Machine Learning in Complex Organic Mixtures: Applying Domain Knowledge Allows for Meaningful Performance with Small Datasets

Katelyn Le,[a‡] Jagoš R. Radović,[b‡] Justin L. MacCallum,[a] Stephen R. Larter,[c] Jeffrey F. Van Humbeck[a]*

[a] Department of Chemistry, University of Calgary, 2500 University Drive NW, Calgary, Alberta, Canada, T2N 1N4.
[b] Center for Petroleum Geochemistry (CPG), Department of Earth and Atmospheric Sciences, University of Houston, Houston, Texas 77204-5007, United States
[c] Department of Earth, Energy, and Environment, University of Calgary, 2500 University Drive NW, Calgary, Alberta, Canada T2N 1N4

*Corresponding author email: jeffrey.vanhumbec1@ucalgary.ca

‡These authors contributed equally.

Abstract

The ability to quantify individual components of complex mixtures is a challenge found throughout the life and physical sciences. An improved capacity to generate large datasets along with the uptake of machine-learning (ML) based analysis tools has allowed for various 'omics' disciplines to realize exceptional advances. Other areas of chemistry that deal with complex mixtures often cannot leverage these advances. Environmental samples, for example, can be more difficult to access and the resulting small datasets are less appropriate for unconstrained ML approaches. Herein, we present an approach to address this latter issue. Using a very small environmental dataset—35 high-resolution mass spectra gathered from various solvent extractions of Canadian petroleum fractions—we show that the application of specific domain knowledge can lead to ML models with notable performance.

Complex mixtures are found throughout multiple areas of chemistry. The ability to relate the observed properties of such mixtures to their molecular composition is a shared goal across fields. Many exceptional advances—particularly in the various '-omics' disciplines—have relied in large part on two specific developments. High-throughput experimentation using sensitive analytical tools has increased the total number of experiments that can be performed and the quantity of data available from each observation.[1] Second, machine learning strategies have made these vast quantities of data processable.[2,3] However, the specific tools that enable this workflow in biomedicine do not necessarily apply to all complex mixtures. In situations with less accessible data out-of-the-box ML tools may be inappropriate, and the thoughtful application of domain knowledge to support data analysis may be critical. The use of physical models to guide the development of ML workflows is termed 'scientific machine learning'[4] and has been shown in multiple contexts to improve performance.[5] We were interested to see whether domain knowledge could be used to help predict the behavior of complex mixtures with limited training data.

Petroleum samples are a complex mixture of significant economic relevance. Recent work to make the analysis of petroleum more molecular in its focus has been aptly termed 'petroleomics'.[6] Within this field, the chemistry of asphaltenes is of particular interest due to their adverse impact during petroleum processing and distribution.[7,8]

Asphaltenes are defined as the fraction of crude petroleum that is soluble in toluene and insoluble in *n*-alkanes[9,10] and are problematic in industry due to their tendency to form deposits during storage and transport.[11,12] On a molecular level, asphaltenes are an ultracomplex mixture whose individual components may include polyaromatic motifs with possible alkyl branching and bridging; acidic and basic functional groups; heteroatoms (e.g. O, N, S); and chelated trace metals.[13] A number of research groups have attempted to increase practical knowledge with the bottom-up synthesis of compounds intended to model various archetypical asphaltenes.[14–17] A further difficulty is presented by the diverse distribution of specific species within asphaltenes, which vary not only with different geographic sources but even between samples within a region.[18,19] Two asphaltenes sourced from Alberta, Canada were investigated herein (Figure 1). Having 4,428 and 4,438 identified molecular formulae, respectively, these samples unquestionably represent complex mixtures. The similarity of those numbers also underestimates their relative difference from one another. Together, the two samples contain 6,458 unique formulae—meaning that these mixtures have roughly equal amounts of shared and unique ions.

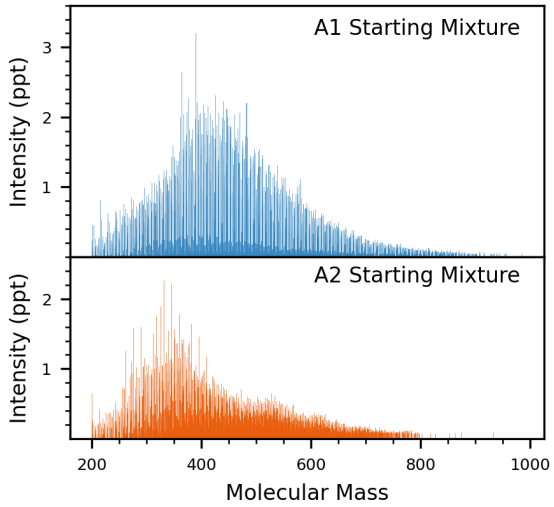


Figure 1. Mass spectrometry data for two asphaltene mixtures of Western Canadian origin.

