Norwegian University of Life Sciences
Faculty of Science and Technology
Department of Data Science

Philosophiae Doctor (PhD)
Thesis 2023:35

New understanding of gas hydrate phenomena and natural inhibitors in crude oil systems through mass spectrometry and machine learning

Ny forståelse av gasshydratfenomener og naturlige inhibitorer i råoljesystemer gjennom massespektrometri og maskinlæring

Elise Lunde Gjelsvik

New understanding of gas hydrate phenomena and natural inhibitors in crude oil systems through mass spectrometry and machine learning

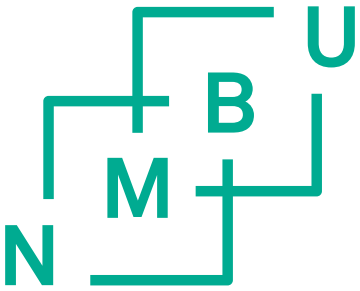
Ny forståelse av gasshydratfenomener og naturlige inhibitorer i råoljesystemer gjennom massespektrometri og maskinlæring

Philosophiae Doctor (PhD) Thesis

Elise Lunde Gjelsvik

Norwegian University of Life Sciences
Faculty of Science and Technology
Department of Data Science

Ås (2023)



Thesis number 2023:35
ISSN 1894-6402
ISBN 978-82-575-2063-2

Acknowledgements

The research presented in this thesis was conducted from March 2020 to February 2023 at the Faculty of Science and Technology, Norwegian University of Life Sciences (NMBU). The work was in collaboration with SINTEF and funded by the Research Council of Norway (Project number: 294636) and industrial partners Equinor ASA, OMV (Norge) AS, Wintershall DEA Norge and TotalEnergies. The research was supervised by main supervisor Kristin Tøndel and co-supervisor Martin Fossen.

Firstly I want to express my gratitude towards my main supervisor Kristin for guidance through this PhD project. Thank you for always answering all my questions, the valuable feedback throughout, and the support. I would also like to thank my co-supervisor Martin for all the help with understanding oil chemistry and performing all the experiments. Thank you both for sharing so much of your knowledge with me in these last years, for believing in me and supporting me through this journey. Also thank you for proof reading my thesis.

Thank you to Anders Brunsvik at SINTEF for conducting all the FT-ICR MS measurements, sharing your knowledge and helping me analyse the spectra and interpret peaks. Also thank you to Anders and SINTEF Industry in Trondheim for having me visit, and letting me borrow an office during all my trips to Trondheim the past three years.

Thanks to all the project partners for valuable discussions, input and sharing your immense knowledge of gas hydrate systems and chemistry.

A big thanks to family and friends for support and distractions from the PhD bubble whenever needed.

Elise Lunde Gjelsvik

Ås, February 2023

Abstract

Gas hydrates represent one of the main flow assurance issues in the oil and gas industry as they can cause complete blockage of pipelines and process equipment, forcing shut downs. Previous studies have shown that some crude oils form hydrates that do not agglomerate or deposit, but remain as transportable dispersions. This is commonly believed to be due to naturally occurring components present in the crude oil, however, despite decades of research, their exact structures have not yet been determined. Some studies have suggested that these components are present in the acid fractions of the oils or are related to the asphaltene content of the oils. Crude oils are among the worlds most complex organic mixtures and can contain up to 100 000 different constituents, making them difficult to characterise using traditional mass spectrometers. The high mass accuracy of Fourier Transform Ion Cyclotron Resonance Mass Spectrometry (FT-ICR MS) yields a resolution greater than traditional techniques, making FT-ICR MS able to characterise crude oils to a greater extent, and possibly identify hydrate active components.

FT-ICR MS spectra usually contain tens of thousands of peaks, and data treatment methods able to find underlying relationships in big data sets are required. Machine learning and multivariate statistics include many methods suitable for big data. A literature review identified a number of promising methods, and the current status for the use of machine learning for analysis of gas hydrates and FT-ICR MS data was analysed. The literature study revealed that although many studies have used machine learning to predict thermodynamic properties of gas hydrates, very little work have been done in analysing gas hydrate related samples measured by FT-ICR MS.

In order to aid their identification, a successive accumulation procedure for increasing the concentrations of hydrate active components was developed by SINTEF. Comparison of the mass spectra from spiked and unspiked samples revealed some peaks that increased in intensity over the spiking levels. Several classification meth-

ods were used in combination with variable selection, and peaks related to hydrate formation were identified. The corresponding molecular formulas were determined, and the peaks were assumed to be related to asphaltenes, naphthenes and polyethylene glycol. To aid the characterisation of the oils, infrared spectroscopy (both Fourier Transform infrared and near infrared) was combined with FT-ICR MS in a multiblock analysis to predict the density of crude oils. Two different strategies for data fusion were attempted, and sequential fusion of the blocks achieved the highest prediction accuracy both before and after reducing the dimensions of the data sets by variable selection.

As crude oils have such complex matrixes, samples are often very different, and many methods are not able to handle high degrees of variations or non-linearities between the samples. Hierarchical cluster-based partial least squares regression (HC-PLSR) clusters the data and builds local models within each cluster. HC-PLSR can thus handle non-linearities between clusters, but as PLSR is a linear model the data is still required to be locally linear. HC-PLSR was therefore expanded into deep learning (HC-CNN and HC-RNN) and SVR (HC-SVR). The deep learning-based models outperformed HC-PLSR for a data set predicting average molecular weights from hydrolysed raw materials.

The analysis of the FT-ICR MS spectra revealed that the large amounts of information contained in the data (due to the high resolution) can disturb the predictive models, but the use of variable selection counteracts this effect. Several methods from machine learning and multivariate statistics were proven valuable for prediction of various parameters from FT-ICR MS using both classification and regression methods.

Sammendrag

Gasshydrater er et av hovedproblemene for Flow assurance i olje- og gassnæringen ettersom at de kan forårsake blokkeringer i oljerørledninger og prosessutstyr som krever at systemet må stenges ned. Tidligere studier har vist at noen råoljer danner hydrater som ikke agglomererer eller avsetter, men som forblir som transporterbare dispersjoner. Dette antas å være på grunn av naturlig forekommende komponenter til stede i råoljen, men til tross for årevis med forskning er deres nøyaktige strukturer enda ikke bestemt i detalj. Noen studier har indikert at disse komponentene kan stamme fra syrefraksjonene i oljen eller være relatert til asfalteninnholdet i oljene. Råoljer er blant verdens mest komplekse organiske blandinger og kan inneholde opptil 100 000 forskjellige bestanddeler, som gjør dem vanskelig å karakterisere ved bruk av tradisjonelle massespektrometre. Den høye masseoppløsningen Fourier-transform ion syklotron resonans massespektrometri (FT-ICR MS) gir en høyere oppløsning enn tradisjonelle teknikker, som gjør FT-ICR MS i stand til å karakterisere råoljer i større grad og muligens identifisere hydrataktive komponenter.

FT-ICR MS spektre inneholder vanligvis titusenvis av topper, og det er nødvendig å bruke databehandlingsmetoder i stand til å håndtere store datasett, med muligheter til å finne underliggende forhold for å analysere spektrene. Maskinlæring og multivariat statistikk har mange metoder som er passende for store datasett. En litteratur studie identifiserte flere metoder og den nåværende statusen for bruken av maskinlæring for analyse av gasshydrater og FT-ICR MS data. Litteraturstudien viste at selv om mange studier har brukt maskinlæring til å predikere termodynamiske egenskaper for gasshydrater, har lite arbeid blitt gjort med å analysere gasshydrat relaterte prøver målt med FT-ICR MS.

For å bistå identifikasjonen ble en suksessiv akkumuleringsprosedyre for å øke konsentrasjonene av hydrataktive komponenter utviklet av SINTEF. Sammenligninger av massespektrene fra spikede og uspikede prøver viste at noen topper økte sammen med spikingnivåene. Flere klassifikasjonsmetoder ble brukt i kombinasjon med

variabelseleksjon for å identifisere topper relatert til hydratformasjon. Molekylformler ble bestemt og toppene ble antatt å være relatert til asfaltener, naftener og polyetylenglykol. For å bistå karakteriseringen av oljene ble infrarød spektroskopi inkludert med FT-ICR MS i en multiblokk analyse for å predikere tettheten til råoljene. To forskjellige strategier for datafusjonering ble testet og sekvensiell fusjonering av blokkene oppnådde den høyeste prediksjonsnøyaktigheten både før og etter reduksjon av datasettene med bruk av variabelseleksjon.

Ettersom råoljer har så kompleks sammensetning, er prøvene ofte veldig forskjellige og mange metoder er ikke egnet for å håndtere store variasjoner eller ikke-lineariteter mellom prøvene. Hierarchical cluster-based partial least squares regression (HC-PLSR) grupperer dataene og lager lokale modeller for hver gruppe. HC-PLSR kan dermed håndtere ikke-lineariteter mellom gruppene, men siden PLSR er en lokal modell må dataene fortsatt være lokalt lineære. HC-PLSR ble derfor utvidet til convolutional neural networks (HC-CNN) og recurrent neural networks (HC-RNN) og support vector regression (HC-SVR). Disse dyp læring metodene utkonkurrerte HC-PLSR for et datasett som predikerte gjennomsnittlig molekylvekt fra hydrolyserte råmaterialer.

Analysen av FT-ICR MS spektre viste at spektrene inneholder veldig mye informasjon. Disse store mengdene med data kan forstyrre prediksjonsmodeller, men bruken av variabelseleksjon motvirket denne effekten. Flere metoder fra maskinlæring og multivariat statistikk har blitt vist å være nyttige for prediksjon av flere parametere from FT-ICR MS data ved bruk av både klassifisering og regresjon.

List of papers

Paper I

Gjelsvik E. L., Fossen M., Tøndel K. (2023) Current overview and way forward for the use of machine learning in the field of petroleum gas hydrates, *Fuel*, 334(Part 2), 126696, doi: 10.1016/j.fuel.2022.126696

Paper II

Gjelsvik E. L., Fossen M., Brunsvik A., Tøndel K. (2022) Identifying components related to hydrate formation using machine learning-based variable selection, *TEKNA 33rd Oil Field Chemistry Symposium*, Geilo, Norway

Paper III

Gjelsvik E. L., Fossen M., Brunsvik A., Tøndel K. (2022) Using machine learning-based variable selection to identify components related to hydrate formation *PLoS ONE* 17(8): e0273084, doi: 10.1371/journal.pone.0273084

Paper IV

Gjelsvik E. L., Fossen M., Brunsvik A., Liland K. H., Tøndel K., Multiblock analysis combining data from FT-ICR MS, FTIR and NIR spectroscopy improves prediction of the density of crude oils, *Submitted to Applied Spectroscopy*

Paper V

Gjelsvik E. L., Tøndel K., Hierarchical cluster-based deep learning, *Manuscript*

Additional Scientific Work

Conference Papers

Gjelsvik E. L., Fossen M.,. Brunsvik A,Tøndel K., Machine learning as a basis for better understanding of flow assurance through FT-ICR-MS analysis of gas hydrates. 32nd Oil Field Chemical Symposium, TEKNA (2021)

Gjelsvik E. L., Fossen M., Brunsvik A., Tøndel K., Exploring the possibilities of a regression model for the prediction of wetting index from crude oils. 34th Oil Field Chemical Symposium, TEKNA (2023) Geilo, Norway

Oral presentations

Gjelsvik E. L., Fossen M., Brunsvik A., Tøndel K., Towards a Machine Learning Based Procedure for Interpretation of Mass Spectra for Better Understanding of Hydrate Phenomena in Oil Systems. 17th Scandinavian Symposium on Chemometrics (2021) Aalborg, Denmark

Gjelsvik E. L., Fossen M., Brunsvik A., Tøndel K., Developing Machine Learning Models for Identifying Chemical Components from Wide and Short FT-ICR Mass Spectrometry Data. Data Analysis in Spectroscopy and Imaging Young Scientists Workshop (2022) Ås, Norway

Gjelsvik E. L., Fossen M., Brunsvik A., Weging S., Tøndel K., Utilization of Machine Learning on FT-ICR MS Spectra for Improved Understanding and Prediction of the Properties of Hydrate-active Components. European Conference on Gas Hydrates (2022) Lyon, France

Fossen M., Brunsvik A., Gjelsvik E. L., Wolden M., Lund A., Tøndel K., A New High Pressure Method for Successive Accumulation of Hydrate Active Components. European Conference on Gas Hydrates (2022) Lyon, France

Contents

List of Figures

List of Tables

Abbreviations

AA	Anti-agglomerant
ANN	Artificial Neural Network
API	American Petroleum Institute gravity scale
APPI	Atmospheric Pressure Photoionisation
ARN	Tetrameric acids
CNN	Convolutional Neural Network
DBE	Double Bond Equivalent
DT	Decision Tree
EISC	Extended Inverted Signal Correction
ESI	Electrospray Ionisation
EMSC	Extended Multiplicative Signal Correction
FCM	Fuzzy-C-Means
FT-ICR MS	Fourier Transform Ion Cyclotron Resonance Mass Spectrometry
FTIR	Fourier Transform Infrared Spectroscopy
H/C	Hydrogen-Carbon ratio
HC-CNN	Hierarchical Cluster-based Convolutional Neural Networks
HC-PLSR	Hierarchical Cluster-based Partial Least Squares Regression
HC-RNN	Hierarchical Cluster-based Recurrent Neural Networks
HC-SVR	Hierarchical Cluster-based Support Vector Regression
HPLC	High Performance Liquid Chromatography
ICR	Ion Cyclotron Resonance
IR	Infrared
ISC	Inverted Scatter Correction
KHI	Kinetic Hydrate Inhibitor
LASSO	Least Absolute Shrinkage and Selection Operator
LDA	Linear Discriminant Analysis
LDHI	Low Dose Hydrate Inhibitor
LS	Least Squares
MB-PLSR	Multiblock Partial Least Squares Regression

MS	Mass Spectrometry
MSC	Multiplicative Scatter Correction
m/z	Mass-to-Charge ratio
NIR	Near-infrared Spectroscopy
OLS	Ordinary Least Squares
PC	Principal Component
PCA	Principal Component Analysis
PEG	Polyethylene Glycol
PLSR	Partial Least Squares Regression
RF	Random Forest
RNN	Recurrent Neural Networks
S/N	Signal-to-noise ratio
SO-PLSR	Sequential Orthogonal Partial Least Squares Regression
SPC	Spectral Clustering
SVM	Support Vector Machine
SVR	Support Vector Regression
THI	Thermodynamic inhibitor
UV	Ultraviolet
VIP	Variable Importance in Projection
WI	Wetting Index
XGBoost	eXtreme Gradient Boosting

1 Introduction

During transportation of oil and gas from the well to the production site, ice formations can occur in the pipelines and these ice formations are known as gas hydrates. This section will go through the chemistry involved in the formation of gas hydrates, the oil industries´ effort to avoid them and how the work presented in this thesis can contribute to an increased understanding of their formation.

1.1 Gas hydrates

Gas hydrates are crystalline structures where smaller guest molecules are trapped inside cages formed by water molecules through hydrogen bonding at low temperatures and high pressures [1, 2]. The gasses involved are usually light hydrocarbons such as methane, ethane, propane, and iso-butane, in addition to carbon dioxide (CO_2) and hydrogen sulphide (H_2S), where one or more enter the hydrate cages during formation. Hydrates are mainly formed as one of three crystallographic structures described by the comprising number of cages, where the smallest is dodecahedron referred to as sI, the second tetrakaidecahedron referred to as sII, and hexakaidecahedron referred to as sH, but hydrates can also exist in other sizes and shapes [3]. The gas composition is of relevance to gas hydrate formation, deciding which hydrate structures are formed, which gasses that enter the hydrate cages and the overall thermodynamics. Gas hydrates can lead to complete blockage of pipelines and production equipment and are therefore among the main flow assurance issues in the oil and gas industry.

Although gas hydrates have been known for over 200 hundred years, with Sir Humphry Davy's discovery in 1810 [4], Hammerschmidt was the first to acknowledge their presence and start explaining the "freezing" of water in gas pipelines [5]. This marks the start of the modern research on hydrate thermodynamics and in later years hydrate formation kinetics [6]. Gas hydrates in the oil and gas industry have, since their discovery, been treated with addition of chemicals or by operating out-

side the hydrate region by controlling the pressure and/or temperature. Currently, the most common strategy for hydrate inhibition is the use of thermodynamic inhibitors (THIs). THIs are chemicals such as methanol, ethanol or glycols, and these inhibitors shift the hydrate curve towards lower pressures at hydrate inducing temperatures, enabling production at lower temperatures without the formation of gas hydrates [7, 8]. For this type of inhibitor to work, 20-50 % relative to the mass of the water phase is needed. Its premise is that without sufficient amount of THI, hydrate formation is expected and the inhibitor is therefore always present in the pipelines. Another promising strategy is the use of low dose hydrate inhibitors (LDHI) which consist of kinetic hydrate inhibitors (KHI) and anti-agglomerants (AAs) [9]. The purpose of the AAs is to form a slurry of gas hydrates dispersed in the oil phase, which can be transported through the pipelines without the particles aggregating together or depositing on the pipe surface. AAs can perform at higher subcoolings than KHIs, making them applicable even for deep water use. Subcooling is the process of lowering the temperature of a liquid below its freezing point, without forming solids. Different types of AAs exist, but they are usually surfactants that either stabilise an emulsion or bind to the hydrate surface and alter it from hydrophilic to hydrophobic, and thereby disrupt growth [10]. The latter of the two types is the most commonly used, and usually consists of quaternary ammonium surfactants where the part that binds to hydrates consists of to or more *n*-butyl, *n*-pentyl or iso-pentyl groups [7]. A KHI on the other hand, binds to the hydrate surface, decreasing the crystal formation process by preventing growth, in order to delay formation long enough to reach the storage facility without causing blockage [11, 12]. KHI formulations consist of water-soluble polymers with functional groups that can create hydrogen-bonds to water molecules or gas hydrate surfaces, and a hydrophobic group either adjacent to or bonded directly to each amide group [13]. However, for an LDHI to be efficient, it must be surface active and able to adsorb to the surface or interact with the hydrate cages of the dispersed hydrate particles. They are also highly dependent on the composition of the oil, where high contents of some components can depreciate the effect of the LDHIs. A typical concentration for an LDHI injection is 0.1-1 wt % relative to the water phase.

However, operating outside the hydrate region is not always possible, and the addition of chemicals poses an environmental threat and an increase in production costs. Moreover, the addition of chemicals can deteriorate the quality of the oil. As drilling and oil extraction technologies have improved, oilfields in subsea areas where hydrate forming conditions are frequent have become more common. The oil industry is therefore in need of an environmentally friendly, non-destructive solution which can deal with subsea conditions.

When blockage of a pipeline occurs, the production has to shut down while the gas hydrate plug is removed. Removal of hydrates is associated with high risk and high cost, and is usually handled by dissociating the plug using either depressurisation, injection of chemicals that generate heat, thermal methods such as electrical heating, or mechanical methods such as coiled tubing or drilling [2, 14]. During the dissociation process, the hydrate plug can detach from the pipe wall and move down the pipeline. In extreme cases detached hydrates can act as projectiles, damaging pipes or equipment, they are therefore treated with great care and should be avoided.

1.2 Production chemistry

Production chemistry constitutes managing the chemical reactions of the phases produced between the reservoir and refinery, with the aim of maximising the flow assurance in the system. Flow assurance is the process of ensuring successful and economical flow from reservoir to end user. Production chemistry deals with issues such as fouling problems; deposition of unwanted matter in a system, problems related to physical properties; foams, emulsions or viscous samples, corrosion related issues and environmental issues; oil discharge or for instance hydrogen sulphide gas. Although non-chemical approaches such as heating, insulation, filtering or altering the flow can be used in some cases, chemical additives often have to be applied in addition to fully resolve or rectify these issues. Production chemicals are mainly classified as inhibitors which minimise fouling or remove deposits, process aids to improve the separation of gas from liquids and water from oil, corrosion inhibitors or chemicals with other benefits such as environmental compliance. During the transportation of oil and gas from well to reservoir for instance, chemicals such as drag reducers, depressants, odorising additives, hydrate inhibitors, surfactants, corrosion inhibitors, scale inhibitors and paraffin inhibitors are added [15]. Many of the production chemicals are environmentally challenging, toxic to the surrounding environment and with restricted or slow biodegradability. In recent years, the focus on hazardous production chemicals has increased and most regions now have environmental regulations. However, these regulations vary between regions. Norway is part of a Harmonised Mandatory Control Scheme valid for the North-East Atlantic areas, which includes regular testing of produced water for toxic components and residues of production chemicals, to keep them under the legal levels [16]. These regulations are valid for the entire North Sea and are considered to be the most complex environmental regulations for toxicity, biodegradation and bioaccumulation.

Green chemistry is an emerging area, focusing on the development of products and processes which reduce or eliminate the use of hazardous chemicals, with particular

focus on the environmental impact of chemistry. A substantial amount of work has been done to develop production chemicals designed to be "greener", more environmentally friendly, and adhering to the regulations, but they usually have lower performance [10]. Reducing toxicity and bioaccumulation of the chemicals, and increasing the biodegradability are important focus areas.

1.3 Emulsions

Production of crude oils, most often involve co-production of a mixture of gas, water, oil and solid particles, where the oil and water often form emulsions and the solids can be suspended in the liquid. In an emulsion containing colloidal solid particles or droplets, it is likely that one of the liquids will wet the solids more than the other liquids [17]. Emulsions are mixtures of two or more liquids that are normally immiscible, where one phase is dispersed in another, continuous phase. The phase with droplets is referred to as the dispersed phase, while the phase they're suspended in is referred to as the continuous phase. In crude oils, emulsions are either water continuous (oil-in-water) or oil continuous (water-in-oil), and they are stabilised by components inherent in the crude oil or by the addition of chemicals [18]. In a crude oil system, emulsions are created when oil and water come in contact and there is sufficient mixing, or when an emulsifying agent (also called emulsifier) is present which reduces the energy needed to increase the systems interface during mixing, i.e creating droplets. In the oil industry emulsions are normally undesired as the phases have to be separated before refining, which increases the production costs. Surfactants are commonly used as emulsifying agents and added during production to break the emulsions, as well as keep them from forming during extraction and transportation. The surfactants can also be adsorbed to the particle surface of suspensions, where they can alter the wettability of the particle and induce different emulsion behaviours. Figure 1.1 shows a water-in-oil and a oil-in-water emulsion and how a surfactant with a hydrophilic head as hydrophobic tail can adsorb to the droplets. The colloidal particles suspended in the liquid can also adsorb to droplets, and thereby stabilise the emulsions.

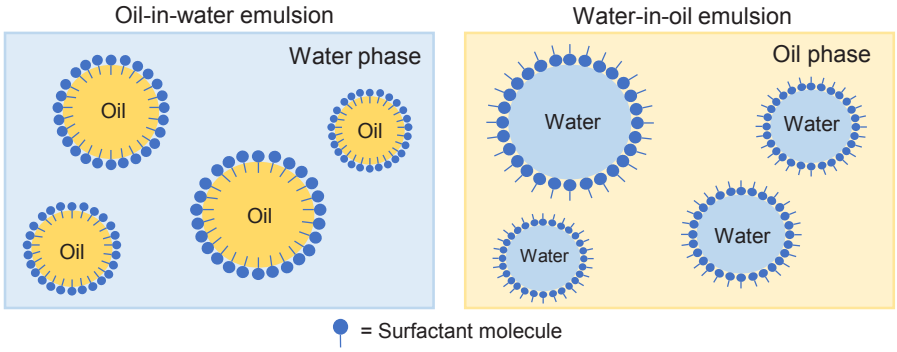


Figure 1.1: Illustration of oil-in-water and water-in-oil emulsions, and how a surfactant with a hydrophilic head and hydrophobic tail will stabilise the emulsions by adsorbing to the oil and water molecules.

The hydrate plugging in crude oil systems has, in addition, previously been related to the wettability state of the hydrate particles when formed, where oil-wet (hydrophobic) particles are associated with transportable dispersions and water-wet (hydrophilic) with aggregation of hydrate particles and a higher potential for plugging [19]. The surface energy, measuring the intramolecular interactions of a solid, has an impact on wetting, adsorption and adhesion on the hydrate, where oil-wetted particles tend to stabilise water-in-oil emulsions while water-wetted particles tend to stabilise oil-in-water emulsions. The wettability of a solid, like a hydrate particle, immersed in oil is determined by the contact angle of a water drop on the horizontal surface under thermal equilibrium. The wettability of a clean hydrate surface would be towards water and thus result in a higher potential for agglomeration and plugging, as gas hydrate particles initially are water-wet. However, naturally occurring components in the crude oil can affect the wettability of hydrate particles by adsorbing to the hydrate, making them oil-wet and altering the hydrate surface towards hydrophobic. Oil-wet hydrate particles will no longer be exposed to water bridging and will disperse into the oil phase as smaller particles, something that will most often alter the inversion point and properties of the emulsion.

Accordingly, the wettability of hydrate particles can be altered when formed in a crude oil system, by interaction (adsorption or inclusion) of hydrate-active components. The degree of alteration of the wettability will depend on the crude oil composition, since variation is attributed to differences in the type and/or amount of surface active components that adsorb to the hydrate surface [20].

1.4 Crude oil chemistry

Crude oils are among the worlds most complex organic mixtures [21] which in addition to the relatively high mass of their components, makes detailed analysis difficult. This is because many traditional analytical techniques does not have high enough resolving power to separate a large number of species, or mass ranges that cover large masses. Crude oils are formed from organic material situated under sedimentary rocks, subjected to heat and high pressures over long time periods. The composition of the oil is therefore dependent on the types of organic material and its constituents. Naturally, different geographical regions have organic materials with varying composition and it has been shown that the composition of the oils is dependent on their geographical origin [22]. However, there are some similarities; all oils are organic and the composition of crude oils normally consists of 83-87 % carbon, 10-14 % hydrogen and varying small amounts of nitrogen, oxygen, sulphur and metals such as nickel and vanadium [23]. The most common constituents are alkanes, cycloalkanes and aromatic hydrocarbons, generally containing between 5 and 40 carbon atoms per molecule. These constituents can all be parts of a homologeous series. A homologeous series is a sequence of compounds with a fixed set of functional groups, which therefore have similar chemical and physical properties, while being separated by a fixed mass unit. E.g. in an alkane series, either straight or branched, from one alkane to the larger alkane next in line, one CH_2, 14u, is added to the molecule.

