

Work Plan: We propose to address the Ca-organic matter complexation problem by computing Ca K-edge XANES spectra with time-dependent density functional theory (TDDFT) and combine the calculated spectra with recently developed ML tools, which we will adapt as needed, to identify sensitive spectral fingerprints that can be used to distinguish different Ca-organic complexes from one another, which ultimately enables us to ascertain the organic matter functional groups participate in Ca bridging. The Tahoma computing cluster at EMSL is an ideal platform for our overall workflow. Our approach is as follows:

1. **Structural Data and Curation:** The structural data set will include structures (as crystallographic information files) obtained from the Cambridge Structural Database [10], which is a repository for metal-organic crystal structures. We will compile structures containing Ca-O and Ca-N bonds (since we expect Ca to bind to either O or N atoms in OM), of which there are ~ 2000 in the database. OM is a mixture of biomolecules (derived from litter, root excretion, and microbial biomass) and their degradation products. As such, it contains amino acids, peptides and some polysaccharides that contain N; as well as carboxylic acids, polysaccharides, and phenolic compounds (lignin, tannin) that exhibit different O functionalities. The structures obtained from the Cambridge Structural Database may not fall into a specific class of biomolecule; however, we will curate a list of structures to include appropriate functional groups and to exclude non-representative structures, for instance, those that contain another metal center (including more than one Ca center), crown ethers, nitrates, and nitriles. We will initially select structures with fewer than 100 atoms. This curated list will exhibit Ca with different coordination numbers and will include Ca bound to the major functional groups present in OM, including carboxylates, catechols, phenols, and amines. We will include structures in which Ca is chelated and in which Ca is bound in a monodentate fashion. The Cambridge Structural Database is the definitive database for molecular structures; however, we can augment a curated list with mineral structures, as needed, from several other databases (e.g. the Inorganic Crystal Structure Database).

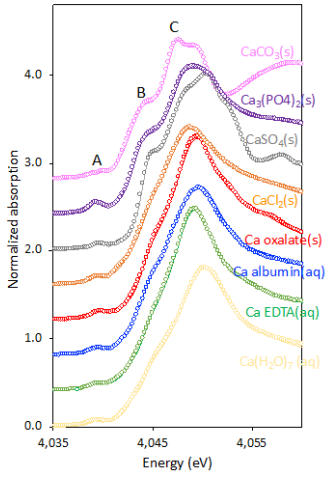

Figure 1: Experimental Ca K-edge XANES spectra for representative reference compounds. The major XANES pre-edge features, A, B and C are shown.

2. **Spectral Data and Generation:** Ca K-edge XANES spectra can exhibit three pre-edge features (A, B, and C) (Figure 1). The lowest energy pre-edge feature (A) is associated with *1s* → *4s*, *4p* transitions, however some *3d* mixing occurs. Previous studies [11] have observed that the intensity of this feature increases in going from 6-coordinate (O_h/D_{4h}) to 8-coordinate (D_{2d}) to 7-coordinate (C_{2v}/C_{3v}) Ca complexes, which was attributed to an increase of O *2p* orbital mixing with the Ca *d* orbitals with decreasing symmetry. The B feature is due to *1s* → *4s*, *4p* transitions. As we can see in Figure 1, there is a large degree of variation in the intensity of this feature among Ca complexes and Ca minerals, which can

be used as a fingerprint to differentiate between different Ca coordination environments. Finally, the C arises from a *1s* → *4p* transition. This suggests that XANES spectroscopy provides insight into the local symmetry of the Ca-organic complex.