Recent machine learning applications towards asphaltenes have typically attempted to predict bulk properties (*e.g.* precipitation/stability) from bulk features (*e.g.* 'SARA' saturates/aromatics/resins/asphaltenes fraction amounts).[20–23] We wanted a relevant process that could be performed on these mixtures that would produce a *quantifiable* and *molecular* output[24]—where we would aim to develop a machine learning approach that could predict such an output. The specific character of asphaltenes presents an additional challenge for obtaining data of this nature. Asphaltenes form aggregates, even at low concentrations and in solvents where they are functionally soluble.[25–28] During aggregate assembly, species may become trapped within or adsorbed to newly formed macrostructures which effectively occludes them from analysis. To obtain a true representation of the molecular profile of these mixtures, it is often essential to apply a pretreatment. Fractionation with various amounts of precipitating and solubilizing solvent is one method that is well-studied and relatively efficient.[29] Previous experiments have quantified the effects of extraction through NMR, elemental analysis, and high-resolution mass spectrometry.[11,30–35] We felt that the particular nature of HRMS petroleomics datasets, i.e. the presence of series of repeatable homologous units, made it an ideal technique

matched to our overall project goals, resulting in an ML approach potentially applicable to other HRMS datasets with comparable properties, e.g. from the analysis of complex natural organic matter samples.

Thus, we constructed a dataset as shown in Figure 2. From **A1** and **A2**, we performed 35 solvent extractions to generate fractions **E1**-**E35**. Each extraction used toluene or CH_2Cl_2 as the non-polar component along with one of four polar modifiers (*i.e.* isopropanol, acetone, acetic acid, and triethylamine). While the starting mixtures and extracted fractions certainly share some similarities in terms of their quasi-Gaussian distribution, there was a measurable change observed that became our target for ML prediction. As mentioned, solubilizing solvents can release occluded molecules in asphaltene samples.[29] We observed this in our own data. Where our starting mixture featured 6,458 identified ions, the set of 35 extractions resulted in over 11,000— adding an additional challenge for prediction.

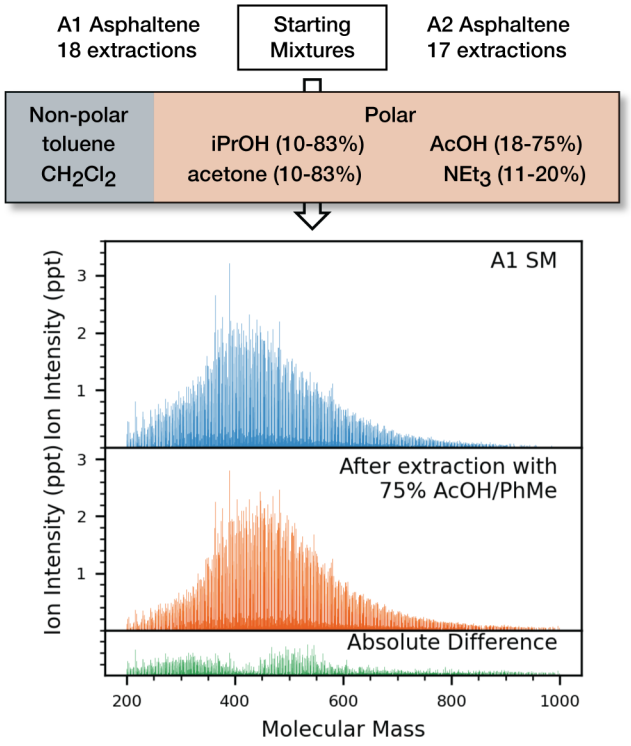
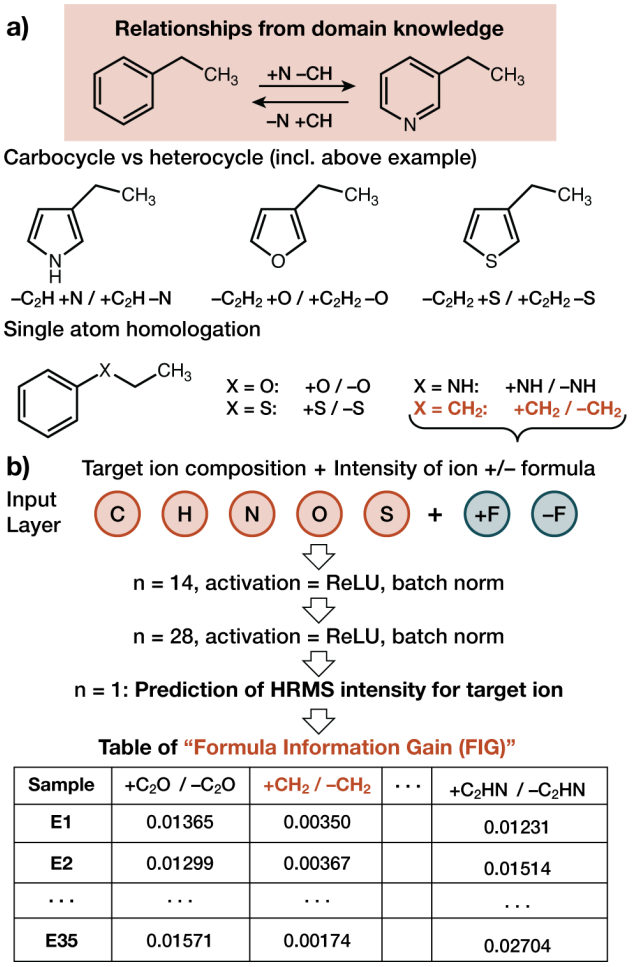