One way to gain an understanding of the molecular structures of crude oil constituents is through the double bond equivalent (DBE). The DBE is a measure of the number of double bonds and rings in the molecular formula, and can be calculated for a molecule with elemental formula $C_cH_hN_nO_oS_s$ by

$$DBE = c - \frac{h}{2} + \frac{n}{2} + 1 \tag{1.1}$$

DBE does not however denote double bonds between other elements than carbon, hydrogen and nitrogen, as shown by Equation 1.1. The crude oil constituents can be sorted according to the DBEs, and along with the hydrogen-carbon (H/C) ratios and the molecular formulas, interpretations of the molecular structures for each component can begin to emerge.

Due to the complexity of the oils, it is often expedient to separate the chemical constituents of interest from the remainder. One method for dividing the crude oils

into distinct fractions is the SARA (Saturates, Aromatics, Resins and Asphaltenes) fractionation scheme developed by Jewel *et al* [24]. In SARA, the oil sample is separated into four main chemical classes based on solubility and polarity, as listed in Table 1.1. The resin and asphaltene fractions are similar, but the asphaltene fraction has higher molecular weight [25, 26]. The asphaltene fraction also contains the largest percentage of heteroatoms (oxygen, sulphur and nitrogen) and organometallic constituents (nickel, vanadium and iron) [27].

Table 1.1: The four SARA fractions and the chemical composition of each fraction.

Constituent	*Composition*
Asphaltenes	Condensed aromatic rings with heteroatoms and cyclic unsaturated compounds (with alternating double bonds and substituted alkyl chains)
Resins	Polar compounds often containing heteroatoms (such as nitrogen, oxygen or sulphur)
Aromatics	Aromatic rings (unsaturated rings)
Saturates	Saturated hydrocarbons (straight chained or branched) non-polar compounds

Crude oils are often categorised according to their geographical origin, the American Petroleum Institute (API) gravity scale and sulphur content. For instance, a crude oil is considered light if it has low density and heavy if it has high density. Density is inversely related to the API scale, where the lighter the oil is, the higher the API gravity. API gravity is a measure of the density of a petroleum liquid relative to water, through its specific gravity, i.e. a measure of a liquids density compared to water. The API gravity for a sample is calculated by

$$^{\circ}API = \left(\frac{141.5}{\text{specific gravity}}\right) - 131.5 \tag{1.2}$$

The API can to a degree also be related to the amount of aromatics and compounds containing heteroatoms, like asphaltenes and resins [28]. A high content of aromatics is related to heavy oils and lower APIs, while light oils have a higher alkane content [29]. The differences in composition between the density fractions make the oils easily distinguishable, and are important for the oil price. The higher the alkane content of the oil is, the more gasoline and diesel fuel can be produced. Consequently, light oils receive a higher price than heavy oils. Additionally, crude oils are considered as "sweet" when they have a low content of sulphur and "sour" if they contain substantial amounts of sulphur. Sweet oils are more desirable because they require less refining to reduce the sulphur content which is toxic to the environ-

ment and corrosive. Accordingly, all the above contribute to the unique molecular composition of each crude oil.

1.4.1 Asphaltenes

The asphaltene fraction constitute the most complicated components of crude oils, they are highly surface active in addition to being the largest, densest and most polar [30]. They do not have a specific chemical formula, and the individual molecules vary in the number of atoms contained in the structure. Asphaltenes can be regarded as a collective term for a solubility fraction of molecules with similar solubilities. The exact molecular structures for this group of molecules are difficult to determine, but are generally regarded to comprise of mainly poly-aromatic ring structures with heteroatoms (oxygen, nitrogen and sulphur), small amounts of metallic constituents (nickel, vanadium and iron) and aliphatic side chains. One of the reasons for the difficulties in determination of molecular structures is that the molecules aggregate in solution. The degree of aggregation is highly dependent on the source of the asphaltene, the surrounding chemistry and physical environment, and the underlying mechanisms are still unclear [31]. However, after the development of high resolution methods such as Fourier Transform Ion Cyclotron Resonance Mass Spectrometry (FT-ICR MS), asphaltene structures have been slightly more illuminated. The current general consensus is that the masses of asphaltenes mainly lie in the range 500 Da to 1000 Da with an average of ∼750 Da [30]. It is worth to note that for asphaltene characterisation, thousands of different molecular structures exists, and this can simply be regarded an average of all the individual ashpaltene molecules.

1.4.2 Naphthenic acids

Naphthenic compounds are found in the resin fraction and consist of a complex mixture of alkyl-substituted acyclic and cycloaliphatic carboxylic acids with the general formula $C_nH_{2n+z}O_2$ where n corresponds to the number of carbon atoms and z specifies the hydrogen deficiency from ring formation [32]. Naphthenic acids constitute about 2-4 % of the average crude oil composition [33]. Molecular weights between 115-1500 Da have been reported with an average weight between 300-500 Da [34]. In light crude oils naphthenic acids are usually present at low concentrations, while the concentrations are higher in heavier oils [35].

Some naphthenic acids have been shown to contribute to corrosion of metals and formation of metal salts called naphthenates, which can stabilise or even precipitate emulsions [36]. As naphthenic acids have amphiphilic structures, they tend to partition between the oil phase and the water phase, remaining at the interface. This

is however dependent on the molecular size, structure and pH [37].

Additionally, some napththenic acid species have been proven to be toxic and carcinogenic, and can therefore have detrimental effects on the environment. They are especially toxic for aquatic life, and efforts have to be made to avoid leakages from offshore instalments to the surroundings. Tetrameric acids (so called ARNs) are one of the more problematic of the high mass naphthenic acid species, containing four carboxylic groups (8 oxygen atoms). ARNs can form interfacial gels using their four carboxylic endings to cross-link calcium ions and produce an insoluble salt, easily adherable and with high interfacial activity. Because of the high number of carboxylic groups ARNs are among the most toxic naphthenic acid species.

1.5 Naturally occurring hydrate active compounds

Through field experience and laboratory experiments, it became evident that some crude oils form hydrates that do not agglomerate or deposit, but remain as transportable particles [38]. The most common explanation is that this is due to naturally occurring components in the crude oil with hydrate active properties, that render the surface of the particles to be hydrophobic. One possibility is that these compounds have the ability to adsorb to the hydrate surface, preventing the agglomeration of the hydrates [39]. Another possibility is that parts of a molecule, for example butyl/pentyl groups, penetrate open $5^{12}6^4$ cavities on the hydrate surface and can even become embedded in the hydrate surface as the hydrate grows around the alkyl groups [7]. Despite decades of investigation, their exact structures have not yet been determined in detail.

However, some previous studies have suggested that these natural inhibitors may be contained in the petroleum acid fraction [38, 40, 41, 42, 43, 44, 45] which has also shown surface activity towards hydrate surfaces. Borgund et al. [41] and Erstad et al. [45] showed experimentally the anti-agglomerating properties of some petroleum acid fractions, and therein naphthenic acid compounds [41]. Naphthenic acids have also been shown to stabilise water-in-oil emulsions [46].

The acidic constituents in crude oils have been shown to be products of biodegradation [47] and studies have indicated that biodegradation of the oil in the reservoir may be necessary for the formation of water-wet hydrates [39]. It has further been suggested that the level of the biodegradation is an important factor for the hydrate plugging tendency of the crude oils [20]. During aerobic biodegradation, aliphatic hydrocarbons are consumed by bacteria, producing carbonic acids. These acids are,

by their amphiphilic nature, surface active compounds and can dramatically change the properties of the oils. Anaerobically degraded oils on the other hand, have proved to alter the wettability of the gas hydrates to be oil-wet, leading to a lower tendency for the gas hydrates to agglomerate or deposit to the pipe wall, and thus a lower plugging tendency [45]. Additionally, Høiland et al. [20] showed that some of the components acting as natural hydrate inhibitors are produced by the bacteria and that the type of inhibitor compound present is more important than the amount [48, 49].

Similarly, the asphaltene fractions are known to possess self-aggregating properties that can stabilise some crude oil systems [50], and some asphaltenes can alter the plugging potential of hydrates [51, 30]. It has been shown that the asphaltene fractions able to stabilise systems prone to form transportable slurries often are more polar, with higher oxygen content, higher acidity and lower DBEs [52]. Other studies have suggested that the possible hydrate activity of asphaltenes is related to their sulfoxide content [53].

1.6 Research aim and objectives

This thesis is part of a larger project with the primary objective to develop new fundamental knowledge of how natural components in crude oils affect gas hydrate properties. This knowledge can be used for increased safety and reduced costs related to the management of gas hydrates during oil production. The main goal of this work was to develop methods that could contribute to identify naturally occurring hydrate active components from FT-ICR MS spectra. To accomplish this, a set of secondary objectives with measurable goals were defined:

- Identify hydrate-active oil components from FT-ICR MS spectra of crude oils
- Develop new data science methods for correlating FT-ICR MS spectra to crude oil properties
- Correlate structures and concentrations of hydrate-active components to the wetting properties of the oil
- Transfer the accumulated knowledge and developed methods to project partners for further use

The exact structures of these hydrate active components have not yet been determined in detail, although many hypotheses exist regarding their compositions. The

complexity of the matrix of crude oils makes their identification a challenge, and FT-ICR MS is one of the few mass spectrometers able to handle such highly complex samples. FT-ICR MS allows for more detailed analysis of crude oils, which could make it possible to quantify masses which previously have been undetected. However, this high mass accuracy results in very high-dimensional data. In recent years, chemometrics have become more frequently used by chemists with the increase in the sensitivity of instruments and the following increase in data amounts [54]. Data analysis for large, complex data sets is not as commonly used in the oil and gas industry, but the use has increased in recent years [55]. The main aim of this project was therefore to use machine learning to come one step closer to the identification of hydrate active components. To investigate the current state of the use of machine learning in the field of gas hydrates, a literature study was performed. The results are shown in **Paper I** [56], where a text mining study of the Scopus database [57] was performed followed by a text analysis study to identify the main topics in the extracted articles.

An additional goal was to develop new methods to facilitate the extraction and identification of hydrate active components. A successive accumulation procedure (spiking) with the aim of accumulating the components was developed with the hope of increasing their concentration [58, 59]. In **Paper II** [60] and **Paper III** [59] variable selection was used to identify the variables related to hydrate formation from FT-ICR MS spectra and their molecular formulas were interpreted. Variable selection was continued into **Paper IV** where the densities of crude oils were predicted from FT-ICR MS, Fourier Transform Infrared (FTIR) and Near-infrared (NIR) spectroscopy. NIR and FTIR measurements were included in the study to assess whether they could provide any additional information or improve the models, as FT-ICR MS yielded poor prediction accuracy. The variable importance in projection (VIP) for each variable was used to identify important variables and evaluate their chemistry. Two multi-block techniques using different ways of fusing the data were compared to Partial Least Squares Regression (PLSR) in an attempt to use the additional information from the IR measurements to improve the accuracy and interpretability of the model.

Another part of the project was to develop new, generally applicable, machine learning methods. In **Paper V** the Hierarchical Cluster-based Partial Least Squares Regression (HC-PLSR) developed by Tøndel *et al.* [61], was expanded into deep learning using convolutional neural networks (CNNs), recurrent neural networks (RNNs) and support vector regression (SVR). These methods should model non-linear and heterogeneous data better than PLSR, which is a linear regression method. How-

ever, neural network based methods need substantial amounts of data for training to create stable and well validated local models, and the project did not generate sufficient gas hydrate related data to test these methods. A previously published data set containing FTIR measurements of raw materials from chicken, turkey, salmon and mackerel was therefore used in **Paper V**. When more data has been generated on oil systems, HC-CNN, HC-RNN and HC-SVR should be valuable methods for analysis of crude oil data as well.

1.6.1 Time consumption and chemical usage reduction

One of the main goals of chemometrics is to use statistical methods for fast and reliable analysis of spectroscopic data to replace time consuming wet chemistry methods, thereby reducing chemical consumption. As described previously in this chapter, hydrate inhibition uses large amounts of chemicals where many of them are toxic to the surrounding environment. However, naturally occurring components inhibiting hydrate formation are already present in the oil, and if they could be utilised, the need for addition of chemicals could be reduced. Many of the analysis methods for determining crude oil properties are still complicated, as well as time and chemical consuming. For instance, the developed method for measuring the wetting properties for a crude oil takes on average one week and it is difficult to ascertain the accuracy of the measurements in contrast to spectroscopic techniques, which are fast and highly reproducible. One such spectroscopic technique is FT-ICR MS, and relating crude oil and hydrate properties to FT-ICR MS spectra will allow for faster determination of the plugging potential when the crude oil forms hydrates. In the following chapters the mass accuracy and advantages of FT-ICR MS will be demonstrated.

1.6.2 Further use of developed methods

The work presented in this thesis was part of a knowledge-building project to increase the understanding of chemical properties of crude oils, specifically related to naturally occurring hydrate active components. The project was a collaboration between SINTEF, the Norwegian University of Life Sciences, and industrial partners Equinor ASA, OMV (Norge) AS, Wintershall DEA Norge and TotalEnergies funded by the Research Council of Norway. Funding was received from the PETROMAKS 2 program aiming at petroleum-oriented knowledge-building projects for industry, with the project number: 294636 and project title “New Hydrate Management: New understanding of hydrate phenomena in oil systems to enable safe operation within the hydrate zone”. An important part of this work was therefore to transfer the

acquired knowledge and developed methods to all project partners for further use. It was therefore of great importance to develop efficient, well performing models, and easily understandable code which can also be used by personnel having little programming expertise in the future.

2 Theory

To identify the naturally occurring hydrate active components, oil samples were measured using FT-ICR MS and machine learning was applied to analyse the resulting data. Infrared spectroscopy was also utilised for crude oil characterisation to compare it´s effectiveness to that of FT-ICR MS. In this chapter an introduction into all applied methods are given.

2.1 Mass spectrometry

Mass spectrometry (MS) is an analytical technique where ions are produced and measured by their the mass-to-charge (m/z) ratio and presented in a mass spectrum. For the m/z-ratio in mass spectrometry, the m refers to the mass of the ion while z refers to the number of charges of the ion. Therefore, when the number of charges on the ion is one, the m/z-ratio equals the molecular mass of the ion. In MS, a sample is ionised into charged fragments before being separated by subjection to an electric or magnetic field. This is the ion analyser which separates the ions according to their mass. The resolution of the MS technique is determined in the analyser, and this is dependent on the type used and its geometry. Various types of analysers exist, and they are often separated into two broad groups. One group contains scanning analysers, where ions of different masses are transmitted successively along a time scale. This group consists of magnetic sector instruments which only allow ions with a given m/z ratio to pass through at a time, or quadrupole instruments with oscillating electric fields [62]. Another group generally allows for simultaneous transmission of all ions, and includes the dispersive magnetic analyser, time-of-flight analysers and the trapped-ion mass analysers consisting of ion traps, ion cyclotron resonance and orbitrap instruments. In the final step, the desired ions selected in the analyser are detected in a detector and presented in the mass spectrum, which is a record of the abundance of each ion reaching the detector plotted against the ions m/z values.

2.2 FT-ICR MS

Fourier-transform ion cyclotron resonance mass spectrometry (FT-ICR MS) [63] is a type of mass analyser that determines the m/z-ratio of ions based on the cyclotron frequency of the ion in a fixed magnetic field. The mass accuracy for FT-ICR MS is sub ppm and the mass spectral resolution can be above 10 million (at m/z=400), which allows identification of a large number of different polar and non-polar groups [64, 65, 66]. In an FT-ICR MS analysis, ions are detected simultaneously within a detecting interval by the ion cyclotron resonance frequency they produce when they rotate in a magnetic field. This provides an increase in signal-to-noise (S/N) ratio compared to traditional mass spectrometers. A schematic illustration of how an FT-ICR MS instrument works is shown in Figure 2.1. A sample is introduced and ionised before the ions are trapped in a magnetic field by electric trapping plates. The ions are then excited at their resonant cyclotron frequencies to a larger cyclotron radius by an oscillating electric field orthogonal to the magnetic field. When the ions are rotating at their cyclotron frequency they create a charge which is detected by the detection plates when the ions come in close proximity. The resulting signal is a transient or interferogram, consisting of a superposition of sine waves. The transients are the frequencies of the ion oscillations, measured as time-domain signals and presented in an interferogram. The signals are then extracted by performing Fourier transformation, first into a frequency spectrum, and after mass correction, to a mass spectrum.

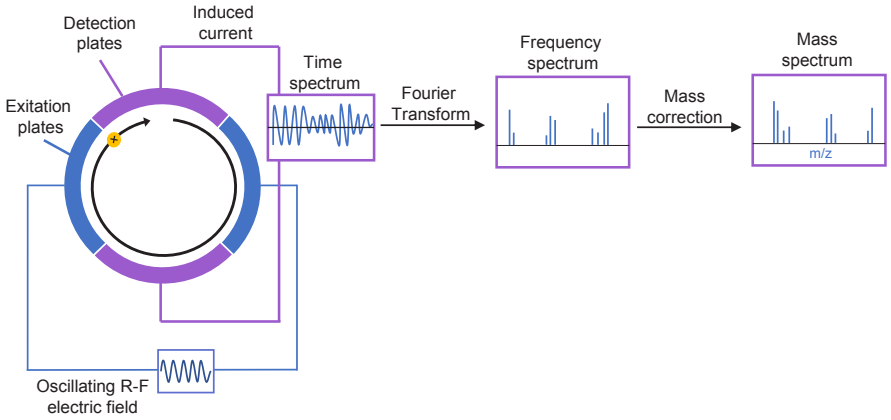


Figure 2.1: Schematic illustration of a FT-ICR MS instrument with the ion trapping, detection, signal generation and conversion.

2.2.1 General principle

Trajectories of ions are curved in a magnetic field, and if the velocity of the ion is low and the field is intense, the radius of the trajectory becomes small [62]. This means that the ions become "trapped" on a circular trajectory in the magnetic field. When an ion of mass m and charge q (the cyclotron charge of the ion) moves in a spatially uniform magnetic field B with a velocity of ν, it rotates around the magnetic field direction [67] as shown in figure 2.2.

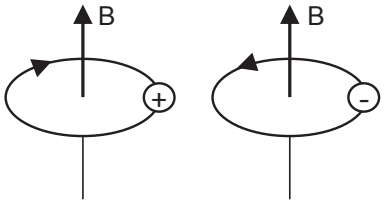


Figure 2.2: Ion cyclotron motion for positive and negatively charged ions moving in a magnetic field, B

The ions rotate in a plane perpendicular to the direction of the spatially uniform magnetic field B, and positive and negative ions orbit in opposite directions. The cyclotron rotational frequency (ω_c) is given by

$$\omega_c = \frac{qB}{m} \tag{2.1}$$

The ion completes a circular trajectory of $2\pi r$ with a frequency (ν_c) which is calculated as

$$\nu_c = \frac{\omega_c}{2\pi r} \tag{2.2}$$

All ions of a given m/z-ratio rotate at the same ion cyclotron resonance (ICR) frequency, which is independent of the velocity, a property that makes ICR especially amenable for mass spectrometry. The ion frequency is relatively insensitive to kinetic energy, meaning that focusing the translational energy (the energy of the ions due to their translational motion) is not essential for precise determination of the ions m/z-ratio. Therefore, the centripetal force (F), the force acting on an object in a curvlinear motion directed towards the axis of rotation or centre of curvature, for

an ion with mass m becomes

$$F = q\nu B \tag{2.3}$$

Correspondingly, the ions centrifugal force ($F^{'}$), which is a pseudo force acting in a circular motion along the radius, directed away from the centre, becomes

$$F^{'} = \frac{m\nu_c^2}{r} \tag{2.4}$$

The ion then stabilises on a trajectory resulting from the balance of these two forces

$$q\nu_c B = \frac{m\nu_c^2}{r} \tag{2.5}$$

This is equal to

$$qB = \frac{m\nu_c}{r}, \tag{2.6}$$

relating back to the relationship in Equation 2.1. As the quadrupolar electrical field used to trap the ions is in an axial direction, this relationship is only approximate. The axial electrical trapping results in axial oscillations within the trap, with the (angular) frequency

$$\omega_t = \sqrt{\frac{q\alpha}{m}}, \tag{2.7}$$

where α is a constant similar to the spring constant of a harmonic oscillator and is dependent on the applied voltage, and the dimensions and geometry of the ion trap. The applied electric field and the resulting axial harmonic motion reduce the cyclotron frequency and introduce a second radial motion, magnetron motion, which occurs at the magnetron frequency. The cyclotron motion is still the used frequency, but the relationship between Equation 2.1 and 2.7 is not exact because of the magnetron motion. The natural angular frequencies of motion are

$$\omega_{\pm} = \frac{\omega_c}{2} \pm \sqrt{\left(\frac{\omega_c}{2}\right)^2 - \frac{\omega_t^2}{2}}, \tag{2.8}$$

where ω_t is the trapping frequency due to the axial electrical trapping, ω_+ is the reduced cyclotron (angular) frequency and ω_- is the magnetron (angular) frequency. ω_+ is typically measured in FT-ICR.

The advantage of the FT-ICR as a mass analyser, is that the m/z-ratio is experimentally manifested as a frequency. As frequencies can be measured more accurately than other experimental parameters, FT-ICR MS is able to achieve higher resolution and thereby also higher mass accuracy than other types of mass measurements [67].

2.2.2 Resolution of FT-ICR MS spectra

The resolution of FT-ICR MS is often described as the full width of a spectral peak at half-maximum peak height ($\Delta m_{50\%}$) for FT-ICR MS spectra in the mass domain. The resolving power for a molecule's FT-ICR MS signal is therefore defined as $m/\Delta m_{50\%}$ constituting the mass resolving power [68]. The frequency resolving power is, based on the relationship in Equation 2.1, equal to the mass resolving power except from a negative sign

$$\frac{\omega}{\Delta\omega_{50\%}} = -\frac{m}{\Delta m_{50\%}} \tag{2.9}$$

The mass resolving power can also be thought of as the number of cyclotron orbits an ion makes during the data acquisition period [69]. It is therefore desirable to confine the ions to the ion trap as long as possible after excitation to achieve maximal mass resolving power.

2.2.3 The Fourier Transform (FT)

The FT is a mathematical transformation that decomposes functions depending on space or time into functions depending on spatial frequency or temporal frequency. For each frequency, the magnitude (absolute value) of the complex value represents the amplitude of a constituent complex sinusoid with that frequency, and the argument of the complex value represents that complex sinusoid's phase offset. If a frequency is not present, the transform has a value of 0 for that frequency [68].

2.2.4 Ionisation techniques

Several different ionisation techniques can be used in combination with FT-ICR MS, and for crude oils the most common are Electrospray Ionisation (ESI) and Atmospheric Pressure Photoionisation (APPI). ESI is achieved by applying a high

voltage to a liquid passing through a capillary tube inducing highly charged droplets [70, 62]. The liquid, often 1-10 μL, is passed through a capillary needle where a potential difference of typically 3-6 kV is applied between the end of the capillary and a cylindrical electrode approximately 0,3-2 cm apart. The liquid becomes a fine mist with highly charged droplets when leaving the capillary, and can either be positively or negatively charged depending on the applied voltage. In positive mode, formic acid is added to the liquid to aid ionisation by protonating basic neutrals, while in negative mode ammonium hydroxide is added to deprotonate acidic neutrals, resulting in lower background noise. ESI can produce multiple charges of the ions and the number of ^{13}C isotope peaks appearing within a single unit on the m/z scale defines the number of charges on the ion [71].

APPI is performed by exposing the sample to photons emitted from an ultraviolet (UV) light source [72, 64, 73]. The liquid consisting of analyte and solvent is vaporised by a nebuliser probe, a probe which disperses the sample, creating a fine mist under temperatures as high as 500°C. The molecules are then ionised using a vacuum UV lamp at atmospheric pressure (105 Pa) and excited, creating the ionised state. Molecules in both the analyte of interest and the solvent are likely to be ionised, and the emission energy of the UV lamp should therefore be in the range between the ionisation potential of the analyte and the ionisation potential of the solvent and air component, thus reducing the amount of impurities. In positive mode, both molecular ($[M^{+\cdot}]$) and protonated ions ($[M+H]^+$) are generated. During negative mode, the ions of the molecular species are produced by either proton abstraction or adduct formation. The predominant ions are the molecular species ions ($[M-H]^-$), which are the ions corresponding to the fatty acids ($R_n - COO^-$) present in the sample [62]. APPI is not suitable for compounds with low thermal stability and is sensitive to aromatic compounds and sulphur containing compounds.

ESI and APPI are popular ionisation techniques as they both can analyse liquids directly from a high performance liquid chromatography (HPLC) column. The two are also defined as soft ionisation techniques because very little fragmentation occurs during ionisation. A disadvantage of ESI is the formation of adducts consisting of analyte and metal ions and this is especially common for compounds with oxygen or sulphur atoms in an orientation where complex formation with alkali metal ions is possible [74]. For crude oil samples, seawater is usually present in the sample matrix and sodium adducts are therefore often observed. When production chemicals are present in the sample, they can infer greatly on the produced mass spectra. If foreign constituents are ionised before introduction to the ionisation source, the ionisation of the analyte in the source is overshadowed. When these ions enter the

detection chamber, they can saturate the detector and suppress the signal from the ionised analyte, leading to lower and in extreme cases no signal. One example of this is polyethylene glycol (PEG) which is a hydrophilic molecule with formula $C_{2n}H_{4n+2}O_{2n+1}$ often used to elongate molecules to alter their solubility. PEGs are common in production chemicals as lubricants, shale stabilisers and demulsifiers [15]. During analysis of the work presented in **Paper II** and **III**, PEG was observed in the mass spectra from ESI(+)-FT-ICR MS and several PEG related structures were selected as important variables. However PEGs have low thermal stability, and are therefore not observed when heating is applied in APPI.

2.3 Infrared spectroscopy (IR) spectroscopy

In IR, the interaction of infrared radiation with matter by absorption, emission or reflection is measured. The spectra are associated with transitions between vibrational energy levels, where functional groups can be identified by their characteristic vibration frequencies [71]. IR is rapid, non-invasive and requires minimal sample preparations. A mass m vibrating with a frequency ν at the end of a fixed bond, illustrates the range of values for the vibrational frequencies of various chemical bonds

$$\nu = \sqrt{\frac{k}{m}} \tag{2.10}$$

Here, k is a measure of the strength of the bond. However, the ends of a chemical bond are not fixed, rather there are two masses (m_1 and m_2) involved where both are able to move. The m in Equation 2.10 is therefore determined by

$$\frac{1}{m} = \frac{1}{m_1} + \frac{1}{m_2} \tag{2.11}$$

Following this, for equal molecules, C-H bonds will have higher stretching frequencies than C-C bonds, which again have higher frequencies than C-halogen bonds and O-H have higher frequencies than O-D. Since k increases with increasing bond order, the relative stretching frequencies of carbon-carbon bonds have the order C≡C > C=C > C−C. These generalisations and Equations 2.10 and 2.11 show how different chemical groups can be separated and shown in an IR spectrum.