In this proposal, in addition to XANES spectra, we will also generate XES spectral data. We will generate the XANES/XES spectra data for all structures in a curated list using electronic structure theory methods – density functional theory (DFT) and time-dependent density functional theory (TDDFT) – with the open-source NWChem computational chemistry program [12], which is developed and maintained at PNNL (https://nwchemgit.github.io/). The curated Ca complexes will first be geometry optimized at the DFT level of the theory. Ground state optimizations will be performed with the B3LYP exchange-correlation functional and suitable all-electron basis sets (6-311G**). This will be followed by TDDFT-based XANES and XES computations of the spectra at the Ca K-edge. For these X-ray spectroscopy calculations, the Sapporo-TZP-2012 all-electron basis set will be used to represent the Ca center, and the 6-311G** basis set for the remaining atoms. The BHLYP exchange-correlation functional will be used for all TDDFT-XANES/XES calculations. We have successfully used this approach to simulate the Ca K-edge XANES in previous studies [13]. The solvent environment, if needed, will be treated with the implicit solvation model. We will also consider explicit solvation if required. The electronic structure methodologies [14,15] in NWChem that we will utilize have been successfully validated and published over a variety of systems. NWChem is available and performs efficiently on the Tahoma cluster at EMSL. A Python-based pipeline [16] has been developed for data generation (Figure 2) that will be adapted, as needed, for the Tahoma cluster. We will augment the computed (synthetic) XANES/XES spectral data with available experimental data from SSRL and other sources.

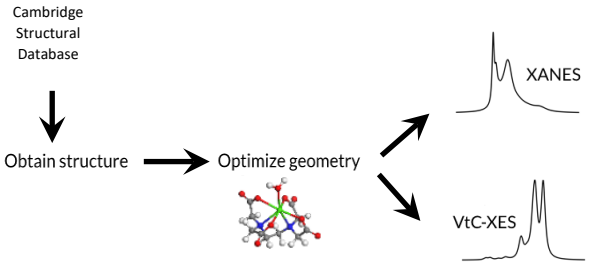

Figure 2: Data generation pipeline schematic

3. **Dimensionality Reduction for XANES/XES:** We have recently discussed dimensionality reduction approaches in the context of X-ray spectroscopies [5]. For completeness, we give a brief overview of the dimensionality reduction approaches we will use. Dimensionality reduction not only helps determine which features in data are most "evident" or variational, but by doing so in a data-driven matter, it also removes potential biases. Lower dimensional representations often yield better classification by addressing the "curse of dimensionality" problem, i.e., everything in a high dimensional space looks far away, so it may be difficult to quantify similarity of points in a high dimensional space [17]. However, selecting the best dimensionality reductional algorithm is closely dependent on both the constraints inherent to the method and the underlying variance of the training data.

   In this proposal, we will use various methods for dimensionality reduction on spectral data to extract spectral similarities and thus determine limits on chemical classes by using these ML-based inferences of structural parameters. Following this recent manuscript [5] , we will investigate three different dimensionality reduction routines: (1) Principal Component Analysis (PCA) [18], which is a fully linear method with an underlying Euclidean metric, (2) a Variational AutoEncoder (VAE) [19], which is a deeply nonlinear method that still has a local metric, and (3) t-distributed Stochastic Neighbor Embedding (t-SNE) [20], a nonlinear embedding that is inherently non-metric. We will explore the utility of these unsupervised machine learning methods to not only analyze the information retained by a reduced-dimensional representation, but most importantly, to identify

classification schemes that are clearly encoded within the spectra. Moreover, both PCA and VAE have the additive benefit of generating a mapping to the reduced-dimensional space which can subsequently be used to map new data onto the derived lower dimensional spaces. This will facilitate applying supervised machine learning to the reduced spaces and thus allow us to quantify the quality of the mapping by calculating the accuracy of supervised classification on a subsequent test set. Furthermore, by progressively decreasing the constraining assumptions of the unsupervised machine learning algorithm, moving from PCA to a VAE to t-SNE, we will not be constrained by the chosen algorithm and thus fully investigate the sensitivity to refined chemical information contained within XANES and XES. For a schematic overview of a VAE architecture, see Figure 3.

4. **Neural Network (NN) for Structural Inferences in Ca-Organic Matter Complexation**

The ability to draw inferences from XANES/XES spectra to structural details is the central problem in advanced X-ray spectroscopies. We would like to ask the following questions. Can we train NNs to identify sensitive spectral fingerprints from the various Ca coordination environments? Do the NNs correctly evaluate the qualitative labels of classification and the quantitative labels of, e.g., bond lengths?