Figure 2. Representative example for one of 35 solvent extractions performed to create our dataset. Plots for all 34 other extractions are available in the Supporting Information.

We suspected that unconstrained ML would be prone to find spurious correlations in this dataset. Our approach to enforcing reasonable limitations during analysis is described in Figure 3. It focused, broadly speaking, on two places where the simple molecular mass of two compounds might not describe their relative solubility. For example, an alkylbenzene and alkylpyridine could be very similar in terms of molecular mass but show divergent behavior when acetic acid is part of the extraction solvent. Conversely, an alkylbenzene and phenyl sulfide with similar side-chain length could perhaps show greater similarity than expected from their difference in mass. In total we selected 23 molecular relationships for investigation, nine are shown in Fig. 3a with the balance in the Supporting Information (Figure SI-7). While a reasonable argument could be made for either a larger or smaller number than this, we believe the performance of our ML method *vide infra* justifies this choice. Further, we felt our selection of a particular set of

molecular relationships could be supported if it were able to pass a positive control test. The generation of data for this test is described in Figure 3b. We were broadly inspired by masked language models (*e.g.* BERT), where a model is trained to predicted a hidden ("masked") word from a body of text, with the necessary context to make that prediction supplied by the surrounding words.[36] For each of our extracted samples E1-E35 we trained a very lightweight fully connected network to predict a masked ion intensity, based only on the observed intensity of the ions found at +/– one of our 23 different formula relationships. We averaged the mean-squared test error for 5 individual trials to create a table of 'Formula Information Gain' (FIG), a description of how much predictive information was held in each molecular relationship for each extracted sample.



Sample	+C_2O / –C_2O	+CH_2 / –CH_2	···	+C_2HN / –C_2HN
E1	0.01365	0.00350		0.01231
E2	0.01299	0.00367		0.01514
···	···	···		···
E35	0.01571	0.00174		0.02704

Figure 3. (a) A representative selection of molecular relationships (of 23 total) that were used in data analysis. (b) A 'masked ion' strategy for rapidly evaluating these 23 relationships in terms of their predictive power.

In the petroleomics field, samples are often quantitatively described by the abundance of heteroatom compound classes (CCs).[37] Our extracted samples contained 22 different CC, ranging from pure hydrocarbons to formulae with 6 heteroatoms (*i.e.* N2S4). Thus, we had a domain-relevant standard for clustering analysis. Figure 4 shows the result of UMAP clustering for all 35 extracted samples, using both CC and our FIG analysis. In both cases, samples derived from **A1** and **A2** are strongly separated: all clusters using CC are entirely homogeneous, and only 2 of 35 samples are in a cluster as a minority species using FIG. Five of the solvent blends used (*i.e.* all four using toluene, and

CH_2Cl_2/acetone) show very strong qualitative similarity between CC and FIG. The two remaining solvent blends (*i.e.* CH_2Cl_2 with isopropanol or acetic acid) each had only 1 sample derived from **A1**, and FIG tends to group those two extractions with **A2** samples. Clustering with other UMAP parameter settings can be found in the Supporting Information (Figures **SI-8** and **SI-9**).