IR covers the region from approximately 14000 cm^{-1} to 20 cm^{-1}, which can be divided into three; the far-infrared region (400-20 cm^{-1}), the mid-infrared region (4000-

400 cm^{-1}) and the near-infrared region (14000-4000 cm^{-1}).

2.3.1 Fourier transform infrared (FTIR) spectroscopy

FTIR covers the mid-infrared region (4000-400 cm^{-1}). In a FTIR spectrometer a source of infrared light, emitting radiation throughout the whole of the selected frequency range, is divided into two beams of equal intensity. One or both of the beams are then passed through a sample. If both are passed through, one of them is made to traverse a longer path. The two beams are then combined, which produces an interference pattern that is the sum of all the interference patterns for each of the wavelengths in the beam [71]. The differences in the two paths are changed systematically so that the interference patterns change, producing a detectable signal varying with optical path differences which are modified by the selective absorption of some frequencies by the sample. The sum of the interference pattern in the time domain is known as an interferogram, containing information about all the frequencies absorbed by the sample. The interferogram is then fourier transformed into a spectrum with absorption plotted against wavenumber. Figure 2.3 shows a schematic illustration of an FTIR spectrometer.

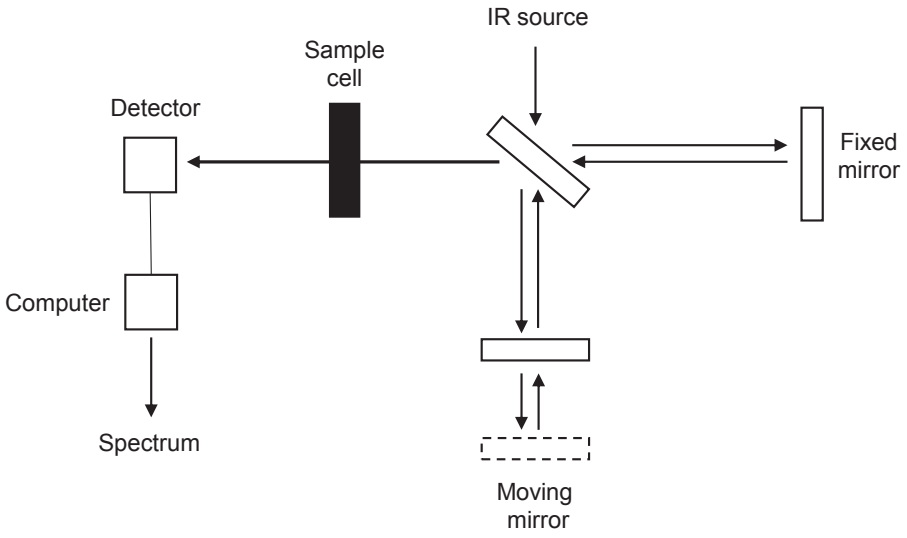


Figure 2.3: Schematic illustration of a FTIR spectrometer with the IR source. How the radiation is passed through the sample using the mirrors creating the interference pattern is shown.

In FTIR spectra of crude oils, the absorption bands from C-H bonds from groups containing aromatics, oxygen, sulphur and nitrogen usually dominate the spectra

[75].

2.3.2 Near-infrared (NIR) spectroscopy

NIR covers the near-infrared region (14000-4000 cm^{-1}). NIR can typically penetrate deeper into a sample than FTIR, but the absorption bands are typically 10-100 times weaker, resulting in a loss in sensitivity. NIR is therefore mainly suitable for quantitative analysis and has more limitations when it comes to identification of chemical groups [76]. Most NIR spectrometers record all wavelengths simultaneously, thereby removing the need for any moving parts such as the moving mirror in Figure 2.3.

The NIR region is useful for crude oil analysis, as many of the absorption bands observed in this region come from combinations or overtones of carbon-hydrogen stretching vibrations, something that makes NIR suited for analysis of hydrocarbon functional groups [77, 78]. The lower specificity observed in NIR when compared to FTIR, is due to overlapping absorption bands, and lower sensitivity due to large variations in chemical groups causing small spectral changes [79]. However, the advantages of NIR for determining physical and chemical properties of crude oils over conventional analyses is well-established [80].

2.3.3 Preprocessing of the spectra

IR spectra are susceptible to non-linearities introduced by light scattering and baseline shifts [81]. Light scattering occurs when electromagnetic radiation is forced to deviate from its trajectory due to localised non-uniformities. As IR spectroscopy is based on measuring the amount of radiation absorbed by the sample, loss of light due to scattering poses a problem for interpretation of the data, as it is difficult to determine if the radiation power is lost due to chemical absorption or scattering. The term "baseline shifts" refers to an additive effect, i.e. an offset, in the spectra, along the absorption axis. This effect is often an indication of variations due to particle size, differences in density or porosity, or the presence of air bubbles. However, many of these unwanted non-linear effects can be completely removed by proper preprocessing of the spectra. The aim of preprocessing methods can for instance be to improve subsequent exploratory analysis and multivariate modelling, and to force the data to obey Beer-Lamberts law [82]. The simplest transformation is the linearisation of transmittance to absorbance ($\log(1/T)$) according to Beer-Lamberts law, which states that the relationship between the concentration and absorbance of the solution is linear. For a transparent sample containing a number of absorbing chemical constituents (J) obeying Beer-Lamberts law, the theoretical chemical absorbance spectrum for a sample (a_i) measured over a range of wavenumbers ($\tilde{\nu}$)

can thus be assumed to be a linear combination of the absorbance contributions of J

$$a_i = c_{i,1}\mathbf{k}_1^T + \cdots + c_{i,j}\mathbf{k}_j^T + \cdots + c_{i,J}\mathbf{k}_J^T, \tag{2.12}$$

where $c_{i,j}$ is the concentration and $\boldsymbol{k_j}$ is the absorbtivity spectrum of the j-th constituent. To approximate the physical effect of scattering, the measured absorbance spectrum ($\mathbf{a}_\mathrm{i}$) for each sample can be modelled as a scaled version of the ideal spectrum. This is the basis for Multiplicative Scatter Correction (MSC) [83, 84], where undesirable scatter effects are removed before data modelling. In the MSC model, physical and chemical contributions to the measured absorbance spectra are separated in accordance with electromagnetic theory. The absorbance spectra for each sample (A_i) is modelled by including a constant baseline offset and a scaling effect

$$A_i(\tilde{\nu}) = a_i + b_i \cdot A_{ref}(\tilde{\nu}) + \epsilon_i(\tilde{\nu}), \tag{2.13}$$

where a is the constant baseline offset, b is the scaling parameter and A_{ref} is a reference spectrum. The reference spectrum should contain the main chemical features of the absorbance. The mean of all spectra in the data set is often used as the reference, as IR spectra of a group of samples often have similar shape. The differences between the measured spectra and the reference are contained in the residuals (ϵ), and can be used as a measure of the chemical variations between spectra. All the parameters are estimated by Least Squares (LS). When the baseline offset and scaling parameters are determined for each sample, the corresponding spectra can be corrected according to

$$A_{i,corrected}(\tilde{\nu}) = \frac{A_{ref} - \epsilon_i(\tilde{\nu})}{b_i} \tag{2.14}$$

However, the light scattering effect depends on the wavenumber $\tilde{\nu}$, and therefore a smooth polynomial wavenumber dependency should be considered. Accordingly, MSC was expanded into Extended Multiplicative Signal Correction (EMSC) [85, 86], which includes a second order polynomial fitting to the reference spectrum and fitting of a baseline of the wavenumber axis. This is done by including $d_i\tilde{\nu}^2$ in the model

$$A_i(\tilde{\nu}) = a_i + b_i \cdot A_{ref}(\tilde{\nu}) + c_i\tilde{\nu} + d_i\tilde{\nu}^2 + \epsilon_i(\tilde{\nu}) \tag{2.15}$$

The additional parameters are again estimated by LS, and are simply added to the model before the spectra are corrected by

$$A_{i,corrected}(\tilde{\nu}) = \frac{A_{ref}(\tilde{\nu}) - a_i - c_i\tilde{\nu} - d_i\tilde{\nu}^2}{b} \tag{2.16}$$

EMSC can also be expanded to correct for *a priori* knowledge from the spectra of interest or spectral interference of interest. This is the advantage of EMSC; that any term, both chemical and physical, which requires correction, can be included as shown by

$$A_i(\tilde{\nu}) = a_i + b_i \cdot A_{ref}(\tilde{\nu}) + c_i\tilde{\nu} + d_i\tilde{\nu}^2 + g_{j,i} \cdot B_j + \epsilon_i(\tilde{\nu}) \tag{2.17}$$

Where B_j can be any term, for instance wavenumber related, weights from some regression, concentrations etc. The term is added to the correction as the polynomial term was added in Equation 2.16. This shows that EMSC is a versatile method, efficient for removing unwanted effects in the spectra. However, the parameters for EMSC are determined using LS regression, which means that collinearities in the spectra to be corrected, can pose an issue.

Additionally, inverse versions of MSC and EMSC have been developed, Inverse Scatter Correction (ISC) and Extended Inverse Signal Correction (EISC), which aim to be even more flexible preprocessing techniques [87]. Instead of regressing $A_i(\tilde{\nu})$ on $A_{ref}(\tilde{\nu})$ and then reversing this model in the signal correction step, the inverted ISC and EISC regress $A_{ref}(\tilde{\nu})$ directly on $A_i(\tilde{\nu})$ and use this model directly in the signal correction step [88]

$$A_{ref}(\tilde{\nu}) = a_i + b_i \cdot A_i(\tilde{\nu}) + \epsilon(\tilde{\nu}) \tag{2.18}$$

The parameters are estimated by LS and correction of the spectra is done by

$$A_{i,corrected} = a_i + b_i A_i(\tilde{\nu}) \tag{2.19}$$

In MSC/EMSC the residuals ϵ_j are minimised horizontally, as the noise is modelled

on the individual spectra, while in ISC/EISC, ϵ_j are minimised vertically, as the noise is modelled on the reference spectra [87]. This difference in how the spectra are modelled also makes the ISC/EISC more computationally demanding, as the corrections for each spectrum has to be calculated individually, as opposed to the MSC/EMSC which uses one reference spectrum for all corrections.

A Savitzky–Golay (SG) filter is a spectral derivation technique consisting of a digital filter which can be applied to data with the aim of smoothing it, in order to increase the precision of the data without distorting the signal [89]. In SG, adjacent data sub-sets are fitted to a low degree polynomial. In order to find the derivative at a centre point, a polynomial is fitted in a symmetric window and the parameters for this polynomial are then calculated as the derivative of any order of this function. The value of the derivative is subsequently used as the derivative estimate for this centre point [81]. This operation is then applied to all points in the spectra sequentially. The highest derivative that can be determined is dependent of the degree of polynomials used during the fitting. As SG uses a symmetric window for smoothing, the number of data point on each side of the center has to be the same. Consequently, a number of points at each end of the spectrum are neglected during the preprocessing. This number of points is equal to the number of points used for smoothing minus one. Usually spectra contain enough information so that the loss of these few point are negligible, however, filter methods using asymmetrical windows are also available.

Finding the optimal preprocessing technique for a data set can be difficult, many methods exits and often have several parameters that needs tuning. This section has only introduced some of the most common techniques. Trial and error has traditionally been the method of determining the best preprocessing, however, ensemble preprocessing methods removing the need for manual determination of parameters, are becoming more frequent [90, 91].

2.4 Interpretation of spectra

Interpretation of the spectra is important to understand the chemical composition of a sample. The positions of the peaks corresponding to different chemical constituents are different for the various spectroscopic techniques, and in this section an overview of how to interpret spectra from FTIR, NIR and FT-ICR MS is presented.

2.4.1 FTIR

In IR, the position of a peak in the spectrum is dependent on the chemical shift of the molecule. A complex molecule has both vibrational modes involving the whole molecule, and localised vibrations of the individual bonds. The localised vibrations are useful for identification of functional groups, in particular the stretching vibrations for O-H and N-H single bonds, double and triple bonds, and in FTIR most of them occur above 1500 cm^{-1}. The remaining single bonds have absorption bands at frequencies below 1500 cm^{-1}, often containing a large number of peaks. The positions of these peaks are characteristic for the molecule, and the composition of peaks in this area can therefore reveal the chemical structure of the sample. This means that the region above 1500 cm^{-1} shows absorption bands that can be related to several functional groups, while the region below gives the characteristic fingerprints of the molecules, and is therefore called the fingerprint region.

The stretching vibrations of single bonds to hydrogen show absorption in the higher frequencies in the spectra due to the low mass of hydrogen. This applies to C-H, O-H and N-H, starting at 2700 cm^{-1} for C-H and increasing in the order above, to approximately 3500 cm^{-1}. However, the C-H peak usually does not reveal important information, as most organic compounds contain C-H bonds. Triple bonded molecules, C≡C and C≡N, usually absorb in the area between 2260-2100 cm^{-1}. Finally, double bonded molecules, C=C, C=O and C=N, absorb in the area between 1800-1650 cm^{-1}.

The absorption bands described above are the vibrations of individual bonds. Many vibrations also exist as coupled vibrations of two or more components in the molecule. Coupled stretching can be divided into asymmetric and symmetric stretching based on the bending modes, i.e. the interaction of the atoms in the compounds. Coupled stretching, both asymmetric and symmetric, can be found in many groups such as primary amines, caboxylic anhydrides, carboxylate ions and nitro groups, all of which have two equal bonds close together. Aromatic rings also have coupled vibrations, and can be identified by two or three bands around 1600-1500 cm^{-1}, corresponding to most six-membered aromatic rings. They also have bands in the fingerprint region at 1225-950 cm^{-1} and at 900 cm^{-1}, and this band was previously used to identify substitution patterns [71].

From the described vibrations it is possible to interpret the peaks in the spectra. Figure 2.4 shows the absorption bands observed in FTIR.

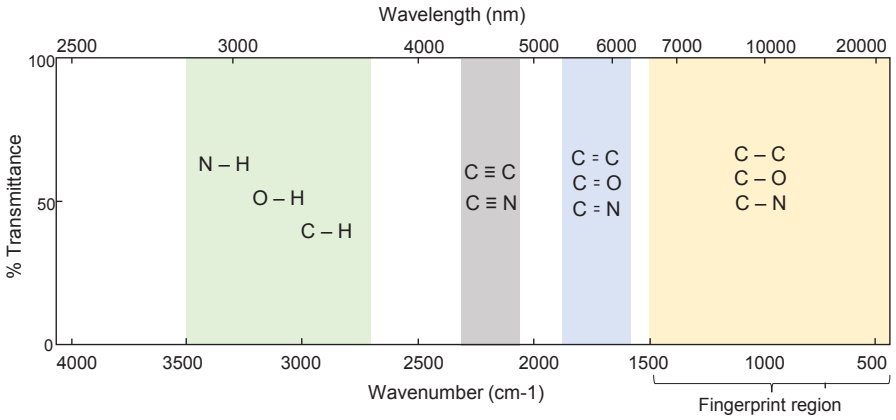


Figure 2.4: Absorption bands in FTIR

2.4.2 NIR

In NIR, the interpretation is not as simple due to broad, overlapping and non-specific absorption bands. The NIR bands are not related to particular molecules, but rather represent molecular bonds, mainly the C-H, O-H and N-H bonds. Additionally, in the NIR range, the overtones occurring when the molecule transitions from the ground state to an exited state are visible as spectral bands. These are referred to as the overtone bands, where two of the exited states have bands occurring in the NIR region. However, the overtone bands are lower than the fundamental bands from the molecule in its ground state.

In a NIR spectrum, the single bond C-H stretching and bending vibrations for CH_2 and CH_3 have absorption bands between 4500-4000 cm^{-1}. The first overtone for the C-H stretching is observed between 6050-5500 cm^{-1}, and a weak absorption is centred at 7000 cm^{-1} for the combination of bending and stretching. The second overtone bonds for C-H stretching are centred at 8000 cm^{-1}. The fundamental vibrations of unsaturated groups absorb in weak bands between 4750-4500 cm^{-1}, which include double and triple bonds. For N-H stretching vibrations the fundamental bands are observed between 4700-4545 cm^{-1}, while the first overtone bonds absorb between 7015-6625 cm^{-1}. Additionally, a second overtone can be observed just below 10 000 cm^{-1}.

Baseline shifts are often observed in NIR. For crude oil analysed using NIR, it is common to see baseline offsets and a slope between 9000-6500 cm^{-1}, which are

characteristic for asphaltene containing samples. These effects correspond to the tail of the absorption bands in the visible region due to transitions of electrons, caused by $\pi - \pi^*$ and n-π^* transition of asphaltene molecules [92]. The offset can also be attributed to light scattering due to asphaltene aggregation [77], and the spectral distortion is increasing along with the asphaltene content.

2.4.3 FT-ICR MS

For MS, the positions of the peaks in the spectra are based on the m/z-ratios of the molecules, and when the formed ion has a charge of 1, the m/z-ratio is directly related to the mass of the molecule.

To be able to interpret peaks in a MS spectrum accurately, their molecular formulas have to be identified using a suitable spectra processing software. Although, there are some chemical groups that have been determined to appear in certain masses. For instance, asphaltenes have an average mass of ∼750 Da, meaning that asphaltenic molecules can be identified by peaks around m/z 750. Another chemical group is naphthenic acids, with an average mass between 300-500 Da [34]. The ARNs have high molecular masses of 1200-1250 Da due to the many carboxylic acid groups. DBE is important for interpretation, since it reveals the number of double bonds or rings in the molecule.

2.5 Choice of spectroscopic method

The various spectroscopic methods are apt at measuring different properties of a sample. For instance, the two ionisation techniques, ESI and APPI, in combination with FT-ICR MS, measure disparate molecules in the sample. ESI is most efficient with polar molecules mainly consisting of heteroatom-containing components [70], while APPI characterises more of the non-polar molecules, which constitutes approximately 90 % of the crude oil components [93]. APPI can also positively charge cycloalkanes and aromatic species to aid their detection. Additionally, the mode of the ionisation has an effect; in postive mode, ESI is able to detect asphaltenes, hydroxyl groups and amines/amide bonds while in negative mode, ESI detects oxygen species containing acidic or carboxylic groups . For APPI, in positive mode NO_x and other nitrogen species are prominent, while negative mode ionises oxygenated groups such as acids, but usually with higher degrees of unsaturation (higher DBEs) than in ESI [94].

For the IR spectroscopic methods, FTIR spectra are usually dominated by the ab-

sorption bands from C-H bonds and groups containing aromatics, sulphur, oxygen and nitrogen [75]. In NIR spectra, functional groups such as methylenic, oleifinic and aromatic C-H bonds are usually more prominent, and the bindings involved are C-H, O-H and N-H.

When determining which methods to use for a given sample, several things have to be taken into consideration. Among these are time, complexity of the sample matrix, the compounds of interest and the desired outcome. Mainly, the IR spectrum identifies functional groups, while a mass spectrum gives the molecular formula. Additional analysis techniques not discussed in this work use the ultraviolet spectrum from which conjugated systems can be identified, and nuclear magnetic resonance spectra which identify how the atoms are connected.

2.6 Machine learning

Machine learning is a sub-field of artificial intelligence devoted to building models that leverage data to improve the performance on a defined set of tasks [95, 96]. The main aim of a machine learning method is to learn information from a data set and then perform accurately on new unseen data. A typical pipeline is shown in Figure 2.5, and consists of collecting and preprocessing of data, training and testing of the model and finally, deploying the model through prediction on new data.

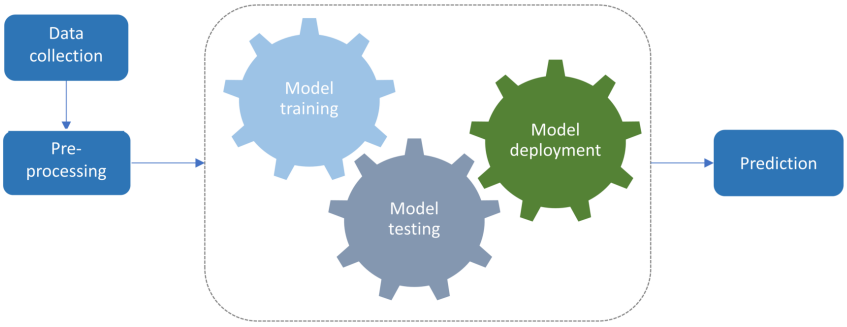


Figure 2.5: Schematic illustration of a machine learning pipeline, with data collection, preprocessing, model training, testing, deployment and prediction. Reprinted from Gjelsvik *et al.* [56].

Machine learning can be divided into two main groups, supervised and unsupervised learning. Unsupervised learning refers to methods which attempt to learn patterns from unlabelled data. This is implemented by the model mimicking the data and then using the errors in the estimated output to correct its weights and bias. The

two most common unsupervised methods are Principal Component Analysis (PCA) and clustering. Supervised learning, on the other hand, deals with labelled data, and the methods aim to map the input variables to the output labels. New samples are predicted based on the input-output information learned from the training data. Supervised learning can again be separated into two categories based on the desired response. When the response is continuous, regression analysis is used, while when the response is a discrete class label, classification is used. A wide range of supervised methods exist based on various algorithms which can be optimised for different types of data. Additionally, some algorithms can be used for both classification and regression tasks with only minor modifications. The following sections will go through some commonly used machine learning and multivariate methods for identification of connections in the data and prediction of desired properties.

2.7 Unsupervised learning

2.7.1 Clustering

Clustering is the exercise of grouping samples in such a way that the samples in the same group, or cluster, are more similar to each other than they are to the samples in another group. It is commonly used for pattern recognition and many different algorithms exist. Clustering is often based on the samples´ proximity to each other, and the main idea is that samples are more related to the nearby ones than samples far away in the sample space, and a cluster can be described as the maximum distance needed to connect parts of the cluster. Figure 2.6 illustrates how clustering methods are applied to find groups of similar samples in the data.

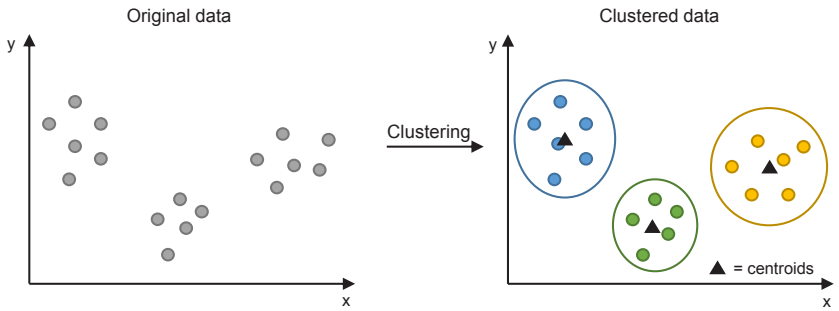


Figure 2.6: Illustration of clustering of a data set with three groups of similar samples.

The different clustering methods are separated based on the algorithms they use to determine cluster distributions. Among the most common methods are con-

nectivity based models, most often referred to as hierarchical clustering [97]. In hierarchical clustering, hierarchies of clusters are built based on either an agglomerative approach, where the samples start as single elements and are merged into clusters based on their distances to each other, or a divisive approach, starting with the full data set as one single cluster and dividing it into different clusters. The merges and splits are determined in a greedy manner, i.e. the locally optimal choice is determined at each point, and the results from the clustering are presented in a dendrogram, with different cluster distributions appearing at different distance thresholds (thresholds for how close samples need to be for them to be defined as members of the same cluster). Thereby a hierarchy of cluster solutions is created. Any valid measure of distance can be used as the metric for determining dissimilarities between samples, and where clusters should be agglomerated or divided. Additionally, a linkage criterion is used to specify the dissimilarities of sample sets as a function of the pairwise distances between observations in the sets. The various linkages and distance metrics have large impacts on the results of the clustering, where the distance determines which samples are most similar, while the linkage criterion determines the shape of the clusters. The most common linkage criteria are single-linkage, which uses the shortest distance between a pair of observations in two clusters, complete linkage, which uses the distance between the farthest pair of observations in two clusters, and average linkage, which adds up the distances between each pair of observations and divides by the number of pairs for an average inter-cluster distance.

Another commonly used approach to clustering is centroid-based clustering, such as K-Means clustering, where each cluster is represented by a central vector. When k clusters are defined, K-Means clustering partitions the observations into k clusters where each observation belongs to the cluster that has the nearest mean [98, 99]. This mean is often referred to as the cluster centroid. K-Means clustering minimises the variance within clusters, determined by the squared distances. The clustering distributions are determined by alternating between two steps. First each observation is assigned to the cluster with the nearest centroid, and then the centroids for each cluster are recalculated as the mean over all cluster members. The algorithm has converged when samples no longer are assigned to new clusters. The centroids are often initialised at random. K-means uses hard clustering, where the samples belong to one cluster, but centroid-based methods can also use soft clustering, where samples receive a likelyhood of belonging to each cluster, as in Fuzzy-C-Means (FCM) clustering [100, 101].

In density-based clustering, the clusters are defined as areas of higher data density

than the rest of the data set. Samples in sparse areas are usually considered as noise or border points, and these areas are used to separate the clusters. The most popular density-based method is Density-based spatial clustering of applications with noise (DBSCAN), where samples with many neighbours (high density) are grouped together and samples alone in low-density areas are marked as outliers [102].

Another clustering approach is Spectral clustering (SPC) [103] which uses the spectrum (eigenvalues) of the similarity matrix of the data to reduce dimensionality, so that the clustering can be done in fewer dimensions. The similarity matrix consists of the relative similarities for each pair of points in the data set found through a kernel with a distance measure. SPC is useful when the structure of the clusters is non-convex, when the centra and spreads of the clusters give a poor description of the properties of the clusters.

2.7.2 Principal Component Analysis (PCA)

PCA is a commonly used dimensionality reduction and visualisation technique where large and high-dimensional data are decomposed into fewer dimensions, while preserving the maximum amount of information [104]. The dimensionality reduction is done by projecting each data point on to the first few Principal Components (PCs). The first PC is the linear combination of the original variables that explains the largest part of the variance in the data set. The second component explains the second largest part of the variance, and is orthogonal to the first PC, and so on until all the variance in the data is explained. PCA can be used prior to other data analysis methods in order to increase accuracy, overview and interpretation. The data set ($\boldsymbol{X}$) is decomposed into a subspace of latent variables representing the main features of variance as shown by

$$\boldsymbol{X} = \bar{\boldsymbol{x}} + \boldsymbol{T}_A \boldsymbol{P}_A^T + \boldsymbol{E}_A, \tag{2.20}$$

where $\boldsymbol{P}_A$ are the loadings and orthonormal eigenvectors of $(\boldsymbol{X} - \bar{\boldsymbol{x}})^T(\boldsymbol{X} - \bar{\boldsymbol{x}})$ minimising the covariance between the $\boldsymbol{X}$-variables after $\boldsymbol{A}$ Principal Components (PCs). The scores ($\boldsymbol{T}_\mathrm{A}$) are orthogonal and calculated by

$$\boldsymbol{T}_A = (\boldsymbol{X} - \boldsymbol{x})\boldsymbol{P}_A \tag{2.21}$$

The error term in Equation 2.20 ($\boldsymbol{E}_A$) is often referred to as the residuals and

contains all the variance that cannot be explained by the model. The residuals are determined from

$$\boldsymbol{E}_A = \boldsymbol{X} - \bar{\boldsymbol{x}} - \boldsymbol{T}_A \boldsymbol{P}_A^T \tag{2.22}$$

Figure 2.7 shows how the two PCs for a data set with two variables will be positioned.