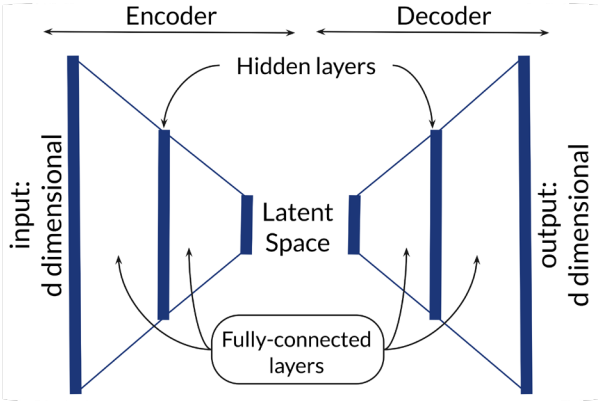

Figure 3: Schematic of the variational autoencoder neural network that will be used to generate reduced dimensional representation of XANES/XES spectra.

We will apply supervised ML to both the spectra and the dimensionally reduced spaces using the classification schemes and regression properties (such as bond length) that were identified via an analysis of the reduced spaces. Thus, we will be predicting structurally relevant information that is manifestly embedded in the Ca XANES/XES spectral data. Each prediction or classification task will have its own ML model, which will be chosen based on the input dimension. For example, classification (such as ligand identity) on the dimensionally reduced spaces will be implemented via K-Nearest Neighbors (KNN) [21]. KNN is both a classification and regression algorithm that categorizes data points based on the other data points in the vicinity, specified by the number of neighbors (k) hyperparameter. However, predictions on the spectra themselves will be implemented via neural networks as neural networks can not only capture the inherent nonlinearities of spectral features, but they also avoid the "curse of dimensionality" that arises when applying KNN on high-dimensional data. Moreover, neural networks can be used for both classification tasks, such as bonding environment identification, and regression tasks, such as bond length predictions. By combining these supervised ML models on the properties identified by the dimensionality reduction routines, we will have an unbiased quantitative evaluation of the sensitivity of both XANES and XES to chemically relevant properties.

Computational Approach.

DFT & TDDFT-XANES Calculations: All DFT and TDDFT-XANES calculations will be performed with the open-source NWChem computational chemistry program developed and maintained at PNNL (https://nwchemgit.github.io/) and installed on the Tahoma computing cluster. We will employ DFT calculations to geometry optimize structures for Ca bound to ligands in the curated structures list. Ground state optimizations will be performed with the B3LYP exchange-correlation functional and suitable all-electron basis sets (6-311G**). With the optimized clusters in hand, TDDFT-XANES calculations will be performed at the Ca K-edge using the restricted excitation window TDDFT approach as implemented in NWChem. For the molecular complexes, the Sapporo-TZP-2012 all-electron basis set will be used to represent the absorbing Ca center, and the 6-311G** basis set for the remaining atoms. For the crystalline systems, the Sapporo-TZP-2012 all-electron basis set will be used for the central Ca absorbing center, whereas the remaining Ca centers will be represented with the Stuttgart RLC ECPs (relativistic large-core effective core potentials), and the remaining atoms will be represented with the 6-311G** basis set. The BHLYP exchange-correlation functional will be used for all TDDFT-XANES calculations. These ground and excited state simulations comprise ~20-100 atoms across all the systems (~2000) that will be considered. We will utilize between 2-4 nodes for these calculations for a maximum of 48 hours. We estimate we will need ~35,000 node hours for DFT and TDDFT-XANES calculations.

Dimensionality Reduction and Neural Network Models: All machine learning models and statistical analysis will be implemented in python using the scikit-learn [7], Keras [8], and Tensorflow [9] packages. During training, deep neural networks implemented in Keras/Tensorflow can easily run on a single GPU to speed up computational time, or they can be implemented using distributed training such that the networks utilize parallelization on multiple GPUs, also increasing computational efficiency. This attribute will be particularly beneficial as we expect to train multiple neural networks – one for each of the various properties identified through dimensionality reduction routines. All other machine learning routines – such as PCA, t-SNE, and KNN – will be implemented using scikit-learn, which has no GPU support. We will utilize 1 node for these calculations (utilizing both CPU and GPGPUs) (10 jobs) for a maximum of 48 hours. We estimate we will need ~1,000 node hours

Computing Resources

IMPORTANT: The EMSL computing systems available to users are not approved for use with sensitive data. The processing, storage, or transmittal of sensitive data (e.g. Personally Identifiable Information, Official Use Only, etc.) is thus prohibited on Tahoma, Cascade and Aurora. Due diligence must be used to prevent inadvertent disclosure of invention, patent, or other sensitive information. It is your responsibility to protect access to the information.