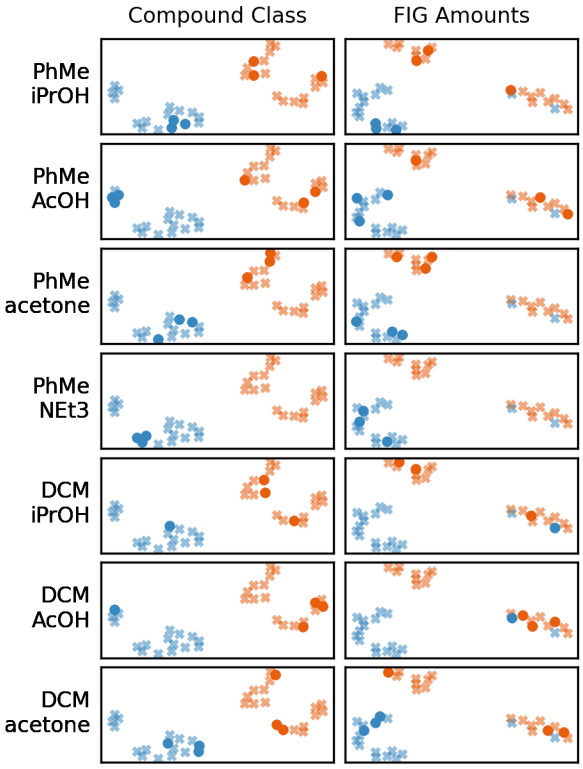


Figure 4. UMAP clustering results using CC and FIG. Blue symbols are samples produced from A1; orange symbols from A2. Circular symbols show extractions using the solvent blend listed in the row label.

The point of this analysis was not to identify whether clustering from CC versus FIG should be preferred. As this analysis required training over 4,000 individual networks (*i.e.* 23 +/– formulae; 35 samples; average of 5 trials) we used absolutely minimal networks that had 518 learnable parameters to reduce computational cost. Our optimized network for predicting extraction results for comparison has 3.2 million parameters (*vide infra*). What was abundantly clear from even this minimal network was the fact that FIG contained real information about the structure of the data in the extracted fractions, just as CC do. The distribution of extracted samples in the FIG clustering is very obviously not a random assortment. With some confidence that FIG could be a useful way to restrict ML methods to focus on meaningful relationships, our approach to quantitatively predict HRMS spectra of extracted fractions using molecular relationships is described in Figure 5.

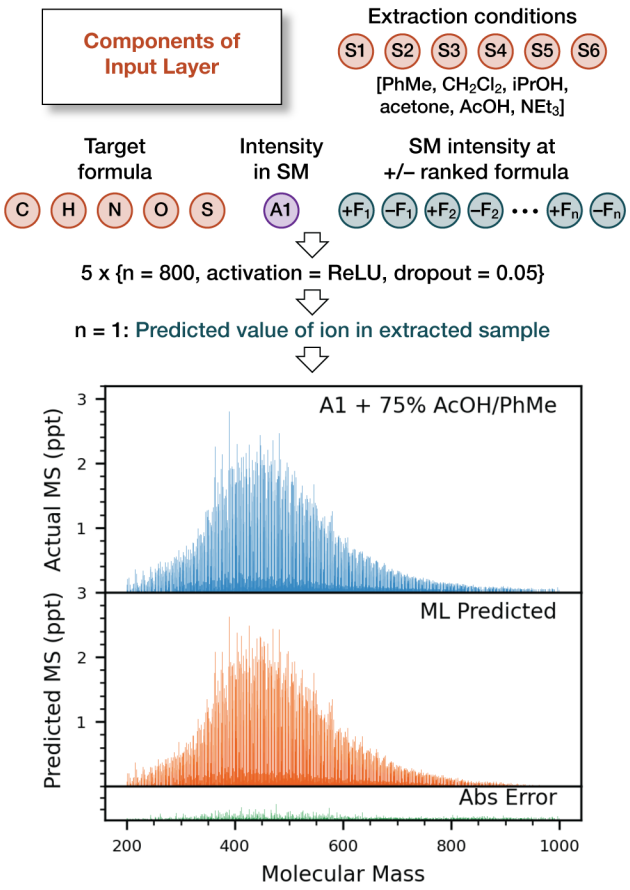


Figure 5. Optimized ML architecture for predicting the composition of an extracted sample. A representative sample (**E35**) analyzed with 11 additional molecular relationships is shown.