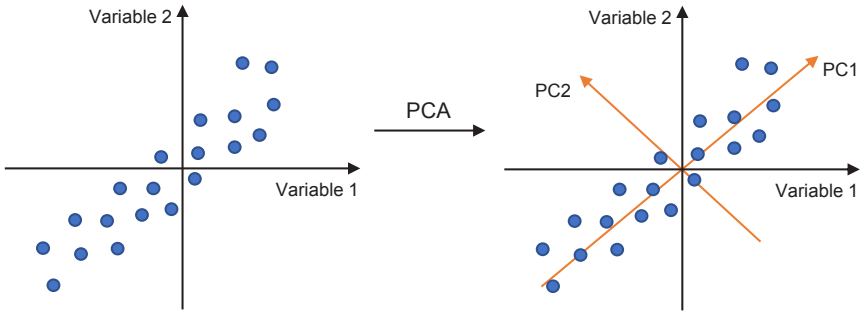


Figure 2.7: Illustration of the orthogonality of the two PCs for a data set with two variables.

PCA is commonly used for pattern recognition, however it is not optimised for separating classes, as it does not give a measure of distance. For such tasks, clustering is the appropriate choice.

2.8 Supervised learning

2.8.1 Ordinary Least Squares (OLS) regression

OLS is a regression method for estimating the unknown parameters in a linear regression model. OLS minimises the sum of squares of the differences between the observed values and the values predicted by the linear function of the independent variables as shown by

$$\boldsymbol{y} = \boldsymbol{X}\beta + \epsilon \tag{2.23}$$

In OLS, the regression coefficients ($\hat{\beta}$) are estimated from

$$\hat{\beta} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y} \tag{2.24}$$

A major drawback with OLS regression is that the matrix inversion used in the calculation of the regression coefficients requires the regressors ($\boldsymbol{X}$) to be linearly independent or uncorrelated. OLS also requires that the number of samples is larger than the number of variables, which is most often not the case when analysing data from e.g. FT-ICR MS. This makes OLS regression unsuitable for many data analysis problems. Two commonly used strategies to overcome this problem are using latent variables which represent linearly independent phenomena and regularisation.

2.8.2 Partial Least Squares Regression (PLSR)

In PLSR [105, 106] the variables are reduced to a smaller set of uncorrelated components, similarly to in PCA, and Least Squares is performed on the reduced data. This is done by decomposing large data sets into a subspace of latent variables (scores and loadings) representing the main features of co-variance between $\boldsymbol{X}$ (regressors) and $\boldsymbol{Y}$ (response). PLSR has the advantage over OLS that it can handle multivariate and multicollinear data in both $\boldsymbol{X}$ and $\boldsymbol{Y}$. The decomposition of $\boldsymbol{X}$ and $\boldsymbol{Y}$ is done simultaneously and iteratively, taking co-linearities in $\boldsymbol{Y}$ into account. For $\boldsymbol{X}$ the decomposition is shown in Equation 2.25 and for $\boldsymbol{Y}$ in Equation 2.26.

$$\boldsymbol{X} = \bar{\boldsymbol{x}} + \boldsymbol{T}_A\boldsymbol{P}_A^T + \boldsymbol{E}_A \tag{2.25}$$

$$\boldsymbol{Y} = \bar{\boldsymbol{y}} + \boldsymbol{U}_A\boldsymbol{Q}_A^T + \boldsymbol{F}_A \tag{2.26}$$

Here $\boldsymbol{A}$ denotes the number of PLS components used and $\boldsymbol{E}_\mathrm{A}$ and $\boldsymbol{F}_\mathrm{A}$ are the error terms using $\boldsymbol{A}$ components. The loading weight matrix ($\boldsymbol{W}_\mathrm{A}$) maximises the covariance between $\boldsymbol{X}$ and $\boldsymbol{Y}$ by maximising the covariance between the scores, $\boldsymbol{T}$ and $\boldsymbol{U}$, with $\boldsymbol{A}$ components. The scores are orthogonal as shown by

$$\boldsymbol{T}_A = \boldsymbol{X}\,\boldsymbol{W}_A(\boldsymbol{P}_A^T\boldsymbol{W}_A)^{-1} \tag{2.27}$$

The loadings for $\boldsymbol{X}$ ($\boldsymbol{P}_\mathrm{A}$) are calculated by Equation 2.28, while the loadings for $\boldsymbol{Y}$ ($\boldsymbol{Q}_\mathrm{A}$) are calculated by Equation 2.29. The direction for the first component ($\boldsymbol{W}_1$) is obtained by maximising the covariance, and the scores along this axis is calculated

by Equation 2.27, before $\boldsymbol{X}$ is regressed onto the estimated scores in order to obtain the loadings. The product of the scores and loadings is then subtracted from $\boldsymbol{X}$. The same procedure is carried out for $\boldsymbol{Y}$ and the $\boldsymbol{Y}$-loadings ($\boldsymbol{Q}_1$). The direction of the second component is found in the same way, only using the residuals after subtraction of the first component instead of the original data. This process continues until the desired number of components ($\boldsymbol{A}$) are extracted, or until the number of components reaches either the number of samples or the number of variables.

$$\boldsymbol{P}_A^T = (\boldsymbol{T}_A^T \boldsymbol{T}_A^T)^{-1} \boldsymbol{T}_A^T (\boldsymbol{X} - \bar{\boldsymbol{x}}) \tag{2.28}$$

$$\boldsymbol{Q}_A^T = (\boldsymbol{T}_A^T \boldsymbol{T}_A^T)^{-1} \boldsymbol{T}_A^T (\boldsymbol{Y} - \bar{\boldsymbol{y}}) \tag{2.29}$$

The error term for $\boldsymbol{X}$ ($\boldsymbol{E}_A$) is calculated as for PCA in Equation 2.22 and the error term for $\boldsymbol{Y}$ ($\boldsymbol{F}_A$) is calculated by

$$\boldsymbol{F}_A = \boldsymbol{Y} - \bar{\boldsymbol{y}} - \boldsymbol{T}_A \boldsymbol{Q}_A^T \tag{2.30}$$

The regression coefficients ($\boldsymbol{B}_\mathrm{A}$), which are measures of the impact of variations in the various regressors on the respective response variables, are calculated by

$$\boldsymbol{B}_A = \boldsymbol{W}_A (\boldsymbol{P}_A^T \boldsymbol{W}_A)^{-1} \boldsymbol{Q}_A^T \tag{2.31}$$

Prediction of $\boldsymbol{Y}$ for a new sample ($\boldsymbol{X}_\mathrm{new}$) is then obtained by Equation 2.32, where $\boldsymbol{b}_0$ is the intercept.

$$\boldsymbol{Y}_{pred} = \boldsymbol{b}_0 + \boldsymbol{X}_{new} \boldsymbol{B}_A \tag{2.32}$$

PLSR is particularly well suited for data sets where the number of variables substantially exceeds the number of samples and when there is multicollinearity among the variables.

2.8.3 Hierarchical cluster-based regression

In local modelling, a set of models are built based on the data where each model represents a sub-space of the problem space. The sub-space can for instance be

a cluster or some other grouping of the data, and the problem space is where the analysis is performed, possibly based on a set of rules or a set of regression models.

Hierarchical Cluster-based Partial Least Squares Regression (HC-PLSR) [61] is a locally linear extension of PLSR where samples are separated into clusters and PLSR models are created for each cluster. HC-PLSR utilises the abilities of clustering methods to collect similar samples into clusters to achieve models with increased prediction abilities. The advantage of PLSR is the efficient and fast ways it finds latent variables in the data. However, PLSR is a linear model, and can therefore struggle to perform well on data with non-linear interrelationships. Although HC-PLSR allows for modelling of non-linearities between the clusters, the data is still required to be locally linear for the local PLSR models to perform well. Additionally, the input space the data lies in can exist in different planes, be high-dimensional, low-dimensional, linear, non-linear etc. Simple non-linearities can be handled by the addition of polynomial terms in the regressor matrix, or by the local modelling, but some types of non-linearities cannot be modelled by PLSR. In such cases, there is a need for methods able to handle more complex structures. Neural networks and support vector machines are powerful for modelling large non-linear data sets, but neural networks often require deep and complex networks to achieve adequate prediction. Neural networks are described in detail in section 2.10. Deep networks contains numerous parameters and need large amounts of computational power to converge. For local modelling, smaller networks can be implemented without loosing predictability, and in most cases both predictability and interpretability are improved. Thus, there are many advantages of expanding local modelling into deep learning to handle non-linear data.

In HC-PLSR, a global PLSR model is first fitted to the data set with the optimal number of components determined by cross validation. The X scores from the global PLSR model are then clustered using FCM, and a local model is built within each cluster, with its parameters determined by cross validation. New samples are projected into the global model and classified into one of the clusters based on their X scores, and finally predicted using the selected local model. The local modelling procedure is shown in Figure 2.8.

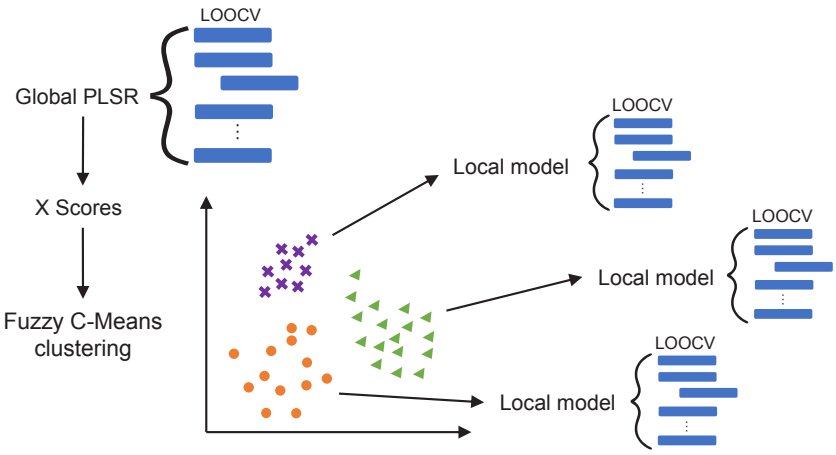


Figure 2.8: Illustration of the local modelling procedure.

Real world data have unknown underlying structures, and the aim of machine learning models is often to identify and model these structures. Local modelling based on clustering merits from dividing the observations into groups without using any previously known information, and then building models based on similarities in the data. Thereby, these underlying structures are somewhat directly modelled, and can be interpreted from the differences between the local models.

2.8.4 Support Vector Machines (SVMs)

SVMs [107] are supervised learning methods that analyse data for classification or regression analysis. SVMs are well suited for learning tasks where the number of variables is large compared to the number of observations in the training set. SVMs are automatically regularised using Tikhonov regularisation, also named Ridge regression, which is explained in section 2.8.7.

For classification, SVMs construct a hyperplane or a set of hyperplanes in a high dimensional space to separate the observations into two groups [108]. SVMs map the training data to points in space where the gap between the two groups is maximised. New test samples are then mapped into the same space and predicted into a class based on their positions in relation to this gap. The gap is quantified by the hyperplane, and the goal is to find the hyperplane that has the largest distance (margin) to the nearest data point belonging to any of the two classes. The margin is defined as the distance between the separating hyperplane (decision boundary)

and the training samples that are closest to this hyperplane. Data points that lie on the margin are known as support vector points, and the solution is represented as a linear combination of only these points. Decision boundaries with large margins tend to have a lower generalisation error, while decision boundaries with small margins are more prone to overfitting. SVMs can also be used for classification in multi-class (more than two classes) problems by distinguishing between one of the labels and the rest (one-versus-all) or between every pairs of classes (one-versus-one). For one-versus-all, classification of new samples is done by a winner-takes-all strategy, where the classifier with the highest-output function assigns the class. For one-versus-one, classification follows a max-wins strategy, where every classifies assigns the new sample to one of the two classes, and the class with the most assigned samples determines the classification.

SVMs can be applied to nonlinear classification problems by using the so-called kernel trick, where the original space is mapped into a much higher-dimensional space where the observations can be more easily separated. To achieve this, a mapping function φ is used, as shown in Figure 2.9. The hyperplanes in the higher-dimensional space are defined as the set of points whose dot product with a vector in that space is constant.

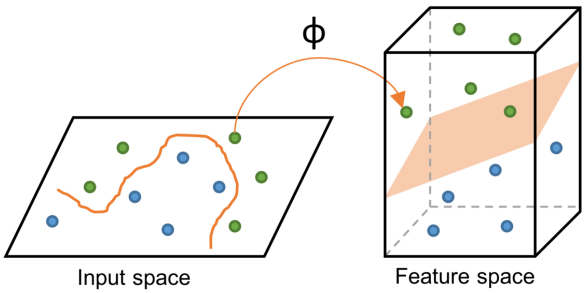


Figure 2.9: The kernel trick to handle non-linear problems. Reprinted from Gjelsvik *et al.* [56].

In Support Vector Regression (SVR), the hyperplane is the line that is used to predict the continuous output, shown in Figure 2.10. SVR basically considers the points that are within the decision boundary lines and the regression line is then the hyperplane that has a maximum number of points.

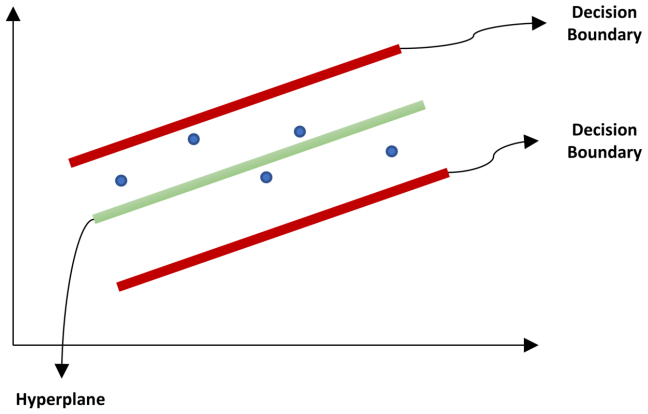


Figure 2.10: Illustration of the hyperplane and decision boundaries in SVR. Reprinted from Gjelsvik *et al.* [56].

2.8.5 Decision Trees (DTs)

DTs [109, 110] are attractive models when interpretability is important, and consist of a tree root, internal nodes, branches and leaf nodes. DTs ask a series of questions, and generate decision rules based on these. The model seeks to find the smallest set of rules that is consistent with the training data. In general, the rules have the form: *if* $condition_1$ *and* $condition_2$ *and* $condition_3$ *then outcome.*

The rules are chosen to divide observations into segments that have the largest difference with respect to the target variable. Thus the rule selects both the variable and the best breaking point to separate the resulting subgroups maximally. The breaking points of variables are found using significance testing (F- or Chi-square with Bonferroni corrections) or reduction in variance criteria. To avoid overfitting, one often has to prune the tree by setting a limit for the maximal depth of the tree. A leaf can no longer be split when there are too few observations, the maximum depth (hierarchy of the tree) has been reached, or no significant split can be identified. It is assumed that observations belonging to different classes have different values in at least one of their variables. DTs are usually univariate, since they use splits based on a single feature at each internal node, but methods are available for constructing multivariate trees [111].

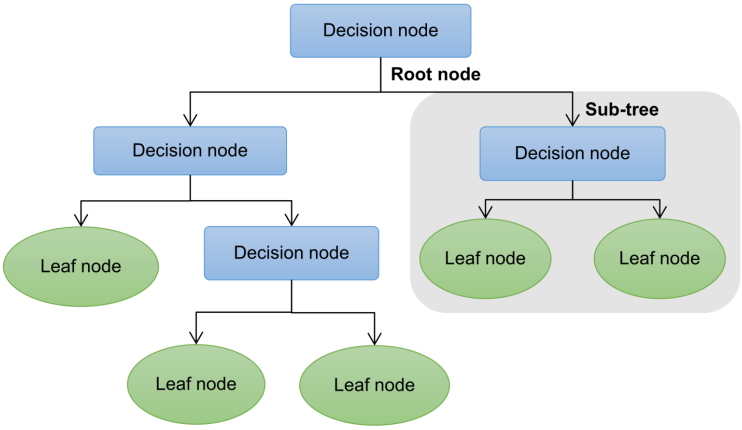


Figure 2.11: Illustration of decision trees. Reprinted from Gjelsvik *et al.* [56].

To improve the prediction of the DT, a boosting method can be applied. Boosting is an ensemble method for improving predictions of a weak learning algorithm [112]. The weak learners are trained sequentially, trying to improve upon its predecessor. When boosting is applied to a tree, each tree is dependent on prior trees and the algorithm learns by fitting the residuals from the prior trees. One example of a boosting method is XGBoost (eXtreme Gradient Boosting) [113]. In XGBoost, trees are built at every iteration, always minimising the prediction error of the classifier, while introducing a penalty function to utilise the computational power more efficiently.

2.8.6 Random Forest (RF)

In DTs, the initial selected split affects the optimality of variables considered for subsequent splits. Ensemble tree models grow trees with varying initial splits, and use either a voting or the average of the predictions for each new data point across all trees. The ensemble is less prone to overfitting and other problems than individual DTs, and generally performs better. RF [114, 115, 116] is an example of such an ensemble tree method. For RF, each tree is based on a random subset of the data and variables, selected by bootstrapping. Bootstrapping is a sampling process where random samples are drawn from the original samples a large number of times, with replacement. Each tree makes a prediction, and for classification problems, the final prediction is determined by majority voting, where the vote distribution can be used to develop a nonparametric probabilistic predictive model. For regression problems, the final prediction is determined by averaging of the individual predictions. When

used for variable selection, the change in prediction accuracy when the values of a feature are randomly permuted among observations (using permutation feature importance explained in section 2.11) gives estimates of the importance of each feature.

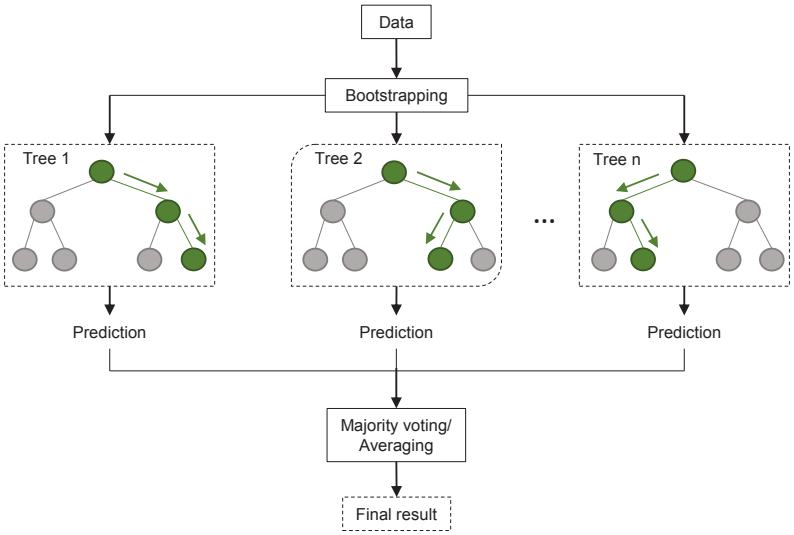


Figure 2.12: Illustration of RF. The variables are selected by bootstrapping and each tree makes a prediction. The final result is determined by majority voting for classification and averaging for regression.

2.8.7 Regularisation-based methods

The aim when using regularisation is to find a balance between a too simple and a too complicated model, the optimal solution shown in Figure 2.13. The model is simplified by adding constraints, i.e. regularisation terms, that shrink the coefficient estimates and minimise the adjusted loss function. Regularisation is often used to reduce overfitting and in highly complex modelling problems. The efficiency of these methods comes from the reduction of the generalisation error of the models, making them more general for prediction of new samples. The shrinkage of the coefficients also makes regularisation very useful for feature selection purposes. This section will present the most commonly used regularisation-based methods.

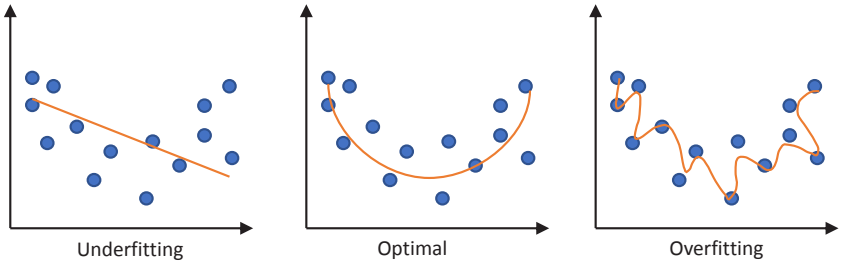


Figure 2.13: Illustration of underfitting and overfitting. Regularisation strives to find the optimal solution as a balance between underfitting and overfitting.

2.8.8 Ridge Regression

In Ridge regression [117], the sum of the squares of the regression coefficients (β) is forced to be less than a fixed value, which shrinks the size of the coefficients. Ridge regression is often used when the variables are highly correlated. In these cases the ridge coefficient estimates are often more precise than those from OLS, as the variance and mean square estimators are often smaller. The aim of OLS is to minimise

$$RSS_{OLS} = \sum_{i=1}^{n} \left(\boldsymbol{y}_i - \beta_0 - \sum_{j=1}^{p} \beta_j \boldsymbol{x}_{ij} \right)^2, \tag{2.33}$$

while in Ridge regression a regularisation term is added, resulting in the minimisation of

$$RSS_{Ridge} = \sum_{i=1}^{n} \left(\boldsymbol{y}_i - \beta_0 - \sum_{j=1}^{p} \beta_j \boldsymbol{x}_{ij} \right)^2 + \lambda \sum_{i=1}^{p} \beta_j^2, \tag{2.34}$$

where $\lambda \geq 0$ is a penalty term which is often found by cross-validation. This gives Equation 2.35 and 2.36.

$$B_{OLS} = (\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{Y} \tag{2.35}$$

$$B_{Ridge} = (\boldsymbol{X}^T \boldsymbol{X} + \lambda I)^{-1} \boldsymbol{X}^T \boldsymbol{Y} \tag{2.36}$$

Hence, Ridge regression handles multicollinearity in the regressor ($\boldsymbol{X}$) matrix, while OLS regression does not. Ridge is also referred to as L2-regularisation or Tikhonov regularisation.

2.8.9 LASSO

In LASSO (least absolute shrinkage and selection operator) [118], the estimates of the regression coefficients are obtained using L1-constrained least squares. This forces the sum of the absolute values of the regression coefficients to be less than a fixed value, which forces certain coefficients to be set to zero. LASSO is a feature selection method, since variables having zero regression coefficients are omitted from the model. Both variable selection and regularisation are used to select a reduced set of the known covariates, enhancing the prediction accuracy and interpretability of the model. In LASSO the regression solution is found by minimising

$$RSS_{LASSO} = \sum_{i=1}^{n} \left(\boldsymbol{y}_i - \beta_0 - \sum_{j=1}^{p} \beta_j \boldsymbol{x}_{ij} \right)^2 + \lambda \sum_{i=1}^{p} \beta_j \tag{2.37}$$

2.8.10 Elastic Net

Elastic net [119] combines the L1 and L2 penalties of the Ridge and LASSO methods linearly. For large data sets where the number of variables is much larger then the number of samples, LASSO exhibits some limitations. The number of selected variables saturates around the number of samples, and LASSO often selects only one variable from groups of highly correlated variables, ignoring the remainder. In Elastic net, these limitations are surpassed by the addition of the quadratic part (β_j^2) from Ridge in the penalty term. This penalty term makes the loss function more convex, therefore giving it a unique minimum. The regression solution for Elastic net is found by minimising

$$RSS_{EN} = \sum_{i=1}^{n} \left(\boldsymbol{y}_i - \beta_0 \sum_{j=1}^{p} \beta_j \boldsymbol{x}_{ij} \right)^2 + \lambda_1 \sum_{i=1}^{p} \beta_j^2 + \lambda_2 \sum_{i=1}^{p} \beta_j \tag{2.38}$$

In Elastic net, highly correlated variables will tend to have similar regression coefficients, which creates a grouping effect that is desirable in many applications.

2.9 Multiblock

Multiblock analysis is a way of utilising several different data sources to gain a deeper understanding of samples. The aim of multiblock methods is to find complementary information from multiple sources of data to improve the predictive accuracy or interpretability of the models [120]. Multi-block data can be multiple analyses of the same samples, for instance with different analytical tools to achieve a better understanding of physical or chemical properties. Figure 2.14 shows how the data can be structured in a multiblock analysis.

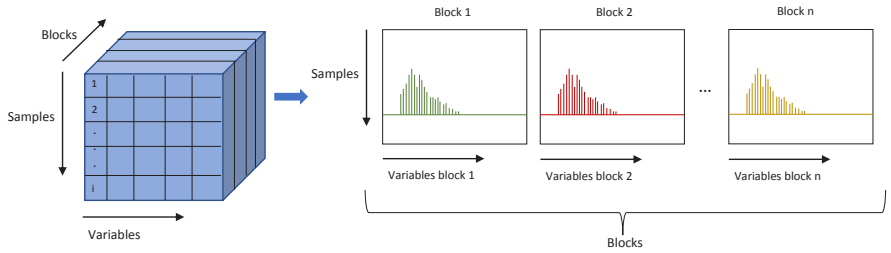


Figure 2.14: Illustration of how the data can be structured for multiblock analysis, with i samples and n blocks.

Several different techniques exist for combining data for multiblock analysis. The data can for instance be concatenated according to a shared mode, for instance if the same samples are measured using different analytical techniques, sample is the shared mode, or the same variables are measured and then variable is the shared mode. Another way is to analyse the data sequentially, extracting important information from one block before moving to the next block. The difference between the methods lies in how the constraints are applied during the decomposition, leading to different orthogonality properties and thereby different independence of the common and distinctive parts [121]. The idea behind finding common and distinct variation in the blocks of data is to separate and quantify the different sources of variation spread across all blocks [122]. The interpretation of the different sources of variation can then lead to a reconstruction of the system. Common variation can be comprehended as the variation associated between data sets while distinct variation can be regarded as the variation which is unique for each data set.