☒ By checking this box, I am confirming that participants on this proposal will NOT process, store, or transmit sensitive data (e.g. Personally Identifiable Information, Official Use Only, etc.) on Tahoma, Cascade or Aurora.

Total CPU Hours Request for first year of proposal: 35,000 node hours					
Total GPGPU Hours Request for first year of proposal: 1,000 node hours					
Total Data Archive Request for first year of proposal:					
Software Details	**Node Request (CPUs or GPGPUs)**	**Estimated # of jobs**	**Estimated Node Hours**	**Expertise of your investigators for these requests**	**EMSL Support Requested** *Specific Needs (e.g., compiling code, libraries needed, help running jobs, etc.)*
NWChem	2-4 nodes	2000	35,000	Expert User	Compiling code, libraries, model building
Machine Learning Python Tools	1 node (CPU/GPGPUs)	10	1,000	Expert User	Installing libraries

Notes:

Tahoma allocations are awarded in units of wall-clock time expressed in node-hours. Tahoma's 160 CPU nodes each have 36 (3.1 GHz) Intel Xeon processor cores with 384 GB of memory and 2 TB of flash storage. Consequently, 10,000 Tahoma CPU node-hours are equal to 360,000 processor core-hours. Tahoma's 24 GPGPU nodes each have 36 processor cores and 2 Nvidia v100 GPGPUs, 1536 GB of memory, and 7 TB of flash storage. Tahoma's 10 PB global file system is capable of 100 Gigabyte/sec bandwidth. Tahoma can deliver a total of 1,500,000 node-hours per year.

Upon successful review and approval of a proposal, computing resources will be allocated for analysis and archiving of experimental data generated at EMSL.

Appendix 1: References

1. C. Rasmussen, et al, *Biogeochemistry*, *137*(3), pp.297-306.
2. R. Mikutta, et al, *Geochimica et Cosmochimica Acta*, *71*(10), pp.2569-2590.
3. J.Timoshenko, et al, *ACS Catalysis*, *9*(11), 10192 (2019)
4. C.Zheng,et al, *Patterns*, *1*(2),100013 (2020)
5. S. Tetef, et al, *PCCP (in review, 2021),* chemrxiv.org/engage/chemrxiv/article-details/60cbb47ffca4902445c7ac4b
6. github.com/Seidler-Lab/Sulfur-ML/tree/v1.0.0
7. G. V. Fabian Pedregosa, et al, *Journal of Machine Learning Research*, 12, 2825 (2011)
8. F. a. o. Chollet, et al, *keras.io* (2015)
9. A. A. Martín Abadi, et al, *tensorflow.org* (2015)
10. C. R. Groom,et al, *Acta Cryst*, B72, 171-179 (2016)
11. V. Martin-Diaconescu, et al, *Inorganic Chemistry 54* (4), 1283-1292 (2015)
12. E. Aprà, et al, *J. Chem Phys*, *152*, 184102 (2020), github.com/nwchemgit/nwchem
13. K. Henzler, et al, *Science Advances*, *4*(1), p.eaao6283 (2018)
14. K. Lopata, et al, *J. Chem. Theory Comput. 8*, 3284 (2012)
15. Y. Zhang, et al, *J. Chem. Theory Comput, 11*(12),5804 (2015)
16. github.com/Seidler-Lab/nwxpl
17. P. Indyk, et al, *ACM Press*, 604 (1998)
18. S. Wold, et al, *Chemometrics and Intelligent Laboratory Systems*, 2, 37 (1987)
19. A. Rocchetto, et al, *npj Quantum Information*, 4, 28 (2018)
20. L. van der Maaten, et al, *Journal of Machine Learning Research*, 9, 2579 (2008)
21. J. Goldberger, et al, *Advances in Neural Information Processing Systems*, 17, 513 (2005)

Appendix 3: Active Collaborator List

CONFLICT OF INTEREST LIST				
Name	**Key Co-Author**	**Collaborator**	**Advisee / Advisor**	**Other**