The inputs for this model included: (i) the CHNOS formula of the target ion; (ii) the intensity of that ion in the starting mixture; (iii) the solvent blend used for extraction. The creation of our FIG table allowed us to have a starting 'rank' for how informative we might expect a given molecular relationship to be. During hyperparameter optimization (number of layers, width per layer, learning rate, dropout amount, batch normalization) we used from zero to five molecular relationships, including each new relationship in rank order. For example, during the generation of the FIG table, we found +/– CH_2 was the most informative relationship—as might be expected given this homologous series is well known in petroleum chemistry[38]—and so it was included in any optimization run that involved one or more relationships. This optimization produced the architecture shown (Fig. 5). To fully evaluate the effectiveness of our approach, we performed leave-one-out cross-validation (LOO-CV) with increasing numbers of molecular relationships from zero to fifteen included, keeping all other network parameters identical. A prediction that resulted in a modest level of error (*i.e.* a rank of 13th, where 1st represents the most error and 35th the least) is shown in Figure 5. Immediately, it can be seen that our approach successfully predicts most of the change that occurred during solvent extraction (*c.f.* Fig 2. vs. Fig. 5).

A minor challenge in evaluating this approach was in identifying an appropriate benchmark. To the best of our knowledge, the quantitative prediction of HRMS ion intensity for a petroleum data set comparable to ours has not

been attempted. We reasoned that the following measures provide meaningful context, and their quantitative performance is shown in Figure 6. First, asphaltene samples have by definition *already been classified according to their solubility*. It may be the case that the distribution of compounds will be relatively unchanged by further extraction, so one would expect the starting spectrum to persist unchanged (*i.e.* Fig 6. 'Persist'). The bimodal distribution seen here requires explanation. During LOO-CV, we considered every ion observed in all training set samples to make up the 'total ion set' that needed to be predicted. The size of the total ion set ranged from 11,163 to 11,252 formulae. However, individual extraction samples themselves only show between 2,833 to 6,722 ions. So, from 40.0–74.8% of ion intensities are zero depending on the sample. When an ion was absent in the starting mixture for a given extraction zero was predicted, and this means the majority of the 'Persist' distribution has perfect accuracy. The upper part of the bimodal distribution, represents a more appropriate benchmark for an actual prediction of a non-zero intensity ion.

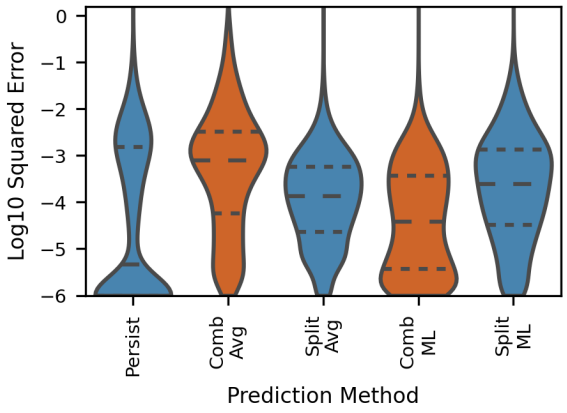


Figure 6. Results for optimized ML methods and benchmark methods for predicting extracted sample HRMS data. Orange plots represent cases where all samples were analyzed together; blue plots represent cases where samples were split based on which starting mixture was used.

Alternatively, one could presume that the solubility behaviors of individual components of A1 and A2 could be relatively insensitive to the *specific* nature of the solvent applied. The least soluble components would be universally precipitated, with others universally soluble. If that were the case, for any given ion a good prediction could be made by taking the average value of that ion in every other sample. This 'Combined Average' (Fig. 6 'Comb Avg') was worse that assuming persistence in the samples. A1 and A2 are quite different, as judged from Fig. 1, so rather than averaging all extractions to make a prediction, we split our samples into those derived from A1 and A2 (Fig. 6 'Split Avg'). Significantly more accurate predictions result. For our ML approach, one could also envision using samples extracted from both A1 and A2 in the training data, or only those from the same starting mix as the held-out test sample. We did exactly this (Fig. 6 'Comb ML' and 'Split ML') and were very surprised to see the *opposite* behavior. LOO-CV predictions were significantly more accurate if training data from both A1 and A2 were included, despite the significant differences in these starting mixtures.

Complete results from LOO-CV are shown in Figure 7. Total error is minimized when 11 formula relationships are included, and this was the number used to generate the data presented in Figure 6. Broadly speaking, we see three

different types of specific behavior. Several of samples are very accurately predicted, even without the inclusion of additional relationships (*e.g.* E6, 10). Three samples are challenging to predict, regardless of the amount of information provided (*i.e.* E2, 12, 28). Most interesting is the behavior of many samples that are difficult to predict with no additional information, but show a dramatic improvement in performance as more relationships are included (*e.g.* E11, 19, 27, 29, 32).