Many sources of data requires preprocessing in order to achieve optimal analyses as discussed previously in this work. For multiblock analysis, preprocessing of data is divided into inter-block and intra-block preprocessing. It is an important aspect of

analysis, as some multiblock methods tend to favour the block with larger variations, causing model bias [120]. Proper weighting and scaling of the blocks can therefore increase the model interpretation and the predictive accuracy.

2.9.1 Multiblock Partial Least Squares Regression (MB-PLSR)

In MB-PLSR, global scores are extracted by maximising the covariance with the response variables, and the extracted global scores are then used in ordinary least squares regression to obtain the predictive models [123, 124]. The data sets are fused by concatenating the individual blocks, after dividing by the square root of the number of variables in each block ($\sqrt{J_m}$). MB-PLSR with super-score deflation of the response starts with an ordinary PLSR on the concatenated blocks, followed by a block-wise extraction of block-weights, block-scores and block-loadings [125]. The prediction is obtained from the PLSR model on the concatenated blocks along with the super-weights, -scores, -loadings and $\boldsymbol{Y}$-scores and -loadings. The block-loading weights ($\boldsymbol{w}_{\mathrm{m}}$) are then obtained from the original block data ($\boldsymbol{X}_{\mathrm{m}}$) by

$$\boldsymbol{w}_m = \frac{\boldsymbol{X}_m^T \boldsymbol{u}}{(\boldsymbol{u}^T \boldsymbol{u})}, \tag{2.39}$$

where $\boldsymbol{u}$ are the $\boldsymbol{y}$ scores. The block-scores ($\boldsymbol{t}_{\mathrm{m}}$) are obtained from

$$\boldsymbol{t}_m = \frac{X_m}{\sqrt{J_m}} \boldsymbol{w}_m^*, \tag{2.40}$$

where $\boldsymbol{J}_{\mathrm{m}}$ are the variables for block $\boldsymbol{m}$ and $\boldsymbol{w}^*{}_{\mathrm{m}}$ are the normalised weights ($\boldsymbol{w}_m^* = \boldsymbol{w}_m / \parallel \boldsymbol{w}_m \parallel$). Finally, the block-loadings ($\boldsymbol{p}_{\mathrm{m}}$) are obtained by Equation 2.41.

$$\boldsymbol{p}_m = \frac{X_m^T}{\sqrt{J_m}} \cdot \frac{\boldsymbol{t}_m}{\boldsymbol{t}_m^T \boldsymbol{t}_m} \tag{2.41}$$

Figure 2.15 shows a schematic illustration of how the super-weights, -scores, -loadings and block-weights and -scores are calculated in MB-PLSR.

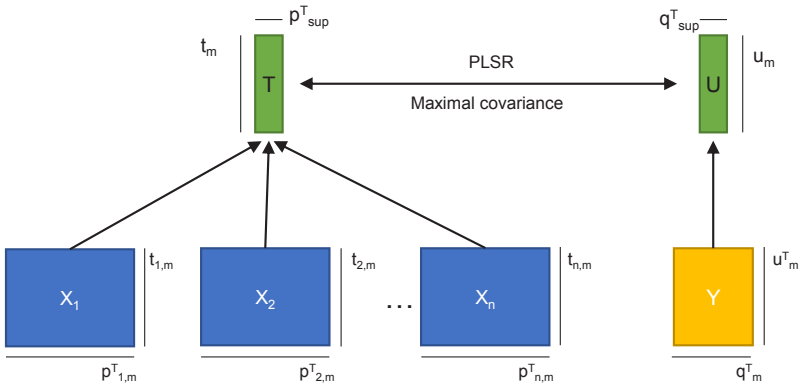


Figure 2.15: Schematic illustration of MB-PLSR, where the data are concatenated by a shared sample mode, and super-scores and -loadings are calculated from the blocks to achieve maximum covariance.

MB-PLSR uses the same number of components for all the blocks. In cases where the dimensionalities of the blocks are very different, the number of components may not be optimal for all the blocks which can complicate the interpretation [126]. However, this also makes for simpler models predicting only on one set of components, making the modelling less susceptible for overfitting.

2.9.2 Sequential Orthogonal Partial Least Squares Regression (SO-PLSR)

In SO-PLSR, the blocks of data are incorporated one at a time to evaluate their incremental contribution by letting the method sequentially search for improvements of predictions using additional and orthogonal information provided by the subsequent blocks [127, 128]. This is done by applying PLSR to the first block, and extracting the scores ($\boldsymbol{T}_1$) and loadings ($\boldsymbol{P}_1$) for block 1, followed by an orthogonalisation of the second block ($\boldsymbol{X}_2$) as shown in Equation 2.42 for $\boldsymbol{X}_2$ and for $\boldsymbol{Y}$ in Equation 2.43.

$$X_{2,orth} = (\boldsymbol{I} - \boldsymbol{T}_1(\boldsymbol{T}_1^T\boldsymbol{T}_1)^{-1}\boldsymbol{T}_1^T)\boldsymbol{X}_2 \tag{2.42}$$

$$\boldsymbol{Y}_{orth} = (\boldsymbol{I} - \boldsymbol{T}_1(\boldsymbol{T}_1^T\boldsymbol{T}_1)^{-1}\boldsymbol{T}_1^T)\boldsymbol{Y} \tag{2.43}$$

In the next step, a new PLSR model is fitted to the $\boldsymbol{Y}$-residuals from the first PLSR,

and the orthogonalised $\boldsymbol{X}_{2,\,\mathrm{orth}}$. This step is repeated for all additional blocks, with the addition that all previous blocks are included in the orthogonalisation step. Block one and two are concatenated for this purpose, $T_{12} = [T_1\ T_2]$ and T_{12} used for orthogonalisation of block three ($\boldsymbol{X}_3$) by Equation 2.44 and $\boldsymbol{Y}$ in Equation 2.45.

$$\boldsymbol{X}_{3,orth} = (\boldsymbol{I} - \boldsymbol{T}_{12}(\boldsymbol{T}_{12}^T\boldsymbol{T}_{12})^{-1}\boldsymbol{T}_{12}^T)\boldsymbol{X}_3 \tag{2.44}$$

$$\boldsymbol{Y}_{orth*} = \boldsymbol{Y}_{orth} - (\boldsymbol{I} - \boldsymbol{T}_2(\boldsymbol{T}_2^T\boldsymbol{T}_2)^{-1}\boldsymbol{T}_2^T)\,\boldsymbol{Y} \tag{2.45}$$

$\boldsymbol{Y}$ is computed by summing the predictions of the individual regressions by Equation 2.46.

$$\boldsymbol{Y}_{pred} = \boldsymbol{T}_1\boldsymbol{Q}_1^T + \boldsymbol{T}_2\boldsymbol{Q}_2^T + \boldsymbol{T}_3\boldsymbol{Q}_3^T + \boldsymbol{F} \tag{2.46}$$

Figure 2.16 shows all the steps in the SO-PLSR algorithm where PLSR models are fitted, and how the orthogonalisation is implemented.

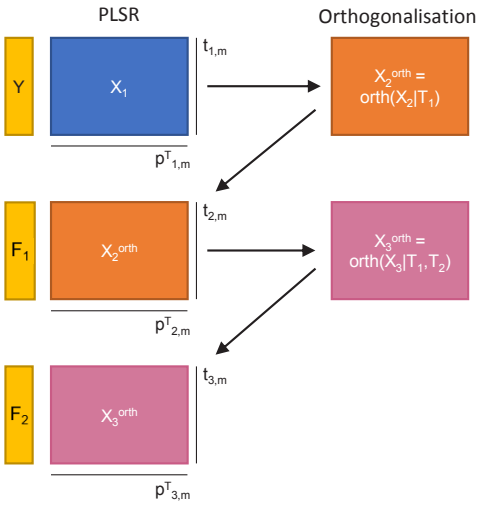


Figure 2.16: Schematic illustration of the SO-PLSR algorithm, starting with a PLSR model from which the scores are used to orthogonalise the second block, which is then fitted to a new PLSR model. The scores from the first and second PLSR model are used to orthogonalise the third block before it is fitted to a new PLSR model.

SO-PLSR is designed to handle blocks of different complexity and type; it can handle

both varying numbers of variables and differences in dimensionality. Additionally, SO-PLSR is invariant to block-scaling. The order of the blocks are of importance in SO-PLSR, contrary to in MB-PLSR, and changing the order will have an impact on the solution.

2.10 Deep learning

Deep learning models are machine learning models based on artificial neural networks, which aim at mimicking the structure and decision making of the human brain. Artificial neural networks (ANNs) [129, 130, 131] are computing systems consisting of nodes called artificial neurons, which have connections between them that are often initialised at random and adjusted by backpropagation [132, 133]. ANNs use neurons in multiple layers to progressively extract higher level features from the data, and then backpropagation uses the prediction error to calculate the gradient of the loss function with respect to the weights in the network. The neurons are typically placed in an input layer, one or more hidden layers and an output layer. In each layer, the input data is transformed into a more abstract and composite representation, learning from the data in each step. A widely used type of composition is the nonlinear weighted sum given by

$$f(x) = K\left(\sum_i w_i g_i(x)\right), \tag{2.47}$$

where K is the activation function, w_i are the weights and g_i are the different functions that are combined in the network. The weights are numerical values attached to each input variable conveying the importance of that corresponding variable when predicting the final output. Weights are calculated at each neuron by applying a specific function to the input values received from the previous layer, also determining how strongly each of the neurons affect the others. Weights with high values indicate that the variables have high impacts on the target value and contrary, the variables with low weight have a low impact on the target. The bias shifts the activation function which introduces non-linearities in the model. The learning consists of iteratively adjusting the weights and bias. Finally, the summation function combines the weights and the inputs to calculate the output. How each part of the network works together to predict the response ($\boldsymbol{y}_{\text{pred}}$) is shown in Figure 2.17.

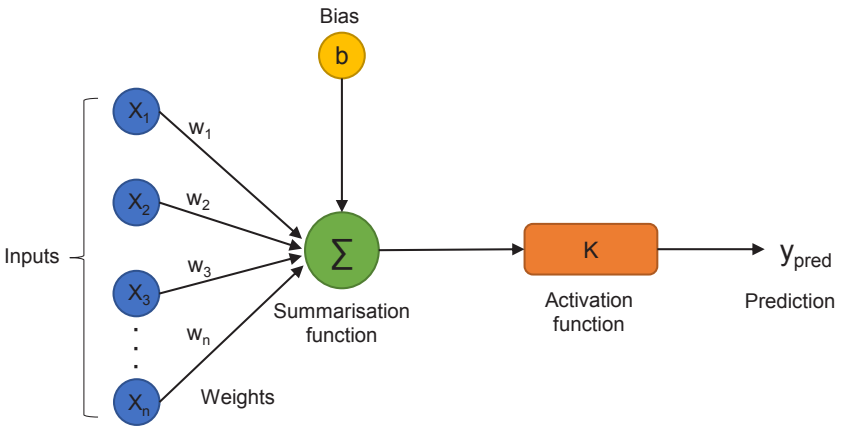


Figure 2.17: Illustration of the composition of an ANN, how the input data is transformed by the various functions combined in the network and the activation function to predict the output.

A typical ANN is trained with experimental data where the output usually is a non-linear function of the input data after learning a pattern, and creating a prediction model [134]. ANNs are self learning, meaning that the network can adjust weights when a new situation is introduced, leading to more flexible predictions than traditional regression models. Still, the network needs manual tuning to determine the optimal number of layers and neurons, which yields good predictions without over-fitting. Deep Neural Networks [135] are ANNs with multiple hidden layers between the input and output layers, as for instance the network shown in Figure 2.18.

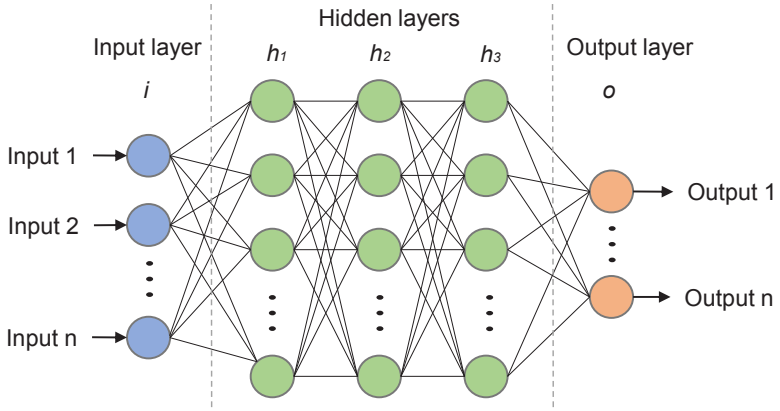


Figure 2.18: Schematic illustration of a neural network with an input layer, hidden layers and an output layer. Reprinted from Gjelsvik *et al.* [56].

Neural networks are used in a wide range of tasks, from computer vision to speech recognition, video games and medical diagnosis. Several new specialised versions of neural networks have emerged to handle these various tasks more efficiently. For instance Convolutional Neural Networks (CNNs) for computer vision and Recurrent Neural Networks (RNNs) for applications such as language modelling.

2.10.1 Convolutional Neural Networks (CNNs)

CNNs are deep neural networks which use convolutions to extract information in one or more hidden layers [136, 134]. CNNs are regularized versions of fully connected networks and consist of an input layer, hidden layers (mainly convolutional layers, pooling and fully connected layers) and an output layer, as illustrated in Figure 2.19. In the convolutional layers, the data is organised in a feature map where the weights are connected to the previous layer. In a CNN the vectors of weights and bias are called filters and represent the particular features of the input, and are used to filter for patterns in the data. The pooling layer semantically merges similar features, reducing the dimension of the representation [134]. Commonly used in pattern recognition, CNNs are good feature extractors as they learn the most important features by themselves. In contrast to PLSR, CNNs can handle non-linear data.

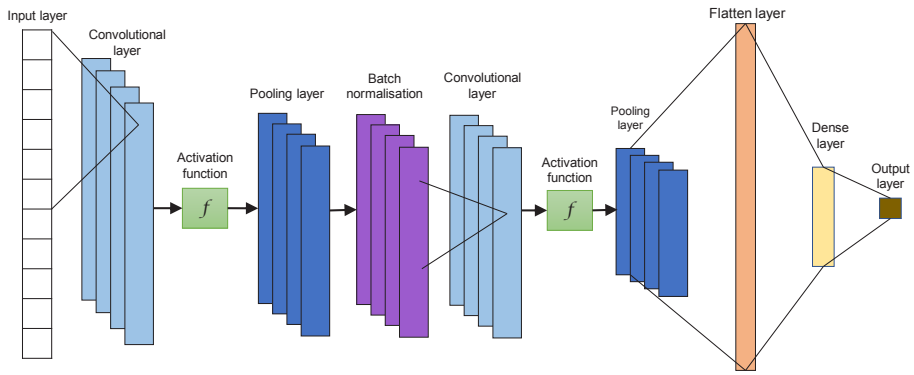


Figure 2.19: Schematic illustration of a CNN with an input layer, convolutional layers with an appropriate activation function, pooling layers, batch normalisation, a flatten layer, a dense layer and an output layer.

Pattern recognition is of high interest in analysis of spectroscopic data as it is assumed that different parts of the spectra are chemically connected, and these connections are often of high importance.

2.10.2 *Recurrent Neural Networks (RNNs)*

RNNs are often used for sequential or time-series data, feeding the output from one layer as input to the next layer [137, 138]. Like CNNs, the RNNs learn from the training input, but the RNNs use internal states (memory) to impact inputs and outputs with previous information. Therefore, RNNs have strong capabilities of capturing contextual data from a sequence. In an RNN, the input sequence is processed one element at a time with the memory in the hidden units retaining information on all the elements in the sequence [134]. Figure 2.20 shows how the recurrence happens in a RNN; the hidden layers are structured in the same way as in Figure 2.18, but with each layer updating with the previous information.

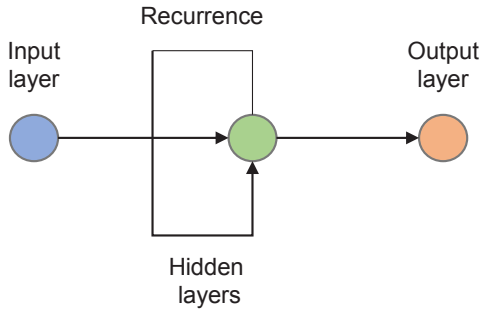


Figure 2.20: Schematic illustration of a RNN with an input layer, hidden layers and an output layer. How the recurrence functions is also shown.

In the case of chemical spectra, the peaks are often connected to adjacent peaks or can appear in certain patterns, which is why pattern recognition methods such as CNNs and RNNs can give good prediction models.

2.11 Variable selection and feature importance

For data sets with a large number of variables or features, it is often assumed that some variables are irrelevant or redundant and can be removed without loss of information. The term variable is more common in statistics, while the term features is common in machine learning. They both refer to the columns of a data set and are used interchangeably in this thesis.

Variable selection is the process of selecting a subset of relevant variables to use when constructing a model. Removing redundant variables can improve the prediction ability of the model, ease interpretation and reduce the computational cost

during modelling. For spectroscopic data, a large number of molecules present in the samples are detected, but in many cases only a few of them are actually related to the response. Identification of the relevant variables in the data set is therefore of high interest. Several regression and classification methods have built-in feature importance measures which can be utilised to select those which have the strongest relationships to the response. For instance, in regularisation, the constraints give weights to the variables, and in LASSO, the regression coefficients of unimportant variables are even set to zero. PLSR however, does not have such an in-built measure and Variable Importance in Projection (VIP) is therefore often used for variable selection in PLSR models [139].

2.11.1 Variable Importance in Projection (VIP)

VIP scores are calculated as the weighted sum of squares for the PLSR weights, which take the amount of explained variance in $\boldsymbol{Y}$ into account for each extracted latent variable. VIP scores can therefore be used to select the variables that contribute the most to the explanation of the variance in $\boldsymbol{Y}$. Since the variance explained by each component can be calculated by $\boldsymbol{q}_j^2 \boldsymbol{t}_j^T \boldsymbol{t}_j$, the VIP score for a variable K can be calculated by

$$VIP_K = \sqrt{n \frac{\sum_{j=1}^{A} \boldsymbol{q}_j^2 \boldsymbol{t}_j^T \boldsymbol{t}_j \left(\frac{\boldsymbol{w}_{kj}}{\|\boldsymbol{w}_j\|} \right)^2}{\sum_{j=1}^{A} \boldsymbol{q}_j^2 \boldsymbol{t}_j^T \boldsymbol{t}_j}}, \tag{2.48}$$

where $(\boldsymbol{w}_{kj} / \parallel \boldsymbol{w}_j \parallel)^2$ represents the importance of the k-th variable, wherein $\boldsymbol{w}_\mathrm{j}$ is the weight vector, $\boldsymbol{w}_\mathrm{kj}$ is the k-th element of $\boldsymbol{w}_\mathrm{j}$. Additionally, $\boldsymbol{q}_\mathrm{j}$ are the loadings and $\boldsymbol{t}_\mathrm{j}$ is the score vector from PLSR with A components. Variables with VIP scores greater then 1 are generally considered as important, however this limit is sensitive to non-relevant information in $\boldsymbol{X}$ [140] and should therefore be evaluated individually for each task.

2.11.2 Permutation feature importance

Permutation feature importance [115] is another measure of the importance each variable has for the response, which is often used for methods without built-in variable selection. Permutation feature importance is a model inspection technique that identifies important variables based on changes in the prediction accuracy when a variable is randomly shuffled (permuted). If the prediction accuracy of the model decreases significantly when a variable is randomly shuffled, this indicates that the

variable is important for the model's ability to predict the response. Similarly, if the prediction accuracy is unaffected when a variable is randomly shuffled, the variable is not important for the prediction. The importance of the variables is calculated from

$$\boldsymbol{i}_j = \boldsymbol{s} - \frac{1}{\boldsymbol{K}} \sum_{k=1}^{K} \boldsymbol{s}_{k,j}, \tag{2.49}$$

where $\boldsymbol{s}$ is the reference prediction accuracy of the model with the original features, $\boldsymbol{s}_{k,j}$ is the prediction accuracy of the models with shuffled variables and $\boldsymbol{K}$ is the number of variables.

However, the work does not stop with simply identifying a variable as important. The main interest for many problems is to identify what makes this variable important. By investigating the selected variables' position in the spectra, chemical groups can be identified, which can lay the foundation for new knowledge of the problem at hand.

3 Experimental methods

To identify hydrate active components and relate FT-ICR MS spectra to crude oil properties associated with hydrate formation, experimental data was needed. Two methods for determining hydrate chemistry were further developed from previously established experimental procedures. Additionally, the density of the crude oils was measured and related to spectroscopic data. In this section all experimental methods used to generate data are described, and their shortcomings are pinpointed.

3.1 Fluid systems

The fluid systems used for the hydrate experiments consisted of water, a crude oil and a synthetic natural gas, to imitate the contents of an oil pipeline. The water phase consisted of 3.5% NaCl in tap water. This was done to only introduce monovalent ions, and thereby generalising the water chemistry compared to the water during production, avoiding any unwanted reactions by bivalent ions such as Ca^{2+} [36]. Ca^{2+} can bind to acidic groups in the oils and this could have an effect on the detection of hydrate active components if these are present in the acid fractions of the oils. The ARNs can also cross-link with Ca^{2+} and precipitate to disturb the experiments. There are small amounts of Ca in tap water (on average 20 mg/L in Trondheim where the studies were conducted), but since these amounts are smaller than in production water, it was assumed that they would not cause any unwanted reactions. The gas phase consisted of a mixture of 86/8/6 mol% of methane, ethane and propane respectively (Linde Gas AS) with a mixture tolerance (accepted deviation from target value) of 10% and an analysis uncertainty of 2%. This gas phase was chosen as it contains the gases most commonly involved in gas hydrate formation.

Hydrate formation was performed experimentally by SINTEF, using a 200 mL high-pressure Sapphire Cell (Top Industrie), shown in Figure 3.1, placed inside a temperature controlled chamber. The temperature was measured using a PT-100 element

positioned at the bottom of the cell. A connected stirrer mixed the phases to create a fully dispersed system. The cell was fitted with a Hy-Lok FT Micron Tee Filter with a 150 μm sintered stainless steel filter element. A probe inserted from the top was used to measure the conductivity in the liquid phase. Gas filling was controlled using an IN-FLOW HI-Press MFC mass flow controller (Bronkhorst). A Bosch video camera was used for monitoring and capturing videos of the cell. On average five different water cuts were tested for each oil to determine where the hydrate formation occurred. The water cut is the ratio of water compared to the total volume of the system.

Figure 3.1: Wetting Index cell used for hydrate formation experiments with stirrer, temperature regulation, pump and camera. Picture by Martin Fossen.

3.1.1 Successive accumulation of hydrate active components

A successive accumulation procedure was developed by Fossen *et al.* [58], based on the procedure by Borgund *et al.* [39], with the aim of increasing the concentration of the hydrate active components. A schematic illustration of the procedure is shown in Figure 3.2. The procedure started with a fresh oil sample which was added to the

cell with the water phase at a given water cut and pressurised with a hydrocarbon gas phase. The pressure was set to 65 bar and the temperature was lowered to 2°C while stirring the liquid to ensure a homogeneous dispersion. By cooling the system at high pressure, the hydrate formation region will eventually be reached and with subcooling, the system will form hydrates. The systems were, in most cases, kept at low temperature over night, to ensure hydrate formation to approach towards equilibrium. The phase not associated with hydrates, referred to as the bulk phase, was drained through the bottom of the cell. The pressure difference of the cell and the ambient pressure conditions outside the cell were the driving forces for draining. The hydrate phase was retained by the filter, so that only water and oil not associated to hydrates were drained. After this draining of the bulk phase, the cell was depressurised and the temperature was increased, causing dissociation of the hydrate phase remaining in the cell. This resulted in a now liquified hydrate phase, consisting of an oil and a water phase that had been associated to the gas hydrates. The hydrate phase was then mixed with fresh oil and water at a ratio ensuring the same water cut as the previous run, and the hydrate formation and draining procedure were repeated. Small samples were taken from both the bulk phase and the hydrate phase at each step for analysis by FT-ICR MS.

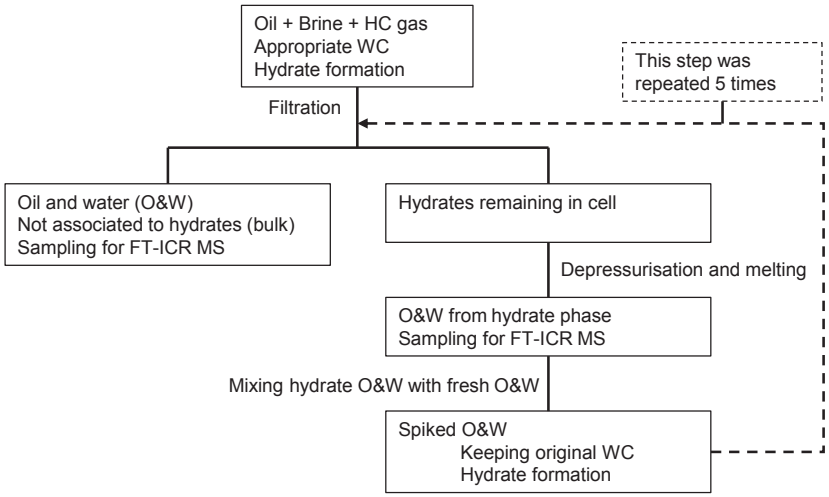


Figure 3.2: Schematic illustration of the successive accumulation experiment for spiking of the hydrate phase.

As each hydrate formation step was allowed to proceed over night, each accumulation cycle lasted one day. The oils tested using this procedure were accumulated at least four times, and all steps required manual handling. Consequently, this is a time and

resource demanding procedure.

3.1.2 Wetting Index experiments

The wettability of the hydrate particles can be expressed as the Wetting Index (WI) for an oil, which is a measure of the emulsion inversion point with and without hydrates present. A WI procedure for determining the emulsion inversion point for hydrates was developed by Høiland *et al.* [141] and improved by Fossen *et al.* [58]. The emulsion inversion point is the point where the phases shift from one being the dominant phase to the other being the dominant phase. Oil- and water-continuous emulsions are defined as homogeneous samples, and the inversion from oil- to water-continuous is followed by an abrupt change in the viscosity of the oil-continuous emulsion to an intermediate region where the samples have the appearance of an inhomogeneous mix of both emulsions. This intermediate region is often referred to as the inversion range. The further transition to water-continuous is quantified as the amount of water needed for the emulsion to change from an inhomogenous mixture to a homogeneous one. The emulsion inversion point is regarded as the volume fraction of water at which the emulsion converts from oil-continuous to the intermediate region.