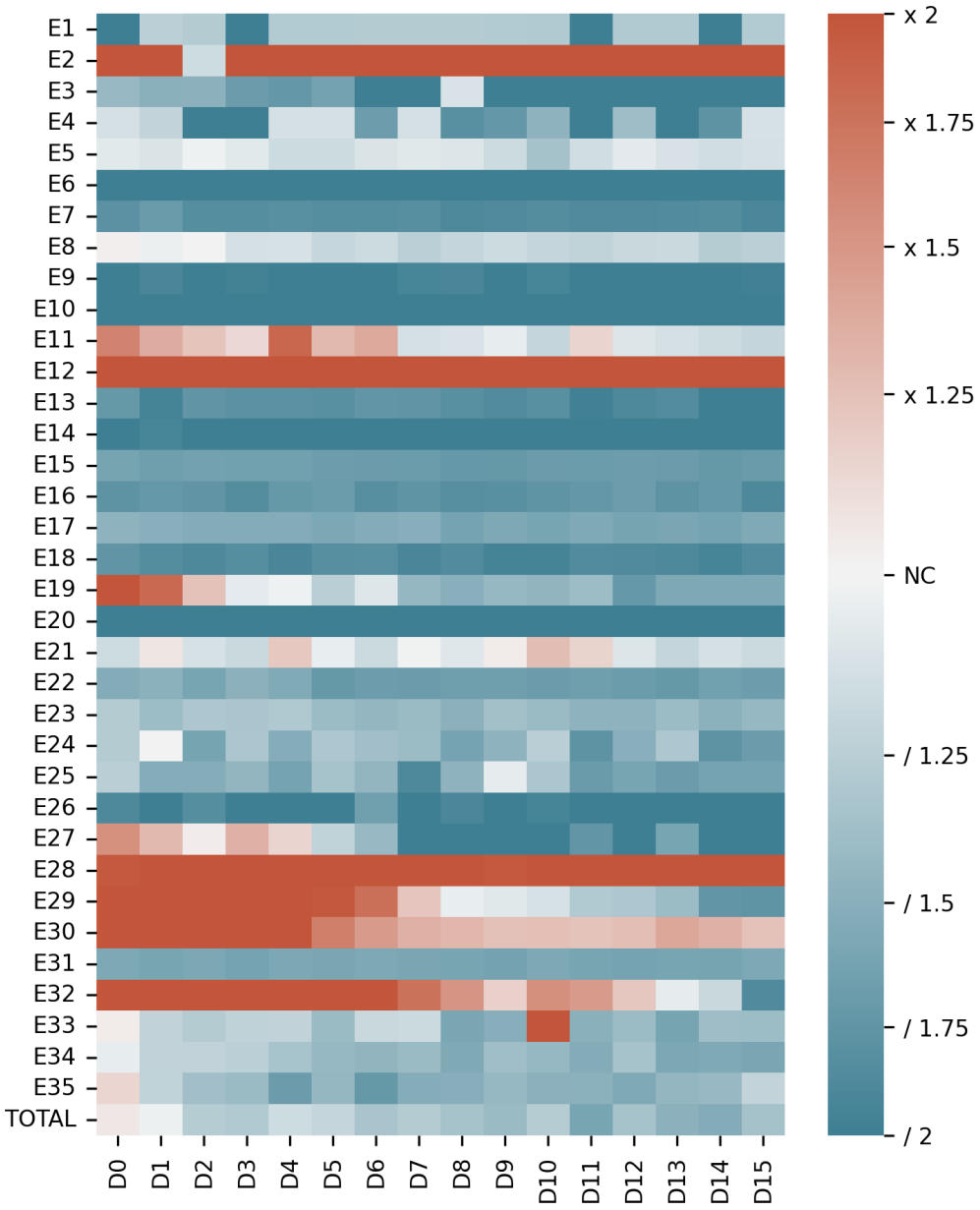


Figure 7. Performance of the 'Combined ML' method during LOO-CV for each sample (E1-E35) with between zero and fifteen additional context points provided (D0-D15). Blue shades represent ML predictions with less error than the split average (deepest shade is halved); orange represents more error than the split average (deepest shade is doubled).

In summary, we have demonstrated that a dataset containing only 35 experimental examples of very complex organic mixtures is amenable to both supervised and unsupervised machine learning approaches. Our HRMS-based petroleomics approach generated quantitative and molecular information for those samples, which could potentially be a double-edged sword. When applying a per-ion prediction strategy, our dataset now contains over 350,000 individual ions. However, an unrestricted search for relationships between ions would now need to discern—for each of these predictions—which of the thousands of other observed ions hold important information for predicting the intensity of an extracted ion. By applying domain expertise, we were able to restrict a ML approach to focus on a small number of relationships that had a good chance of holding important information: useful levels of accuracy were the result. We suspect that incorporating domain knowledge in this or other similar ways will be beneficial to ML analysis of complex mixtures in domains across chemistry that have limited data. Beyond petroleum, we think that other complex HRMS datasets that have underlying ordered structure from well-defined chemical relationships and homologous series (*e.g.* oceanic dissolved organic matter[39]) will benefit from this approach.

Supporting Information

General methods, materials, details and all performed extraction conditions. Additional figures that show the absolute difference after extraction (Fig. 2) or ML performance (Fig. 5) for all extractions (PDF). All python files necessary to perform the analysis described in this paper, as well as a selection of trained models and the raw HRMS data can be found on GitHub: github.com/JVH-YYC/Bitumen-ML.

Conflicts of Interest

There are no conflicts to declare.

Acknowledgements

This work was supported by Alberta Innovates through the Carbon Fiber Grand Challenge (Phase I) and by allowing access to the Asphaltene Sample Bank. It was also supported by the Canada First Research Excellence Fund – Global Research Initiative. J.L.M. is supported by the Canada Research Chairs program.

References

(1) Pade, L. R.; Stepler, K. E.; Portero, E. P.; DeLaney, K.; Nemes, P. *Mass. Spectrom. Rev.* **2024**, *43*, 106.

(2) Liebal, U. W.; Phan, A. N. T.; Sudhakar, M.; Raman, K.; Blank, L. M. *Metabolites* **2020**, *10*, 243.

(3) Reel, P. S.; Reel, S.; Pearson, E.; Trucco, E.; Jefferson, E. *Biotechnol. Adv.* **2021**, *49*, 107739.

(4) Baker, N.; Alexander, F.; Bremer, T.; Hagberg, A.; Kevrekidis, Y.; Najm, H.; Parashar, M.; Patra, A.; Sethian, J.; Wild, S.; Willcox, K.; Lee, S. *Workshop Report on Basic Research Needs for Scientific Machine Learning: Core Technologies for Artificial Intelligence*; United States, **2019**. doi: 10.2172/1478744.

(5) Rackauckas, C.; Ma, Y.; Martensen, J.; Warner, C.; Zubov, K.; Supekar, R.; Skinner, D.; Ramadhan, A.; Edelman, *arXiv preprint* arXiv:2001.04385 **2020**.

(6) Marshall, A. G.; Rodgers, R. P. *Proc. Natl. Acad. Sci. USA* **2008**, *105*, 18090.

(7) Akbarzadeh, K.; Hammami, A.; Kharrat, A.; Zhang, D.; Allenson, S.; Creek, J.; Kabir, S.; Jamaluddin, A.; Marshall, A. G.; Rodgers, R. P.; Mullins, O. C.; Solbakken, T. *Oilfield Review* **2007**, *19*, 22.

(8) Buenrostro-Gonzalez, E.; Groenzin, H.; Lira-Galeana, C.; Mullins, O. C. *Energy Fuels* **2001**, *15*, 972.

(9) Speight, J. G. *Oil Gas Sci. Technol.* **2004**, *59*, 467.

(10) Adams, J. J. *Energy Fuels* **2014**, *28*, 2831.

(11) Buckley, J. S.; Hirasaki, G. J.; Liu, Y.; Von Drasek, S.; Wang, J. X.; Gill, B. S. *Pet. Sci. Technol.* **1998**, *16*, 251.

(12) Klein, G. C.; Kim, S.; Rodgers, R. P.; Marshall, A. G.; Yen, A.; Asomaning, S. *Energy Fuels* **2006**, *20*, 1965.

(13) Chacón-Patiño, M. L.; Rowland, S. M.; Rodgers, R. P. *Energy Fuels* **2018**, *32*, 9106.

(14) Morimoto, M.; Fukatsu, N.; Tanaka, R.; Takanohashi, T.; Kumagai, H.; Morita, T.; Tykwinski, R. R.; Scott, D. E.; Stryker, J. M.; Gray, M. R.; Sato, T.; Yamamoto, H. *Energy Fuels* **2018**, *32*, 11296.