When the emulsion inversion point shifts towards higher water cuts after hydrate formation, the hydrates are oil-wetted and when the water cut shifts to lower water cuts, the hydrates are water-wetted. This follows the principles of Bancroft [142]. The WI is defined as the normalised difference in inversion point with and without hydrates present, represented by a number between -1 and +1. Positive values indicate oil-wetted systems with little or no probability of plugging, and negative values indicate water-wetted systems with a high potential for plugging. The absolute value of the WI number is expected to be of importance, and a higher positive or negative value indicates higher degrees of oil-wetted or water-wetted hydrate particles, respectively.

Experimentally, the WI for each sample was measured as follows: The cell was first filled with oil and water according to the selected water cut, with a total liquid content of 160 ml. Then the cell was pressurised with the hydrocarbon gas phase to 65 bar. Stirring at a rate of 25 Hz was applied to disperse the liquid phases, while cooling the system at constant pressure until hydrates were formed. This process took approximately 12 hours. The cooling rate was on the order of 8 K/hour. Hydrate formation was detected based on the changes in temperature measurements of the liquid phase and visual inspection. In some cases the temperature showed no clear change upon hydrate formation, and then changes in the conductivity and

visual inspection were used to verify hydrate formation instead. The conductivity is the main information needed to determine the WI, and high conductivity indicates a water continuous system, while a low conductivity indicates an oil continuous system. The point of phase inversion ($\Delta\varphi_w^{inv}$) is considered as the volume fraction of water where the emulsion converts from oil continuous to be in the inversion range, and is determined from the conductivity measurements. The inversion point for determining the WI ($\Delta\varphi^*$) is calculated and normalised by

$$\Delta\varphi^* = \frac{\Delta\varphi_w^{inv}}{\Delta\varphi_{max}}, \tag{3.1}$$

where $\Delta\varphi_{max}$ describes the maximum possible variation for the inversion point in systems without hydrates $\Delta\varphi_w^0$. For high conductivities, $\Delta\varphi_w^{inv} > 0$, and $\Delta\varphi_{max}$ is calculated by

$$\Delta\varphi_{max} = 1 - \Delta\varphi_w^0 \tag{3.2}$$

while for low conductivities, $\Delta\varphi_w^{inv} < 0$, and $\Delta\varphi_{max}$ is equal to $\Delta\varphi_w^0$.

3.1.3 Measurement uncertainties

The uncertainty of the WI method has not been systematically tested nor are standard deviations calculated. However, since multiple water cuts in close proximity were tested, uncertainties or irregularities in the determination of the continuous phase were detected and the water cut re-checked by a duplicate test to control the consistency of the method. Additionally, during the WI experiments, it was discovered that oils measured previously in another project 2 years prior, received completely different WIs. One possible explanation for this could be that storing the oil has an effect on the composition of the oil, and that these results might not be representative for a fresh oil. The reason for this deviation was not further investigated. It is difficult to evaluate the accuracy of these measurements as they are heavily dependent on the composition of the crude oils. The external parameters such as temperature, pressure and stirring are controlled to create an equal system for all oils, but hydrate formation is still affected by the oil composition.

3.2 Density measurements

Density was measured using a Sigma 703D by Biolin Scientific using the associated density probe. All samples were measured at room temperature during a period of 1 week. The measurements were conducted as follows: First the density ball was placed on the hook before taring, and the density should then read $1.2e^{-3}$ g/ml, i.e. the density of air. The beaker was then filled with the fluid oil sample and placed so that the stagnant density probe was fully immersed in the sample. Lastly, the density value on the display was recorded. The density was measured three times per sample and the value was reported as the average of the three repeats. Between each measurement, the density probe was cleaned with toluene and acetone and dried with an air pistol. The instrument power was turned on at least 30 minutes before the first sample was measured to reduce variation caused by instrument heating.

3.3 FT-ICR MS analysis

For FT-ICR MS, the samples were prepared by diluting 20 μL of sample in 980 μL dichloromethane (Supelco Suprasolv for gas chromatography MS). From this diluted sample, 20 μL were diluted further in 980 μL of a 1:1 mixture of toluene (Sigma-Aldrich CHROMASOLV for HPLC 99.9%) and methanol (Baker analyzed LC-MS Reagent). The samples were diluted a total of 100 times and from each diluted sample, 100 μL were injected using a High Performance Liquid Chromatography (HPLC) system as the introduction device. This was done to inject a steady flow over a given period of time to assure a good collection of spectra. In this work, the samples were injected over a time period of 10 minutes. The samples were analysed in three parallels each. For each parallel, 220 spectra were collected and the data for each parallel was given as the average over all 220.

3.3.1 Data Preparation

Before data analysis, preprocessing of the raw data is required. The data from each sample was first combined into a bucket table using Bruker Compass ProfileAnalysis 2.1 [143]. Bucketing is the process of removing unwanted variations in peak positions due to changes in shifts during analysis. This is usually done by dividing the spectrum into equally sized parts, namely buckets, integrating the intensity values in each bucket and annotating this value to the bucket [144]. However, when doing so, small peaks adjacent to large peaks become overshadowed and diminished, thus resulting in a loss of sensitivity. Still, FT-ICR MS has such a high mass accuracy that this reduction in sensitivity is negligible. It rather reduces the amount of peaks

and thereby also the dimension of the data set, something that could be beneficial for data analysis. The settings in ProfileAnalysis can be set in numerous ways, and based on previous information, the settings in this work were as follows: Normalisation was set to the sum of bucket values in the analysis, no baseline or smoothing was used, the S/N threshold was set to 4, the relative intensity threshold was 0.01 and the absolute intensity threshold was 100. Based on this, the average peak list was calculated.

Normalisation is a common way to preprocess FT-ICR MS spectra and consists of dividing each spectrum by an estimation of its spectral intensity. This can be done with regards to a selected area, the maximal peak in the spectrum, a specific spectral point, spectral length or the sum of the spectral values. When the normalisation is set to the sum of the bucket values in the analysis, each sample will be normalised to have the same total intensity over all the peaks.

The signal-to-noise ratio (S/N) can be used as a measure for the detector performance, as it conveys information about the lower limit of detection [145]. The smallest signal that can be attributed to an analyte is commonly regarded to have a S/N of 2 of more. During the experiments presented here, the S/N was set to 4 to disregard the smallest peaks as their inclusion would have contributed to more noise in an already large data set.

The threshold is defined as the size of the signal exceeding the noise level, and is used to determine the start and the end of a peak. The ideal threshold should be low enough to trigger the start of a peak when the signal is just slightly higher than the noise, and high enough to prevent noise from being mistaken as a peak. This setting is the intensity threshold in ProfileAnalysis. In this work, the relative intensity threshold was set to 0.01, and the absolute intensity threshold to 100. This means that the signal had to have an intensity of 10% (relative) or 100 (absolute) higher than the noise to be included as a peak.

The focus area of the FT-ICR MS method is set to m/z 250, which is where the largest abundance of peaks are expected to occur. This means that the sensitivity of the instrument decreases over m/z 1000, simply because this area is far away from the focus area, causing more noise in the measured peaks for high masses. It is more difficult to determine molecular formulas for molecules with masses over 1000 Da. Additionally, high masses have a higher possibility to ionise with multiple charges. The system is set to avoid multiple charges, but multiple charges are not completely uncommon in the higher masses.

3.3.2 Irregularities in the spectra

During analysis of the ESI(+) FT-ICR MS spectra, it was discovered that some spectra had peaks with substantially higher intensity than the remaining spectra. These spectra also had fewer peaks and a shape which is not common for oil spectra. During inspection of the peaks in spectra with these effect, it was discovered that the distance between many of the peaks was m/z 44.026 which is the weight of a C_2H_4O molecule. This is also the difference between chains of polyethylene glycol (PEG) molecules, as the differences between molecules in a PEG series is a C_2H_4O group. The hypothesis is therefore that these shape irregularities in the spectra were caused by the presence of PEG in the samples. Figure 3.3 shows an example of an oil spectrum with the shape characteristic for crude oils shown in blue, and an example of a spectrum for a sample likely containing a PEG series shown in red. From Figure 3.3, the differences in intensity between regular oil spectra and spectra containing PEG are shown, with a maximum intensity of 1.4^9 in the oil spectrum and 1.2^{11} in the spectrum containing PEG. During the work in **Paper II** and **Paper III**, peaks with molecular formulas corresponding to PEGs were selected by the variable selection methods. It is not unlikely that other production chemicals could be found in the spectra as well, although in this work only PEGs were revealed.

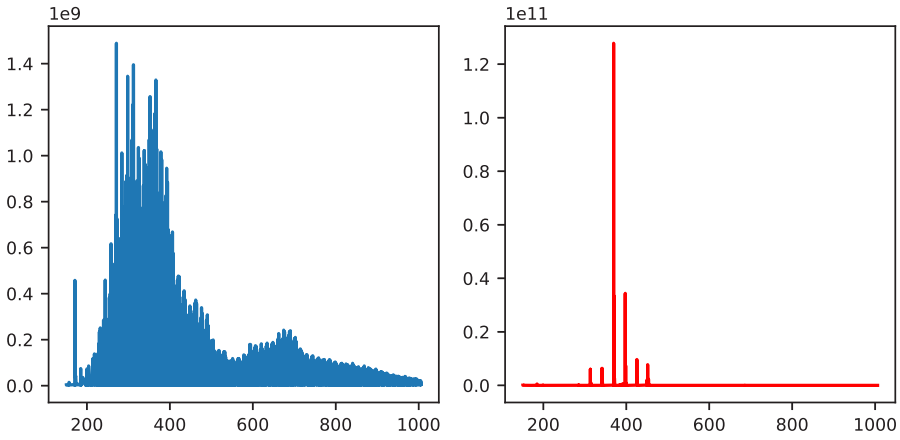


Figure 3.3: Spectrum with the shape characteristic for crude oils in blue and the spectrum of a sample containing PEG in red, with the differences in intensity indicated at the top of the spectra.

Another issue was uncovered when the replicates for the samples were compared. The measurements for the replicates for each sample were not consistent, exposing some type of non-linearity in the system. This is shown in Figure 3.4, where the left plot shows how the data should be if the replicates were in agreement, and the

right plot illustrates the non-linearities present in some of the samples. This could possibly be due to non-linearities in the form of signals with odd-numbered multiples of the fundamental reduced frequency from Equation 2.8, caused by variations in the time-domain signal magnitude as a function of the ion cyclotron orbital radius, referred to as harmonic signals [67]. These harmonic signals can increase the number of peaks in the spectra and interfere with identification of other fundamental signals which are low in magnitude. One hypothesis is that these effects can be remedied in the same way as non-linearities are handled in IR spectra, with preprocessing. However, the underlying cause should be determined first to evaluate whether this could be handled directly. As of now, what causes these effects in some of the samples is unknown.

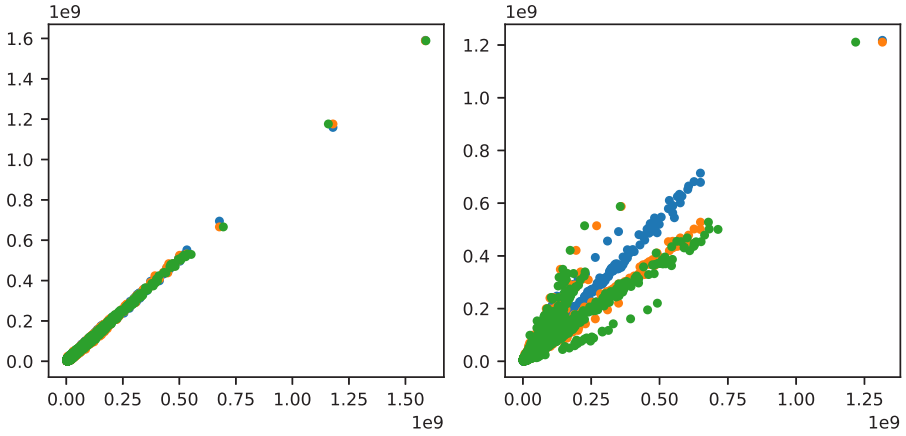


Figure 3.4: The three replicates for two samples plotted against each other. The left plot shows one sample where the replicates are in agreement and the right plot shows a sample with non-linearities present.

3.4 IR analysis

IR analysis was performed by applying 20 μL of crude oil onto the detection window of a PerkinElmer Frontier FTIR/NIR Spectrometer, an instrument capable of measuring all ranges in the IR region. When set to NIR mode, the instrument operates in the range 15 800-2000 cm^{-1} and for FTIR mode it operates from 8300-400 cm^{-1} [146]. The advantages of these methods are that the analysis is fast, requires none or minimal sample preparations and is non-destructive.

The FTIR measurements were taken over the range 4000-800 cm^{-1}, while the NIR measurements were taken over the range 12800-4000 cm^{-1}. The resolution for both

methods was 4 cm^{-1} with 16 scans for each sample. The final spectra for each sample was the average of the 16 scans with a background spectrum subtracted. For data analysis, the spectra from FTIR were combined to one data set and the data from NIR to another. The data sets were then preprocessed by EMSC and/or Savitzky–Golay depending on the degree of removal of non-linearities achieved in each of the data sets.

4 Summary of the papers

In this work, machine learning was used with the aim of identifying naturally occurring components related to hydrate formation, where the high mass accuracy of FT-ICR MS was utilised to obtain detailed spectra of crude oils. In this section a summary of the work presented in the papers and the rational behind each paper will be given. In **Paper I** a literature review revealed that little work has been done to identify naturally occurring hydrate active components from FT-ICR MS spectra using machine learning. Therefore, a feasibility study was done in **Paper II** to test the proficiency of various variable selection methods for this purpose. As this was successful, we expanded on this work by including additional methods in **Paper III**, and several molecules which were related to hydrate formation were identified. However, the prediction accuracies in **Paper II** and **Paper III** were not high, and FTIR and NIR were included in **Paper IV** to further characterise the oils. Lastly, new methods able to handle large variations in the relationships between $\boldsymbol{X}$ and $\boldsymbol{Y}$, and non-linearities were developed in **Paper V**.

4.1 **Paper I**

Current overview and way forward for the use of machine learning in the field of petroleum gas hydrates

To gain an overview of the current status of the use of machine learning in the field of gas hydrates, a literature study was performed. As the main aim of this project was to identify naturally occurring hydrate active components, a literature search was done to find all literature where machine learning had been used to identify hydrate active components from FT-ICR MS spectra, which returned zero publications. Therefore, the search was expanded to all literature regarding machine learning on gas hydrates. FT-ICR MS was included in the search string to establish a link between gas hydrates and analysis of crude oils. Several machine learning methods were hypothesised to be useful in this context, and their names were there-

fore also included in the search. The included methods were Principal Component Analysis, Partial Least Squares Regression, Decision Trees, Random Forest, Artificial Neural Networks, Convolutional Neural Networks, Support Vector Machines, regularisation, Bayesian Networks and K-Nearest Neighbours. The results showed that both the field of gas hydrates and FT-ICR MS had limited publications, however, both showed an increasing trend over the last few years. For gas hydrates, ANNs were most commonly used, while for FT-ICR MS, PCA was the most commonly used method. A text analysis study was performed to group the search results according to different topics.

The text analysis revealed that even though quite restrictive search terms were used, many of the topics in the literature were not relevant for gas hydrates and crude oil analysis. From the relevant topics, several publications from both fields were determined to be of value for future research regarding identification of hydrate active components from crude oils. They were further evaluated with the aim of identifying trends and gaps in the research.

The retrieved literature showed that for gas hydrates, machine learning has mainly been used for determination of thermodynamic properties for hydrate formation. Additionally, many of the papers were based on data sampled from the literature, and in most of the cases, the same data sources were used. This leads to low diversity in the results, and it became apparent that more experimental data is needed. The thermodynamic properties of hydrate formation have been thoroughly evaluated already, and it could therefore now be time for creation of other types of data, for instance data where hydrate formation is related to the chemical properties of the oil. For FT-ICR MS, machine learning has mainly been used for crude oil characterisation, however, most studies did not expand past PCA and visualisation using unsupervised methods. As FT-ICR MS spectra contain such large amounts of data, many machine learning methods could be beneficial to yield a deeper understanding of the spectra.

This review showed that several machine learning methods have been tested and proven effective for analysis of thermodynamic properties of gas hydrate related samples, and for chemical properties of data from FT-ICR MS. However, at the time this search was carried out, no studies existed in the cross section between the two. To fully understand the formation mechanisms of gas hydrates, it is important to understand the chemistry that is involved. We therefore believe that characterising gas hydrate related crude oil samples, could provide valuable insights.

4.2 Paper II

Identifying components related to hydrate formation by machine learning-based variable selection

The FT-ICR MS spectra usually contain between 10 000-25 000 peaks due to the high mass accuracy of the instrument. It is highly unlikely that all these peaks are related to the response, and some peaks could even be noise, reducing the prediction ability of the model. It is therefore of interest to reduce the data set by removing the unimportant variables, and at the same time identify the variables which are important so that their chemistry can be interpreted. Variable selection can be used to find the variables which contribute the most to the response in a data set. In **Paper II**, tree based methods such as Decision Trees, Random Forest, Bagging and Boosting, as well as regularisation based methods such as LASSO and Ridge Regression were tested with the aim of identifying naturally occurring components in crude oils related to hydrate formation. These methods were selected as they are apt at handling data where the number of variables are larger than the number of samples, which is the case for most FT-ICR MS data sets.

Two oil samples (anonymised to A and J2) were subjected to the successive accumulation procedure measuring six spiking levels for oil A and 4 spiking levels for oil J2. The WI was also measured for the two oils, resulting in the value +0.44 for oil J2, indicating that J2 forms transportable hydrates, and 0 for oil A, indicating that hydrates formed in oil A can be either transportable or plugging. Samples were measured in three parallels using ESI(+)-FT-ICR MS, both for bulk samples and hydrate samples. PCA showed that there was a difference between the spiked samples and the crude oils, and between some of the spiking levels. To investigate whether any components were accumulated during the spiking procedure, the mass spectrum for an unspiked sample was subtracted from each spiking level. This should show an increase for peaks where an accumulation had occurred. Some peaks emerged as interesting from this analysis.

Classification was performed with the sample origin as the response, either from bulk (0) or from hydrate (1). The results showed that Boosting yielded the best classification accuracy. For each of the methods, the five variables (m/z-ratios) selected as the most important, the corresponding proposed molecular formulas, DBEs and H/C-ratios were determined, in addition to which oil samples the selected variables appeared in. The results showed that some of the selected variables probably corresponded to asphaltenes and some variables could possibly be sulfoxides. Sul-

foxides have previously been shown to be correlated with agglomeration of hydrates, and could therefore be related to the WI for oil A which indicates that it can form plugging hydrates. This conference paper was written to investigate the possibilities for using variable selection to identify important components from the FT-ICR MS data.

4.3 Paper III

Using machine learning-based variable selection to identify hydrate related compounds from FT-ICR MS spectra

Paper II showed the possibilities for using variable selection for identification of naturally occurring hydrate active components. The study was expanded and one more oil (anonymised to I) was included. The successive accumulation was also performed for oil I, resulting in five spiking levels. The WI for oil I was measured to be 0.31. Bulk samples and hydrate samples from the three oils, oil A, J2 and I, were measured in three parallels each using ESI(+)-FT-ICR MS. PCA showed that the oils were different after the spiking procedure. Comparison of the mass spectra revealed that some peaks were increasing in intensity for oil A and I, however this was not observed for oil J2.

The most promising methods from **Paper II**, Decision Trees, Random Forest, LASSO and Gradient Boosting, were included in this study with an addition of new methods, PLS-DA and XGBoost. As in **Paper II**, classification was performed using the sample origin as the response, i.e. whether they were from the bulk phase (0) or from the hydrate phase (1). PLS-DA was determined to be the best performing classification method over 25 different training and test sets. During this analysis we discovered that the prediction accuracies of the models were very dependent on the composition of the training set. Some combinations of training and test set performed very well and some extremely poorly. The models were therefore validated using several training and test set splits, and a study was performed, splitting the data set into training and test sets 25, 50, 75 and 100 times, to determine a fair number of splits. The results showed that increasing the number of splits above 25 did not affect the standard deviation significantly. The variables selected in all of the 25 models were considered as important, and their molecular formulas, DBEs, H/C-ratios, corresponding adducts and molecular weights were estimated based on the m/z values.

The results showed that several of the variables could correspond to asphaltenic

structures, while many were determined to be PEG stemming from oil chemicals. Both the PCA and the comparisons of the mass spectra showed that the successive accumulation increased the concentration of some compounds in the hydrate phases for oil A and I. Some of these variables were also selected as important from the PLS-DA. The selected variables´ effects on hydrate formation were evaluated with regards to the WI of the crude oils. Additionally, the lack of peaks with increasing intensity in the mass spectra for J2 after accumulation, together with the high positive WI for J2, indicates that this oil is saturated with hydrate active components so that the spiking procedure did not change the composition of the oil.

In summary, our results showed that it is possible to identify components related to hydrate formation. However, more work should go into determining their molecular structures in detail, as the molecular formulas could correspond to a large number of different structures (the larger the mass, the larger the number of possibilities).

A major aim of this project was to establish a connection between the FT-ICR MS spectra and crude oil properties. To accomplish this goal, regression models were built to predict the WI from FT-ICR MS spectra. However, during the project period WI was measured for only 17 oils. All attempts to build useful models were unsuccessful, which can be attributed to the low number of available samples and also the previously discussed uncertainties in the measurements. Building models on few data points makes validation challenging and makes it difficult to see connections and draw conclusions. The spiking data was possible to correlate to the FT-ICR MS spectra using classification, but these data also yielded large deviations, as shown in **Paper III**. Some of these modelling issues can also be attributed to the complexity of the oils and their large geographical differences, in that oils from different locations have such varied chemical compositions that this alone can dominate the analyses.

4.4 Paper IV

Multiblock analysis combining data from FT-ICR MS, FTIR and NIR spectroscopy improves prediction of the density of crude oils

Crude oils are extremely complex with a large number of constituents and therefore difficult to characterise in full, even with the high mass accuracy of FT-ICR MS. The studies in **Paper II** and **III** showed the difficulties in modelling this type of data with mediocre prediction accuracies and large variations in prediction accuracies for different training/test set combinations. FTIR and NIR have previously been used extensively for crude oil characterisation as well [147, 148], and these

methods were therefore compared to APPI(+)-FT-ICR MS in **Paper IV**. Combining different complementary measurement techniques yields a more comprehensive characterisation, as properties not measured by one technique can be captured by another. Previous studies have shown the effect of comparing different ionisation sources [94, 149, 150], however no studies have been done previously to compare crude oil measurements from mass spectrometry to infrared spectroscopy. Measurements from ESI(+)-FT-ICR MS were also supposed to be included in this study, but due to the irregularities discovered between the replicates of the spectra described in section 3.3.2 the ESI(+) measurements were excluded.

APPI(+)-FT-ICR MS, FTIR and NIR were therefore used to characterise 42 crude oils, and the oils densities were measured. Prediction models using Partial Least Squares Regression (PLSR) on the data from each analysis technique (single-block PLSR) and the multiblock methods Multiblock Partial Leas Squares Regression (MB-PLSR) and Sequential Orthogonal Partial Least Squares Regression (SO-PLSR) on fused data were compared to predict the density from the spectra. Additionally, variable selection through variable importance in projection (VIP) was performed to reduce the data set, but more importantly, identify variables of high importance to the response. Density is highly related to the chemical structure of the samples, and it was therefore of interest to evaluate the chemistry of the selected variables to examine whether the three methods highlighted similar compound groups. The two multiblock methods, MB-PLSR and SO-PLSR, use two different procedures for fusing data, and both were compared to single-block PLSR to evaluate the benefit of using data from more than one spectroscopic technique.

The results showed that the multiblock methods did improve the prediction for the three data sources, both before and after variable selection. FTIR received the highest prediction accuracy in the single-block PLSR analysis, indicating a higher correlation between the FTIR peaks and density than in the other two blocks. The overall best performing prediction model was SO-PLSR using the variables selected from the single-block methods.

The selected variables from each method were compared and chemical groups related to density, such as naphthenic acids, aromatic groups and alkanes, were identified. Naphthenic acids and aromatic groups are related to heavy oils and high densities, while larger amounts of alkanes are related to light oils and low densities. The comparison of the selected variables from the three PLSR stategies showed that SO-PLSR consistently selected variables related to high densities, while single-block PLSR and MB-PLSR more often selected variables related to low densities. This is highly interesting since SO-PLSR also outperformed single-block PLSR and MB-

PLSR.

4.5 Paper V

Hierarchical cluster-based deep learning

In **Paper II-III**, PCA showed that the crude oils were vastly different, which was expected due to the highly complex matrix of crude oils. This was illuminated by the low predictive abilities of the models. Local modelling has the abilities of handling large inhomogeneity and variability in data sets, and was therefore believed to be valuable for improving the prediction. Unfortunately, too few samples were available in time to train and test the method on hydrate related samples.

Local modelling is based on the concept of dividing the data into smaller groups and fitting a model to each group, assuming that this will result in better fitting models for data with dissimilar relationships between $\boldsymbol{X}$ and $\boldsymbol{Y}$. The division of samples can be based on *a priori* information, however, this is usually not available for real world data. Hierarchical Cluster-based Partial Least Squares Regression (HC-PLSR) was developed by Tøndel *et al.* and utilises the capabilities of clustering to group similar data points, and creates one PLSR model for each cluster. New samples are classified into one of the clusters and predicted using the corresponding local model. However, PLSR is a linear method and requires that the data is locally linear within each cluster. For some data sets this is not the case, and here HC-PLSR will perform poorly. To overcome these issues with non-linearities in the local models, the HC-PLSR method was expanded into deep learning by the creation of HC-CNN, HC-RNN and HC-SVR. Neural networks handle non-linearities through applying non-linear functions to the weights during training, while SVR handles non-linearities though non-linear kernels.

To evaluate the hierarchical cluster-based deep learning models, FTIR measurements of raw material films from hydrolysed fish and poultry consisting of 28 known subgroups were used. Chicken, turkey, salmon and mackerel samples were hydrolysed with various enzymes, with the aim of predicting the molecular weight during the enzymatic hydrolysis. The results showed that the local deep learning and SVR models outperformed HC-PLSR.