(15) Akbarzadeh, K.; Bressler, D. C.; Wang, J.; Gawrys, K. L.; Gray, M. R.; Kilpatrick, P. K.; Yarranton, H. W. *Energy Fuels* **2005**, *19*, 1268.

(16) Sjöblom, J.; Simon, S.; Xu, Z. *Adv. Colloid. Interface. Sci.* **2015**, *218*, 1.

(17) Tan, X.; Fenniri, H.; Gray, M. R. *Energy Fuels* **2008**, *22*, 715.

(18) Klein, G. C.; Kim, S.; Rodgers, R. P.; Marshall, A. G.; Yen, A. *Energy Fuels* **2006**, *20*, 1973.

(19) Zheng, F.; Shi, Q.; Vallverdu, G. S.; Giusti, P.; Bouyssiere, B. *Processes* **2020**, *8*, 1.

(20) Gao, X.; Dong, P.; Meng, X.; Tian, D.; Wang, X. *SPE Journal* **2023**, *28*, 2065.

(21) Ali, S. I.; Lalji, S. M.; Awan, Z.; Qasim, M.; Alshahrani, T.; Khan, F.; Ullah, S.; Ashraf, A. *Chemom. Intell. Lab. Syst.* **2023**, *235*, 104784.

(22) Asoodeh, M.; Gholami, A.; Bagheripour, P. *Fluid Phase Equilib.* **2014**, *364*, 67.

(23) Moncayo-Riascos, I.; Guerrero-Benavides, C.; Giraldo, J.; Ramírez-Jaramillo, Ó.; Rojas-Ruiz, F. A.; Orrego-Ruiz, J. A.; Cundar, C.; Cañas-Marín, W. A.; Osorio Gallego, R. *Energy Fuels* **2022**, *36*, 14243.

(24) Oldenburg, T. B. P.; Brown, M.; Bennett, B.; Larter, S. R. *Org. Geochem.* **2014**, *75*, 151.

(25) Strausz, O. P.; Peng, P.; Murgich, J. *Energy Fuels* **2002**, *16*, 809.

(26) Porte, G.; Zhou, H.; Lazzeri, V. *Langmuir* **2003**, *19*, 40.

(27) Roux, J. N.; Broseta, D.; Demé, B. *Langmuir* **2001**, *17*, 5085.

(28) Gray, M. R.; Tykwinski, R. R.; Stryker, J. M.; Tan, X. *Energy Fuels* **2011**, *25*, 3125.

(29) Chacón-Patiño, M. L.; Rowland, S. M.; Rodgers, R. P. *Energy Fuels* **2018**, *32*, 314.

(30) Liao, Z.; Zhou, H.; Graciaa, A.; Chrostowska, A.; Creux, P.; Geng, A. *Energy Fuels* **2005**, *19*, 180.

(31) Chacón-Patiño, M. L.; Vesga-Martínez, S. J.; Blanco-Tirado, C.; Orrego-Ruiz, J. A.; Gómez-Escudero, A.; Combariza, M. Y. *Energy Fuels* **2016**, *30*, 4550.

(32) Neumann, A.; Chacón-Patiño, M. L.; Rodgers, R. P.; Rüger, C. P.; Zimmermann, R. *Energy Fuels* **2021**, *35*, 3808.

(33) Strausz, O. P.; Torres, M.; Lown, E. M.; Safarik, I.; Murgich, J. *Energy Fuels* **2006**, *20*, 2013.

(34) Novaki, L. P.; Keppeler, N.; Kwon, M. M. N.; Paulucci, L. T.; De Oliveira, M. C. K.; Meireles, F. A.; Baader, W. J.; El Seoud, O. A. *Energy Fuels* **2019**, *33*, 58.

(35) Gawrys, K. L.; Spiecker, P. M.; Kilpatrick, P. K. *Pet. Sci. Technol.* **2003**, *21*, 461.

(36) Devlin, J.; Chang, M.-W.; Lee, K.; Toutanova, K. BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding. **2019**, arXivID: 1810.04805.

(37) Marshall, A. G.; Rodgers, R. P. *Acc Chem Res* **2004**, *37*, 53.

(38) Stratiev, D.; Shishkova, I.; Tankov, I.; Pavlova, A. *J. Pet. Sci. Eng.* **2019**, *178*, 227.

(39) Longnecker, K.; Kujawinski, E. B. *Rapid Commun. Mass. Spectrom.* **2016**, *30*, 2388.