The advantage of PLSR over deep learning lies in the easy interpretation offered by scores and loadings. Deep learning models are considered as "black boxes" - it is nontrivial to interpret how the neurons work together to create the output. In an attempt to compare the results from the local networks and HC-SVR to HC-PLSR,

methods for identifying important variables for each model were implemented. For HC-CNN and HC-RNN, VarGrad [151, 152] was used, which adds small amounts of noise to each variable and monitors the effect the noise has on the output. Permutation Feature Importance was used for SVR, permuting one variable at a time and evaluating the resulting change in prediction accuracy. The visualisation revealed that samples containing similar chemical groups were clustered together, and the important features from the four methods highlighted these differences in chemical groups. The peaks determined as important were related to known absorption bands for dry-film FTIR spectra. This shows that the local models were able to find clusters and build models based on differences in the chemical composition of the samples without using any prior knowledge.

In order to avoid overfitting, the number of clusters should be kept relatively low. It is also advantageous to keep the number of PLSR components low, to ease the clustering and subsequent interpretation of the models. In this study, HC-PLSR needed a higher number of clusters than HC-CNN, HC-RNN and HC-SVR, as expected, to be able to better handle the heterogeneity in the data.

Paper V is currently a manuscript and we are working on including another test case based on simulated data to provide further proof of concept for the HC-deep learning methods.

5 Conclusion

One of the main contributions of the thesis to the field of gas hydrate research has been to use machine learning to establish a link between FT-ICR MS spectra and the identification of naturally occurring hydrate active components in crude oils. Several crude oil constituents which have previously been related to hydrate formation were during this work identified from the spectra by various variable selection methods. A literature review revealed that no papers were previously published where machine learning was applied to FT-ICR MS spectra of gas hydrate related samples with the intent of identifying naturally occurring hydrate active components. Hence, this field had not been explored much previously.

In this thesis, a number of methods for analysing FT-ICR MS spectra have been demonstrated. FT-ICR MS data can be difficult to work with as the high mass accuracy causes the spectra to contain large amounts of peaks, which not all methods are able to handle. Additionally, in both **Paper II** and **Paper III**, the prediction models based on FT-ICR MS data gave surprisingly low prediction accuracies. It was expected that it would be relatively easy to build models on FT-ICR MS data with all the chemical information contained in the spectra. However, this was not the case, neither for ESI ionised molecules or APPI ionised molecules. In **Paper IV** on the other hand, the effect of reducing the FT-ICR MS data set using variable selection showed a significant increase in the prediction ability of the PLSR model on the APPI(+) FT-ICR MS data. This is a strong indication that the spectra might contain too much information, causing the relevant parts to be overshadowed. This work therefore demonstrated that using variable selection could be highly beneficial for this type of data.

One of the aims of the main project this work is a part of was to utilise the WI measurements to evaluate whether an oil will be plugging or not. As FT-ICR MS spectra contain such vast amounts of information, it was assumed that the spectra also contain the information needed to predict the plugging potential for hydrates.

Before commencement of the data analysis performed in this thesis, it was therefore expected that machine learning models could be built to replace the WI measurements with FT-ICR MS measurements. Since the WI measurements are time and resource consuming, very few tests were completed in the course of the project. Nevertheless, many attempts were made to create a model which could correlate the WI data to FT-ICR MS spectra, but all were unsuccessful. All methods described in this thesis were attempted, with and without variable selection, and even autoencoders, which are profound at reducing noise in the data, were attempted. Attempts to relate the WI to the FTIR and NIR data were also unsuccessful. If any hydrate active components are present in the oils, one hypothesis is that they exist in very low concentrations. Low concentrations can make their identification and quantification difficult amongst the large amounts of peaks in a spectrum. Another possibility is that the components acting as natural inhibitors are not the same for different oils. For one oil it could be asphaltenic compounds which inhibit agglomeration, while in another oil it can be naphthenic acids. Such differences make it difficult for a machine learning model to detect a pattern, and there might not even be a pattern. However, since very few samples were available, it is also difficult to make any conclusions, other oils could have exhibited completely different results, or with more oils there might have been enough of a pattern to detect correlations. These unsuccessful attempts can also indicate that there is not a strong enough relationship between WI and FT-ICR MS spectra to be able to create models.

The thesis further contributed to the field by the creation of new non-linear approaches to local modelling. The work done in this thesis displayed the difficulties in building models on FT-ICR MS spectra from crude oils. It is believed that some of these difficulties arise from the differences between the oil samples. The chemical composition of oils can be very different and composition also effects the oils relationship to the responses used in this work. The previously established local modelling procedure HC-PLSR was therefore believed to be proficient for such data, as its main premise is to separate the data into groups of similar data and build one model for each group. However, PLSR is a liner model, and to handle non-linear data, the HC-PLSR procedure was expanded into deep learning and SVR. When more data is available the proficiency of these methods can also be displayed on samples related to hydrate formation.

6 Suggestions for future work

The main focus of this work was to identify components from FT-ICR MS related to naturally occurring hydrate anti-agglomerants. The results presented in this thesis identify possible groups related to hydrate active components, and describe methods which can be used for this purpose. However, this is a difficult task with many road blocks which have also been illuminated. In this section some suggestions are presented for how the work from this thesis can be taken further.

The variable selection methods tested in **Paper II** and **III** were shown to be effective for identifying components related to naturally occurring hydrate active components and some possible molecular formulas were suggested. To further determine the molecular structures of these compounds, they can be fragmented using the FT-ICR MS. The large masses of some of the selected components imply that their molecular formulas can correspond to a large number of different structures. With fragmentation, the molecules are split up into smaller sections, and their patterns can be used to determine the parent molecule. When some molecules are identified, they can be added to a crude oil, and whether the compound is able to alter the hydrate formation and specifically the wettability of the system can be evaluated.

As indicated above, both the WI and the spiking procedures are quite time consuming, and the amount of available data was therefore limited in this project. The complex matrix of crude oils means that samples and their measured spectra can exhibit large differences, making them difficult to model together. Therefore, the local modelling procedures in **Paper V** were developed in an attempt to improve the predictive abilities of the models. However, we did not have enough oil data to test these methods on hydrate related FT-ICR MS spectra. When more data is available the local models should be tested, as they should be able to better account for complex non-linearities in the data.

An effort should be made to correct the replicate issue observed in the FT-ICR MS spectra. We were not able to ascertain why this is happening during this work,

and further investigation to identify the reason behind could hopefully reveal more answers. Until the reason is discovered, these issues could possibly be rectified by preprocessing methods traditionally used for IR spectra, such as EMSC/EISC. Various preprocessings were tested, where some did reduce the observed differences between the replicas, but further work is needed to evaluate their effect and if the preprocessing improves the subsequent models prediction abilities.

References

[1] C. A. Koh. "Towards a fundamental understanding of natural gas hydrates". *Chemical Society Reviews* 31 (Apr. 2002), pp. 157–167.

[2] E. D. Sloan and C. A. Koh. *Clathrate Hydrates of Natural Gases*. en. 3rd. Chemical Industries series 119. Boca Raton, FL: CRC Press, Taylor & Francis Group, 2008.

[3] E. D. Sloan. "Fundamental principles and applications of natural gas hydrates". *Nature* 426 (Nov. 2003), pp. 353–359.

[4] H. Davy. "I. The Bakerian Lecture. On some of the combinations of oxymuriatic gas and oxygene, and on the chemical relations of these principles, to inflammable bodies". *Philosophical Transactions* 101 (1811), pp. 1–35.

[5] E. G. Hammerschmidt. "Formation of Gas Hydrates in Natural Gas Transmission Lines". *Industrial and Enigneering Chemistry* 26.8 (Aug. 1934), pp. 851–855.

[6] W. Ke, T. M. Svartaas, and D. Chen. "A review of gas hydrate nucleation theories and growth models". *Journal of Natural Gas Science and Engineering* 61 (Jan. 2019), pp. 169–196.

[7] M. A. Kelland. "History of the Development of Low Dosage Hydrate Inhibitors". *Energy Fuels* 20.3 (Apr. 2006), pp. 825–847.

[8] Q. Nasir, H. Suleman, and Y. A. Elsheikh. "A review on the role and impact of various additives as promoters/ inhibitors for gas hydrate formation". *Journal of Natural Gas Science and Engineering* 76 (Feb. 2020), p. 103211.

[9] J.-H. Sa et al. "Investigating the effectiveness of anti-agglomerants in gas hydrates and iceformation". *Fuel* 255 (July 2019), p. 115841.

[10] M. A. Kelland. *Production Chemicals for the Oil and Gas Industry*. English. 2nd Edition. Boca Raton: CRC Press Inc, Apr. 2014.

[11] J. Lederhos et al. "Effective kinetic inhibitors for natural gas hydrates". *Chemical Engineering Science* 51.8 (Oct. 1995), pp. 1221–1229.

[12] S. Shahnazar et al. “Structure, mechanism, and performance evaluation of natural gas hydrate kinetic inhibitors”. *Reviews in Inorganic Chemistry* 38.1 (Apr. 2018), pp. 1–19.

[13] M. Varma-Nair et al. “Thermal analysis of polymer–water interactions and their relation to gas hydrate inhibition”. *Journal of Applied Polymer Science* 103.4 (Feb. 2007), pp. 2642–2653.

[14] S. R. Davies et al. “Hydrate plug dissociation”. *AIChE Journal* 52.12 (Dec. 2006), pp. 4016–4027.

[15] J. K. Fink. *Petroleum Engineer’s Guide to Oil Field Chemicals and Fluids*. 1st ed. Saint Louis: Elsevier Science & Technology, 2011.

[16] OSPAR Commision. *OSPAR Guidelines for Completing the Harmonised Offshore Chemical Notification Format (HOCNF)*. May 2012.

[17] P. Finkle, H. D. Draper, and J. H. Hildebrand. “The theory of emulsification”. *Journal of the American Chemical Society* 45.12 (Dec. 1923), pp. 2780–2788.

[18] J. Sjoblom. *Emulsions and Emulsion Stability*. 2nd ed. Surfactant Science Series 61. Boca Raton: CRC Press, Nov. 2005.

[19] P. Fotland and K. M. Askvik. “Some aspects of hydrate formation and wetting”. en. *Journal of Colloid and Interface Science* 321.1 (May 2008), pp. 130–141.

[20] S. Høiland et al. “Wettability of Freon hydrates in crude oil/brine emulsions”. *Journal of Colloid and Interface Science* 287.1 (July 2005), pp. 217–225.

[21] A. G. Marshall and R. P. Rodgers. “Petroleomics: The Next Grand Challenge for Chemical Analysis”. *Accounts of Chemical Research* 37.1 (Jan. 2004), pp. 53–59.

[22] S. Chiaberge et al. “Classification of crude oil samples through statistical analysis of APPI FTICR mass spectra”. en. *Fuel Processing Technology* 106 (Feb. 2013), pp. 181–185.

[23] J. G. Speight. *The Chemistry and Technology of Petroleum*. 4th ed. Chemical Industries 114. Boca Raton, FL: CRC Press, 2006.

[24] D. M. Jewell et al. “Ion-Exchange, Coordination, and Adsorption Chromatographic Separation of Heavy-End Petroleum Distillates”. *Analytical Chemistry* 44.8 (July 1972), pp. 1391–1395.

[25] H. Groenzin and O. C. Mullins. “Molecular size and structure of asphaltenes from various sources”. *Energy Fuels* 14.3 (Mar. 2000), pp. 677–684.

[26] O. C. Mullins et al. *Asphaltenes, Heavy Oils, and Petroleomics*. 1st ed. Springer-Verlag New York, 2007.

[27] J. Sjöblom et al. “Our current understanding of water-in-crude oil emulsions.: Recent characterization techniques and high pressure performance”. *Advances in Colloid and Interface Science* 100-102 (Feb. 2003), pp. 399–473.

[28] L. Stasiuk and L. Snowdon. “Fluorescence micro-spectrometry of synthetic and natural hydrocarbon fluid inclusions: crude oil chemistry, density and application to petroleum migration”. *Applied Geochemistry* 12.3 (May 1997), pp. 229–241.

[29] M. M. Boduszynski. “Composition of heavy petroleums. 1. Molecular weight, hydrogen deficiency, and heteroatom concentration as a function of atmospheric equivalent boiling point up to 1400. °F (760 °C)”. *Energy Fuels* 1.1 (Jan. 1987), pp. 2–11.

[30] J. J. Adams. “Asphaltene Adsorption, a Literature Review”. *Energy Fuels* 28.5 (Mar. 2014), pp. 2831–2856.

[31] M. L. Chacón-Patiño et al. “Advances in Asphaltene Petroleomics. Part 4. Compositional Trends of Solubility Subfractions Reveal that Polyfunctional Oxygen-Containing Compounds Drive Asphaltene Chemistry”. *Energy Fuels* 34.3 (Mar. 2020), pp. 3013–3030.

[32] J. S. Clemente and P. M. Fedorak. “A review of the occurrence, analyses, toxicity, and biodegradation of naphthenic acids”. *Chemosphere* 60.5 (July 2005), pp. 585–600.

[33] M. P. Barrow et al. “Data Visualization for the Characterization of Naphthenic Acids within Petroleum Samples”. *Energy Fuels* 23.5 (Mar. 2009), pp. 2592–2599.

[34] C. Hurtevent et al. “Production Issues of Acidic Petroleum Crude Oils”. *Emulsions and Emulsion Stability*. 2nd ed. CRC Press, 2005, p. 40.

[35] C. Yang et al. “Characterization of naphthenic acids in crude oils and refined petroleum products”. *Fuel* 255 (Nov. 2019), p. 115849.

[36] H. Magnusson, A.-M. D. Hanneseth, and J. Sjöblom. “Characterization of C80 Naphthenic Acid and Its Calcium Naphthenate”. *Journal of Dispersion Science and Technology* 29.3 (Feb. 2008), pp. 464–473.

[37] A. Bertheussen, S. C. Simon, and J. Sjöblom. “Equilibrium partitioning of naphthenic acids and bases and their consequences on interfacial properties”. *Colloids and Surfaces A: Physicochemical and Engineering Aspects* 529 (Sept. 2017), pp. 45–56.

[38] F. H. Fadnes. “Natural hydrate inhibiting components in crude oils”. *Fluid Phase Equilibria* 117.1-2 (Mar. 1996), pp. 186–192.

[39] A. E. Borgund et al. “Molecular analysis of petroleum derived compounds that adsorb onto gas hydrate surfaces”. *Applied Geochemistry* 24.5 (Jan. 2009), pp. 777–786.

[40] S. Høiland et al. “Wettability of Freon hydrates in crude oil/brine emulsions: the effects of chemical additives.” *5th International Conference in Gas Hydrate*. Vol. 4. Trondheim, Norway, June 2005, pp. 1151–1161.

[41] A. E. Borgund, K. Erstad, and T. Barth. “Fractionation of Crude Oil Acids by HPLC and Characterization of Their Properties and Effects on Gas Hydrate Surfaces”. *Energy Fuels* 21.5 (July 2007), pp. 2816–2826.

[42] P. V. Hemmingsen et al. “Structural Characterization and Interfacial Behavior of Acidic Compounds Extracted from a North Sea Oil”. *Energy Fuels* 20.5 (July 2006), pp. 1980–1987.

[43] P. V. Hemmingsen et al. “Hydrate Plugging Potential of Original and Modified Crude Oils”. *Journal of Dispersion Science and Technology* 28.3 (Apr. 2007), pp. 371–382.

[44] L. Bergfløtd et al. “Chemical Influence on the Formation, Agglomeration, and Natural Transportability of Gas Hydrates. A Multivariate Component Analysis”. *Journal of Dispersion Science and Technology* 25.3 (2004), pp. 355–365.

[45] K. Erstad et al. “Influence of Petroleum Acids on Gas Hydrate Wettability”. *Energy Fuels* 23.4 (Feb. 2009), pp. 2213–2219.

[46] M.-H. Ese and P. K. Kilpatrick. “Stabilization of Water-in-Oil Emulsions by Naphthenic Acids and Their Salts: Model Compounds, Role of pH, and Soap:Acid Ratio”. *Journal of Dispersion Science and Technology* 25.3 (2004), pp. 253–261.

[47] T. Barth et al. “Acidic compounds in biodegraded petroleum”. *Organic Geochemistry* 35.11-12 (Dec. 2004), pp. 1513–1525.

[48] G. Aspenes et al. “The influence of petroleum acids and solid surface energy on pipeline wettability in relation to hydrate deposition”. *Journal of Colloid and Interface Science* 333.2 (May 2009), pp. 533–539.

[49] S. Skiba et al. “Impact of biodegradation of oil on the kinetics of gas hydrate formation and decomposition”. *Journal of Petroleum Science and Engineering* 192 (Mar. 2020), p. 107211.

[50] P. Qiao et al. “Fractionation of Asphaltenes in Understanding Their Role in Petroleum Emulsion Stability and Fouling”. *Energy Fuels* 31.4 (Dec. 2016), pp. 3330–3337.

[51] D. C. Salmin. “The Impact of Synthetic and Natural Surface-Active Components on Hydrate Agglomeration”. en. Doctoral thesis. Golden, Colorado: Colorado School of Mines, 2019.

[52] P. K. Kilpatrick. “Water-in-Crude Oil Emulsion Stabilization: Review and Unanswered Questions”. *Energy Fuels* 26.7 (May 2012), pp. 4017–4026.

[53] F. Yang et al. “Asphaltene Subfractions Responsible for Stabilizing Water-in-Crude Oil Emulsions. Part 2: Molecular Representations and Molecular Dynamics Simulations”. *Energy Fuels* 29.8 (July 2015), pp. 4783–4794.

[54] R. G. Brereton et al. “Chemometrics in analytical chemistry—part II: modeling, validation, and applications”. *Analytical and Bioanalytical Chemistry* 410.26 (Aug. 2018), pp. 6691–6704.

[55] Z. S. Baird and V. Oja. “Predicting fuel properties using chemometrics: a review and an extension to temperature dependent physical properties by using infrared spectroscopy to predict density”. *Chemometrics and Intelligent Laboratory Systems* 158 (Nov. 2016), pp. 41–47.

[56] E. L. Gjelsvik, M. Fossen, and K. Tøndel. “Current overview and way forward for the use of machine learning in the field of petroleum gas hydrates”. *Fuel* 334.2 (Feb. 2023), p. 126696.

[57] S. A. S. AlRyalat, L. W. Malkawi, and S. M. Momani. “Comparing Bibliometric Analysis Using PubMed, Scopus, and Web of Science Databases”. *Journal of Visualized Experiments* 152 (Oct. 2019), p. 12.

[58] M. Fossen, M. Wolden, and A. Brunsvik. “Successive accumulation of naturally occurring hydrate active components and the effect on the wetting properties”. *32nd Oil Field Chemistry Symposium 2021*. TEKNA, May 2021, p. 16.

[59] E. L. Gjelsvik et al. “Using machine learning-based variable selection to identify hydrate related components from FT-ICR MS spectra”. *PLoS One* 17.8 (Aug. 2022), e0273084.

[60] E. L. Gjelsvik et al. “Identifying components related to hydrate formation using machine learning-based variable selection”. *33rd Oil Field Chemistry Symposium*. Geilo, Norway: TEKNA, Feb. 2022.

[61] K. Tøndel et al. “Hierarchical Cluster-based Partial Least Squares Regression (HC-PLSR) is an efficient tool for metamodelling of nonlinear dynamic models”. *BMC Systems Biology* 5 (June 2011), Article 90.

[62] E. de Hoffmann and V. Stroobant. *Mass spectrometry: principles and applications*. 3rd ed. West Sussex, England: John Wiley and Sons Ltd., Oct. 2012.

[63] M. B. Comisarow and A. G. Marshall. “Fourier transform ion cyclotron resonance spectroscopy”. *Chemical Physics Letters* 25.2 (Mar. 1974), pp. 282–283.

[64] Y. Cho et al. “Developments in FT-ICR MS instrumentation, ionization techniques, and data interpretation methods for petroleomics”. *Mass Spectrometry Reviews* 34.2 (Mar. 2014), pp. 248–263.

[65] M. R. Emmett et al. “Application of micro-electrospray liquid chromatography techniques to FT-ICR MS to enable high-sensitivity biological analysis”. *Journal of the American Society for Mass Spectrometry* 9.4 (Apr. 1998), pp. 333–340.

[66] C. A. Hughey et al. “Kendrick Mass Defect Spectrum: A Compact Visual Analysis for Ultrahigh-Resolution Broadband Mass Spectra”. *Analytical Chemistry* 73.19 (Aug. 2001), pp. 4676–4681.

[67] A. G. Marshall and C. L. Hendrickson. “Fourier transform ion cyclotron resonance detection: principles and experimental configurations”. *International Journal of Mass Spectrometry* 215.1-3 (Apr. 2002), pp. 59–75.

[68] A. G. Marshall, C. L. Hendrickson, and G. S. Jackson. “Fourier transform ion cyclotron resonance mass spectrometry: A primer”. *Mass Spectrometry Reviews* 17.1 (Dec. 1998), pp. 1–35.

[69] A. G. Marshall and C. L. Hendrickson. “Charge reduction lowers mass resolving power for isotopically resolved electrospray ionization Fourier transform ion cyclotron resonance mass spectra”. *Rapid Communications in Mass Spectrometry* 15.3 (Feb. 2001), pp. 232–235.

[70] J. B. Fenn et al. “Electrospray ionization-principles and practice”. *Mass Spectrometry Reviews* 9.1 (Jan. 1990), pp. 37–70.

[71] D. Williams and I. Fleming. *Spectroscopic methods in organic chemistry*. 6th ed. UK: McGraw-Hill Education, 2008.

[72] J. M. Purcell et al. “Atmospheric Pressure Photoionization Fourier Transform Ion Cyclotron Resonance Mass Spectrometry for Complex Mixture Analysis”. *Analytical Chemistry* 78.16 (July 2006), pp. 5906–5912.

[73] T. J. Kauppila, J. A. Syage, and T. Benter. “Recent developments in atmospheric pressure photoionization-mass spectrometry”. *Mass Spectrometry Reviews* 36.3 (May 2015), pp. 423–449.

[74] T. Fujii. *Ion/Molecule Attachment Reactions: Mass Spectrometry*. 1st ed. New York, NY: Springer, 2015.

[75] M. Fossen et al. “Solubility Parameters Based on IR and NIR Spectra: I. Correlation to Polar Solutes and Binary Systems”. *Journal of Dispersion Science and Technology* 26.2 (Sept. 2004), pp. 227–241.

[76] J. Workman Jr. and Lois Weyer. “History of Near-Infrared Applications”. *Practical Guide and Spectral Atlas for Interpretive Near-Infrared Spectroscopy*. 2nd ed. CRC Press, 2012, p. 6.

[77] N. Aske, H. Kallevik, and J. Sjöblom. “Determination of Saturate, Aromatic, Resin, and Asphaltenic (SARA) Components in Crude Oils by Means of Infrared and Near-Infrared Spectroscopy”. *Energy Fuels* 15.5 (Aug. 2001), pp. 1304–1312.

[78] J. J. Kelly and James B. Callis. “Nondestructive analytical procedure for simultaneous estimation of the major classes of hydrocarbon constituents of finished gasolines”. *Analytical Chemistry* 62.14 (July 1990), pp. 1444–1451.

[79] J. Laxalde et al. “Characterisation of heavy oils using near-infrared spectroscopy: Optimisation of pre-processing methods and variable selection”. *Analytica Chimica Acta* 705.1-2 (Oct. 2011), pp. 227–234.

[80] H. Chung. “Applications of Near-Infrared Spectroscopy in Refineries and Important Issues to Address”. *Applied Spectroscopy Reviews* 42.3 (May 2007), pp. 251–285.

[81] Å. Rinnan, F. van den Berg, and S. B. Engelsen. “Review of the most common pre-processing techniques for near-infrared spectra”. *TrAC Trends in Analytical Chemistry* 28.10 (Nov. 2009), pp. 1201–1222.

[82] H. Martens, J. P. Nielsen, and S. B. Engelsen. “Light Scattering and Light Absorbance Separated by Extended Multiplicative Signal Correction. Application to Near-Infrared Transmission Analysis of Powder Mixtures”. *Analytical Chemistry* 75.3 (Feb. 2003), pp. 394–404.

[83] H. Martens, S. Jensen, and P. Geladi. “Multivariate linearity transformations for near infrared reflectance spectroscopy”. *Nordic Symposium on Applied Statistics*. Stavanger, Norway: Stokkand Forlag Publishers, June 1983, pp. 205–234.

[84] P. Geladi, D. MacDougall, and H. Martens. “Linearization and Scatter-Correction for Near-Infrared Reflectance Spectra of Meat”. *Applied Spectroscopy* 39.3 (1985), pp. 491–500.

[85] H. Martens and E. Stark. “Extended multiplicative signal correction and spectral interference subtraction: New preprocessing methods for near infrared spectroscopy”. *Journal of Pharmaceutical and Biomedical Analysis* 9.8 (1991), pp. 625–635.

[86] A. Kohler et al. “Estimating and Correcting Mie Scattering in Synchrotron-Based Microscopic Fourier Transform Infrared Spectra by Extended Multiplicative Signal Correction”. *Applied Spectroscopy* 62.3 (2008), pp. 259–266.

[87] I. S. Helland, T. Næs, and T. Isaksson. “Related versions of the multiplicative scatter correction method for preprocessing spectroscopic data”. *Chemometrics and Intelligent Laboratory Systems* 29.2 (Oct. 1995), pp. 233–241.

[88] D. K. Pedersen et al. “Near-Infrared Absorption and Scattering Separated by Extended Inverted Signal Correction (EISC): Analysis of Near-Infrared Transmittance Spectra of Single Wheat Seeds”. *Applied Spectroscopy* 56.9 (Sept. 2002), pp. 1206–1214.

[89] A. Savitzky and M. J. E. Golay. “Smoothing and Differentiation of Data by Simplified Least Squares Procedures”. *Analytical Chemistry* 36.8 (July 1964), pp. 1627–1639.

[90] J.-M. Roger, A. Biancolillo, and F. Marini. “Sequential preprocessing through ORThogonalization (SPORT) and its application to near infrared spec-

troscopy". *Chemometrics and Intelligent Laboratory Systems* 199 (Apr. 2020), p. 103975.

[91] P. Mishra et al. "New data preprocessing trends based on ensemble of multiple preprocessing techniques". *TrAC Trends in Analytical Chemistry* 132 (Nov. 2020), p. 116045.

[92] O. C. Mullins. "Asphaltenes in crude oil: absorbers and/or scatterers in the near-infrared region?" *Analytical Chemistry* 62.5 (Mar. 1990), pp. 508–514.

[93] A. G. Marshall and R. P. Rodgers. "Petroleomics: Chemistry of the underworld". *Proceedings of the National Academy of Sciences of the United States of America* 105.47 (Nov. 2008), pp. 18090–18095.

[94] M. P. Barrow et al. "Athabasca Oil Sands Process Water: Characterization by Atmospheric Pressure Photoionization and Electrospray Ionization Fourier Transform Ion Cyclotron Resonance Mass Spectrometry". *Analytical Chemistry* 82.9 (May 2010), pp. 3727–3735.

[95] C. M. Bishop. *Pattern Recognition and Machine Learning*. 1st ed. Information Science and Statistics. New York, NY: Springer, 2006.

[96] T. M. Mitchell. *Machine Learning*. 1st ed. McGraw-Hill Series in Computer Science. Artificial Intelligence. McGraw-Hill Education, Oct. 1997.

[97] F. Nielsen. "Hierarchical Clustering". *Introduction to HPC with MPI for Data Science*. 1st ed. Undergraduate Topics in Computer Science. Springer, Feb. 2016, pp. 195–211.

[98] J. MacQueen. "Some methods for classification and analysis of multivariate observations". *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics* Vol. 1 (1967), pp. 281–297.

[99] S. Lloyd. "Least squares quantization in PCM". *IEEE Transactions on Information Theory* 28.2 (Mar. 1982), pp. 129–137.

[100] J. C. Dunn. "A Fuzzy Relative of the ISODATA Process and Its Use in Detecting Compact Well-Separated Clusters". *Journal of Cybernetics* 3.3 (Sept. 1973), pp. 32–57.

[101] J. C. Bezdek. *Pattern Recognition with Fuzzy Objective Function Algorithms*. 1st ed. Advanced Applications in Pattern Recognition. Springer US, 1981.

[102] M. Ester et al. "A density-based algorithm for discovering clusters in large spatial databases with noise". *Second International Conference on Knowledge Discovery and Data Mining*. Portland, Oregon, USA: AAAI Press, Aug. 1996, pp. 226–231.

[103] U. von Luxburg. "A tutorial on spectral clustering". *Statistics and Computing* 17 (Aug. 2007), pp. 395–416.

[104] K. Pearson. "On lines and planes of closest fit to systems of points in space". *Philosophical Magazine* 2 (1901), pp. 559–572.

[105] S. Wold, H. Martens, and H. Wold. “The multivariate calibration problem in chemistry solved by the PLS method”. *Matrix Pencils*. Lecture Notes in Mathematics 973. Berlin, Heidelberg: Springer, 1983, pp. 286–293.

[106] A. Höskuldsson. “PLS regression methods”. *Journal of Chemometrics* 2.3 (June 1988), pp. 211–228.

[107] C. Cortes and V. Vapnik. “Support-vector networks”. *Machine Learning* 20 (Sept. 1995), pp. 273–297.

[108] C. J. Burges. “A Tutorial on Support Vector Machines for Pattern Recognition”. *Data Mining and Knowledge Discovery* 2 (June 1998), pp. 121–167.

[109] J. R. Quinlan. “Simplifying decision trees”. *International Journal of Man-Machine Studies* 27.3 (Sept. 1987), pp. 221–234.

[110] P. E. Utgoff. “Incremental Induction of Decision Trees”. *Machine Learning* 4 (Nov. 1989), pp. 161–186.

[111] C. E. Brodley and P. E. Utgoff. “Multivariate Decision Trees”. *Machine Learning* 19 (Apr. 1995), pp. 45–77.

[112] L. Breiman. “Arcing classifier (with discussion and a rejoinder by the author)”. *Annals of Statistics* 26.3 (1998), pp. 801–849.

[113] T. Chen and C. E. Guestrin. “XGBoost: A Scalable Tree Boosting System”. *KDD ’16: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. San Francisco, CA, USA: Association for Computing Machinery, Aug. 2016, pp. 785–794.

[114] L. Breiman. “Bagging predictors”. *Machine Learning* 24 (Aug. 1996), pp. 123–140.

[115] L. Breiman. “Random Forests”. *Machine Learning* 45 (Oct. 2001), pp. 5–32.

[116] T. K. Ho. “The random subspace method for constructing decision forests”. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 20.8 (Aug. 1998), pp. 832–844.

[117] A. E. Hoerl. “Application of Ridge Analysis to Regression Problems”. *Chemical Engineering Progress* 58.3 (1958), pp. 54–59.

[118] R. Tibshirani. “Regression Shrinkage and Selection Via the Lasso”. *Journal of the Royal Statistical Society: Series B (Methodological)* 58.1 (1996), pp. 267–288.

[119] H. Zou and T. Hastie. “Regularization and variable selection via the elastic net”. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 67.2 (Apr. 2005), pp. 301–320.

[120] P. Mishra et al. “Recent trends in multi-block data analysis in chemometrics for multi-source data integration”. *TrAC Trends in Analytical Chemistry* 137 (Apr. 2021), p. 116206.

[121] A. K. Smilde, T. Næs, and K. H. Liland. *Multiblock Data Fusion in Statistics and Machine Learning: Applications in the Natural and Life Sciences*. 1st ed. John Wiley & Sons, Ltd, Apr. 2022.

[122] A. K. Smilde et al. "Common and distinct components in data fusion". *Journal of Chemometrics* 31.7 (July 2017), e2900.

[123] L. E. Wangen and B. R. Kowalski. "A multiblock partial least squares algorithm for investigating complex chemical systems". *Journal of Chemometrics* 3.1 (Jan. 1989), pp. 3–20.

[124] J. A. Westerhuis, T. Kourti, and J. F. MacGregor. "Analysis of multiblock and hierarchical PCA and PLS models". *Journal of Chemometrics* 12.5 (Dec. 1998), pp. 301–321.

[125] J. A. Westerhuis and A. K. Smilde. "Deflation in multiblock PLS". *Journal of Chemometrics* 15.5 (June 2001), pp. 485–493.

[126] K. Jørgensen, B.-H. Mevik, and T. Næs. "Combining designed experiments with several blocks of spectroscopic data". *Chemometrics and Intelligent Laboratory Systems* 88.2 (Sept. 2007), pp. 154–166.

[127] T. Næs et al. "Path modelling by sequential PLS regression". *Journal of Chemometrics* 25.1 (Jan. 2011), pp. 28–40.

[128] A. Biancolillo and T. Næs. "The sequential and orthogonalized PLS regression for multiblock regression: theory, examples, and extensions". *Data Fusion Methodology and Applications*. Vol. 31. Data Handling in Science and Technology. Elsevier, 2019, pp. 157–177.

[129] C. M. Bishop. *Neural networks for pattern recognition*. 1st. Advanced Texts in Econometrics. 198 Madison Ave. New York, NY, United States: Oxford University Press, Inc., Nov. 1995.

[130] T. Udelhoven, D. Naumann, and J. Schmitt. "Development of a Hierarchical Classification System with Artificial Neural Networks and FT-IR Spectra for the Identification of Bacteria". *Applied Spectroscopy* 54.10 (Oct. 2000).

[131] T. Udelhoven, M. Novozhilovx, and J. Schmitt. "The NeuroDeveloper®: a tool for modular neural classification of spectroscopic data". *Chemometrics and Intelligent Laboratory Systems* 66.2 (June 2003), pp. 219–226.

[132] D. E. Rumelhart, G. E. Hinton, and R. J. Williams. "Learning representations by back-propagating errors". *Nature* 323 (Sept. 1986), pp. 533–536.

[133] D. E. Rumelhart, G. E. Hinton, and R. J. Williams. "8. Learning Internal Representations by Error Propagation". *Parallel Distributed Processing : Explorations in the Microstructure of Cognition*. 11th ed. Vol. Vol 1: Foundations. Cambridge: MIT: Bradford Books, 1986.

[134] Y. LeCun, Y. Bengio, and G. Hinton. "Deep learning". *Nature* 521 (May 2015), pp. 436–444.

[135] J. Schmidhuber. “Deep learning in neural networks: An overview”. *Neural Networks* 61 (Jan. 2015), pp. 85–117.

[136] Y. LeCun et al. “Gradient-based learning applied to document recognition”. *Proceedings of the IEEE* 86.11 (Nov. 1998), pp. 2278–2324.

[137] D. P. Mandic and J. A. Chambers. *Recurrent Neural Networks for Prediction: Learning Algorithms, Architectures and Stability*. Wiley, Aug. 2001.

[138] I. H. Sarker. “Deep Learning: A Comprehensive Overview on Techniques, Taxonomy, Applications and Research Directions”. *SN Computer Science* 2 (Aug. 2021), p. 420.

[139] T. Mehmood et al. “A review of variable selection methods in Partial Least Squares Regression”. *Chemometrics and Intelligent Laboratory Systems* 118 (Aug. 2012), pp. 62–69.

[140] T. N. Tran et al. “Interpretation of variable importance in Partial Least Squares with Significance Multivariate Correlation (sMC)”. *Chemometrics and Intelligent Laboratory Systems* 138 (Nov. 2014), pp. 153–160.

[141] S. Høiland, P. Glénat, and K. M. Askvik. “Wetting Index : A Quantitative Measure Of Indigenous Hydrate Plugging Tendency ; Flow Test Validations.” *7th International Conference on Gas Hydrates*. Edinburgh, UK, July 2011.

[142] W. D. Bancroft. “The Theory of Emulsification, V”. *Journal of Physical Chemistry* 17.6 (June 1913), pp. 501–519.

[143] Bruker Daltonik GmbH. *Bruker Compass ProfileAnalysis*. 2013.

[144] T. De Meyer et al. “NMR-Based Characterization of Metabolic Alterations in Hypertension Using an Adaptive, Intelligent Binning Algorithm”. *Analytical Chemistry* 50.10 (May 2008), pp. 3783–3790.

[145] J. M. Miller. *Chromatography: Concepts and Contrasts*. 2nd ed. Hoboken, New Jersey: Wiley, Aug. 2009.

[146] *PerkinElmer Frontier FT-IR, NIR and FIR Spectroscopy Brochure*. 2011.

[147] M. K. Moro et al. “A review of chemometrics models to predict crude oil properties from nuclear magnetic resonance and infrared spectroscopy”. *Fuel* 303 (Nov. 2021), p. 121283.

[148] E. V. Barros et al. “Characterization of naphthenic acids in crude oil samples – A literature review”. *Fuel* 319 (July 2022), p. 123775.

[149] S. Lababidi and W. Schrader. “Online normal-phase high-performance liquid chromatography/Fourier transform ion cyclotron resonance mass spectrometry: Effects of different ionization methods on the characterization of highly complex crude oil mixtures”. *Rapid Communications in Mass Spectrometry* 28.12 (June 2014), pp. 1345–1352.

[150] S. Chiaberge et al. “Bio-Oil from Waste: A Comprehensive Analytical Study by Soft-Ionization FTICR Mass Spectrometry”. *Energy Fuels* 28.3 (Mar. 2014), pp. 2019–2026.

[151] J. Adebayo et al. “Local Explanation Methods for Deep Neural Networks Lack Sensitivity to Parameter Values”. *arXiv* (Oct. 2018), p. 1810.03307.

[152] J. Adebayo et al. “Sanity Checks for Saliency Maps”. *Advances in Neural Information Processing Systems 31*. Montréal, Canada, 2018, pp. 9505–9515.

Papers

Paper I

Gjelsvik E. L., Fossen M., Tøndel K. (2023) Current overview and way forward for the use of machine learning in the field of petroleum gas hydrates *Fuel* 334(Part 2), 126696, 10.1016/j.fuel.2022.126696

Fuel 334 (2023) 126696

Contents lists available at ScienceDirect

Fuel

journal homepage: www.elsevier.com/locate/fuel

Review article

Current overview and way forward for the use of machine learning in the field of petroleum gas hydrates

Elise Lunde Gjelsvik [a,*], Martin Fossen [b], Kristin Tøndel [a]

[a] Norwegian University of Life Sciences, Faculty of Science and Technology, Ås, Norway
[b] SINTEF AS, Trondheim, Norway

ARTICLE INFO

Keywords:
Gas hydrates
Machine learning
FT-ICR MS
Chemometrics
Crude oil

ABSTRACT

Gas hydrates represent one of the main flow assurance challenges in the oil and gas industry as they can lead to plugging of pipelines and process equipment. In this paper we present a literature study performed to evaluate the current state of the use of machine learning methods within the field of gas hydrates with specific focus on the oil chemistry. A common analysis technique for crude oils is Fourier Transform Ion Cyclotron Resonance Mass Spectrometry (FT-ICR MS) which could be a good approach to achieving a better understanding of the chemical composition of hydrates, and the use of machine learning in the field of FT-ICR MS was therefore also examined. Several machine learning methods were identified as promising, their use in the literature was reviewed and a text analysis study was performed to identify the main topics within the publications. The literature search revealed that the publications on the combination of FT-ICR MS, machine learning and gas hydrates is limited to one. Most of the work on gas hydrates is related to thermodynamics, while FT-ICR MS is mostly used for chemical analysis of oils. However, with the combination of FT-ICR MS and machine learning to evaluate samples related to gas hydrates, it could be possible to improve the understanding of the composition of hydrates and thereby identify hydrate active compounds responsible for the differences between oils forming plugging hydrates and oils forming transportable hydrates.

1. Introduction

Gas hydrates are crystalline structures where smaller guest molecules are trapped in cages formed by water molecules that are held together by hydrogen bonds [1]. Gas hydrates are among the main flow assurance issues when producing oil and gas, especially subsea or in cold locations, because they can lead to complete blockage (plugging) of pipelines and process equipment forcing the operator to shut down the production [2]. The most common, yet very conservative, hydrate strategy states that the positive driving forces for hydrate formation, i.e. high pressure and low temperature, should be avoided. In practice this requires determination of the thermodynamic region where hydrate formation occurs in order to keep the system outside this pressure–temperature region. [3].

For hydrate inhibition on the other hand, the most common strategy is currently the use of thermodynamic inhibitors (THIs). These inhibitors shift the hydrate curve towards higher pressures at hydrate inducing temperatures, enabling production at lower temperatures without the formation of gas hydrates [4,5]. Common inhibitors are organic chemicals, such as methanol and monoethylene glycol (MEG) dosed at concentrations of 20%–50% of the mass relative to the water produced [4]. The premise of their application is that gas hydrate formation is expected, and therefore the inhibitors are always present in the pipelines. Another promising strategy for hydrate management is the injection of low dose hydrate inhibitors (LDHI) [6]. The two main types of LDHIs are the kinetic hydrate inhibitors (KHI) which alter the kinetics during the hydrate formation, and the anti-agglomerants (AAs) which alter wettability of the hydrate particles and prevent them from sticking together. A typical concentration for an LDHI injection is 0.1–1 wt % relative to the water phase [4,7]. For the AAs the purpose is to form a slurry of gas hydrates dispersed in the oil phase that can be transported through the pipelines without the particles aggregating together or depositing to the pipe wall. However, for an AA to be efficient, it must be surface active and able to adsorb to the surface or interact with the hydrate cages of the dispersed hydrate particles. The purpose of KHIs, on the other hand, is to delay the formation of hydrates long enough to reach the storage facility without causing blockage [8]. The KHI binds to the hydrate surface, decreasing the crystal formation process by preventing the growth of hydrate crystals nuclei [9].

* Corresponding author.
E-mail address: elise.lunde.gjelsvik@nmbu.no (E.L. Gjelsvik).

https://doi.org/10.1016/j.fuel.2022.126696
Received 22 July 2022; Received in revised form 12 October 2022; Accepted 7 November 2022
Available online 21 November 2022

However, through laboratory experiments spurred by field experience, it became evident that some crude oils did not experience plugging when gas hydrates were formed [10]. Instead, the hydrates behaved more like dry particles that could be transported without any issues [11]. The explanation set forth was that some crude oils contain naturally occurring components that interact with the gas hydrates rendering the surface of the particles hydrophobic. One hypothesis is that these components have the ability to adsorb to the hydrate surface, preventing agglomeration of hydrates and the potential plugging of the pipeline [12]. Another hypothesis is that parts of a molecule, for example butyl/pentyl groups, penetrate open cavities on the hydrate surface (of $5^{12}\,6^4$ SII cages) and can become embedded in the surface as the hydrate grows around the alkyl groups [4]. The current status of the search for the type and structures of natural hydrate inhibitors is that they have not yet been characterised in detail [2,11–13]. Some previous studies have suggested that these natural inhibitors may be contained in the petroleum acid fraction [11,14–17] which has been shown to include a large amount of naphthenic compounds. Borgund et al. [15] and Erstad et al. [18] showed experimentally the anti-agglomerating properties of some petroleum acid fractions.

Similarly, the asphaltene fractions are known to possess self-agglomerating properties that can stabilise some crude oil systems [19] and some asphaltenes can alter the plugging potential of crude oils [20,21]. It has been shown that the asphaltene fractions able to stabilise systems prone to form transportable slurries are often more polar, with higher oxygen content, higher acidity and lower double bound equivalents (DBEs) [22]. Other studies have suggested that the possible hydrate activity of asphaltenes is related to their sulfoxide content [23].

The overall goal of this review was to establish a baseline for the current status of the use of machine learning in the field of petroleum gas hydrates. A part of this study was to identify work related to naturally occurring hydrate inhibitors in crude oils where machine learning methods have been used. It was, however, shown that this research was extremely limited, resulting in only one publication [24]. Therefore, the methodologies described are related to the thermodynamic aspects of gas hydrates and the chemical analysis of crude oils. Fourier Transform Ion Cyclotron Mass Spectrometry (FT-ICR MS) has a high mass accuracy which could be utilised for analysis of properties related to gas hydrates. FT-ICR MS was therefore included in this review to establish a link between aspects of gas hydrates and analysis of crude oils.

2. Fourier Transform Ion Cyclotron Resonance Mass Spectrometry (FT-ICR MS)

The complex mixtures of crude oils and the relatively high masses of their components make detailed identification difficult with most mass spectrometers. However, with the high mass accuracy of FT-ICR MS, more detailed analysis of crude oil samples are possible [25,26]. In FT-ICR MS the mass-to-charge (m/z) ratio of ions are determined based on the cyclotron frequency of the ions in a fixed magnetic field. The mass accuracy for FT-ICR MS is sub ppm and the mass spectral resolution can be above 10 million (at m/z = 400), which allows identification of a large number of different polar groups [27–29]. In an FT-ICR MS analysis, ions are detected simultaneously within a detecting interval by the ion cyclotron resonance frequency they produce when they rotate in a magnetic field. This provides the increase in signal-to-noise ratio compared to traditional mass spectrometers.

There are several different ionisation techniques to be used in combination with FT-ICR MS. For crude oils, the most common are electrospray ionisation (ESI) and atmospheric pressure photo ionisation (APPI) as they ionise polar compounds efficiently [27,30]. ESI is achieved by applying a high voltage to a liquid passing through a capillary tube inducing highly charged droplets [31]. In positive mode, formic acid is added to the solution aiding ionisation, while in negative mode ammonium hydroxide is added resulting in lower background noise. APPI is performed by exposing the analytes to photons emitted from a UV lamp [27] and in positive mode, both molecular ($[M^{+\cdot}]$) and protonated ions ($[M+H]^+$) are generated. During negative mode, the ions of the molecular species are produced by either proton abstraction or adduct formation. The predominant ions are the molecular species ion ($[M-H]^-$), which is the ion corresponding to the fatty acids ($R_n - COO^-$) present in the sample [31]. APPI is sensitive to aromatic compounds and sulphur containing compounds.

FT-ICR MS has previously been used widely for crude oil characterisation [27,32–38]. For instance, Qian et al. [39,40] showed that positive and negative mode ESI-FT-ICR MS are able to characterise different aspects of crude oils. In negative mode it was identified over 3000 chemical formulas of acids and acidic compounds, while in positive mode over 3000 unique elemental compositions of Nitrogen-Containing Aromatic Compounds were identified, illustrating the high accuracy of FT-ICR MS. The large data sets constituting FT-ICR MS spectra, require data treatment methods able to handle big data and find underlying relationships.

The objective of this review is to provide an overview of the machine learning methods used within the field of gas hydrates, with specific focus on the oil chemistry. First, we performed a text mining study to show the previous research areas of focus and expose potential gaps within. The aim of text mining is to scrape a web page of text related to a predefined keyword. We accessed all relevant publications from the Scopus Search database [41] and the most common and promising methods in literature are discussed. Additionally, methods commonly used for analysis of FT-ICR MS data in other fields which we believe could make valuable contributions to analysis of gas hydrate related samples, were identified. If correlations between hydrate-active components responsible for non-plugging crude oil systems and oil composition can be determined, this can be utilised as a parameter base for improved hydrate management strategies, better decision support tools and pipe flow simulations.

3. Text mining

To achieve an overview of the current status of machine learning methods within the field of petroleum gas hydrates the following questions were defined, of which the answers should give a thorough understanding of the field.

- **Q1**: Within which fields of gas hydrate research are machine learning used?
- **Q2**: What type of machine learning methods are used in the literature?
- **Q3**: What are the challenges in the field of gas hydrates using machine learning?
- **Q4**: How can machine learning improve the field of gas hydrate research?

3.1. Search strategy

For the text mining, we used the Scopus Search API from the pybliometrics library [42] in Python, which searches the Scopus database, containing over 78 million records within the fields of life sciences, social sciences, physical sciences and health sciences [43]. The search can be defined in different ways, searching for keywords, abstracts, title, doi, url, etc. Our approach was to search for selected words within either the keywords, titles or abstracts. To ensure that all relevant references were collected, the resulting literature was compared to results in Web of Science.

First a search was performed with the combination of *gas hydrates*, *FT-ICR MS*, *natural inhibitors* and *machine learning*, resulting in zero publications. The term *natural inhibitor* was removed and a search with *gas hydrates*, *FT-ICR MS* and *machine learning* was performed, which resulted in only one publication, a study performed by the authors of

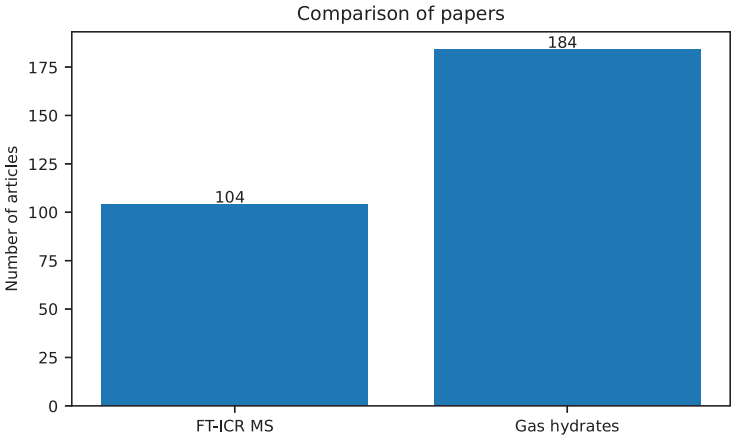


Fig. 1. Comparison of publications on the machine learning methods from Table 1 within the fields of gas hydrates and FT-ICR MS and number of publications retrieved.

Table 1
Overview of searches, each method was searched in combination with *gas hydrate* and *FT-ICR MS* to find all literature related to the methods.

Subjects	Methods
Gas hydrates	Principal Component Analysis (PCA)
FT-ICR MS	Partial Least Squares (PLS)
	Decision Trees (DT)
	Random Forest
	Artificial Neural Network (ANN)
	Support Vector Machine (SVM)
	Convolutional Neural Network (CNN)
	Regularisation/LASSO/Elastic Net/Ridge Regression
	Bayesian Networks (BN)
	K-Nearest neighbours (KNN)

this review [24]. Two new searches were therefore performed with *gas hydrates* plus *machine learning* and *FT-ICR MS* plus *machine learning*. This resulted in 45 publications for gas hydrates and 9 for FT-ICR MS. As very few publications were found, it was assumed that most publications do not use the term machine learning and only mention the methods used. Therefore, several machine learning methods were used as input in new searches. An overview of the methods included is presented in Table 1.

The resulting search phrases were as follows for gas hydrates *'TITLE-ABS-KEY((gas W/1 hydrate*) AND ((machine learning method) OR (method abbreviation)))'* and for FT-ICR MS *'TITLE-ABS-KEY((ft-icr W/1 ms) AND ((machine learning method) OR (method abbreviation)))'*. The 'W/1' ensures that the words are only one term apart and the * allows for different endings of the word, for instance *s* for plural notations. Duplicates of publications were removed.

A search was also performed for natural inhibitors with all the methods mentioned in Table 1 for both gas hydrates and FT-ICR MS, which resulted in zero publications.

To evaluate the use of mass spectrometry (MS) in the field of gas hydrates, a search was performed with *mass spectrometry* and *gas hydrates* which resulted in 2045 publications. To evaluate how many of these that were related to machine learning, a search with the methods presented in Table 1 was performed with both *mass spectrometry* and *gas hydrates*. This search resulted in 11 publications and all the 11 publications were also present in the results from the gas hydrate search with the machine learning methods.

The text mining study revealed that no other review paper exists on the topic of machine learning methods within the field of petroleum hydrates.

Text analysis was performed within the results of the two searches to find trends in the topics mentioned in the publications. The t-distributed stochastic neighbour embedding (t-SNE) technique was used to visualise the data. In t-SNE, similar data are grouped close together based on the stochastic neighbour embedding, while dissimilar data are more distant [44].

4. Results

The results from the two searches, *gas hydrates* and *FT-ICR MS*, with the methods in Table 1, are shown in Fig. 1. From the search of gas hydrates in combination with the methods from Table 1, 184 publications were retrieved and from FT-ICR MS and the methods in Table 1, 104 publications were retrieved. The publications returned by the text mining study are reported in the supporting information.

In Fig. 2 the publications on machine learning methods within the fields of gas hydrates and FT-ICR MS are plotted by publication year. Fig. 2 shows that there has been an increase in machine learning based research within both fields in the recent years. The first publication for gas hydrates was in 1998, and the first paper on use of machine learning within FT-ICR MS is from 2006. As FT-ICR MS has become more publicly available in the recent years, it is not surprising that the amount of publications have increased recently.

The amount of publications within each method is shown in Fig. 3. For gas hydrates, ANN is the most common machine learning method used, followed by SVM and PCA. For FT-ICR MS, the most common method is PCA followed by PLSR, the remaining methods have very few publications each and several of the methods had zero publications.

4.1. Text analysis study

A text analysis was performed, and a t-SNE plot of topics within the gas hydrate publications are shown with three topics in Fig. 4. The most common words for each topic are shown in the word clouds in Fig. 5. Fig. 4 shows that Topic 2 (orange) has the most entries of the three. The word clouds show that Topic 2 contains words such as 'gas', 'hydrate', 'prediction' and words associated with ANNs, 'artificial', 'neural' and 'network'. Topic 3 contains words associated with natural hydrates and some entries of 'network', while Topic 1 contains words associated with seismic and water analysis. From this analysis it is likely that the publications of interest with regards to machine learning and prediction of petroleum gas hydrates are within Topic 2 and natural gas hydrates within Topic 3.

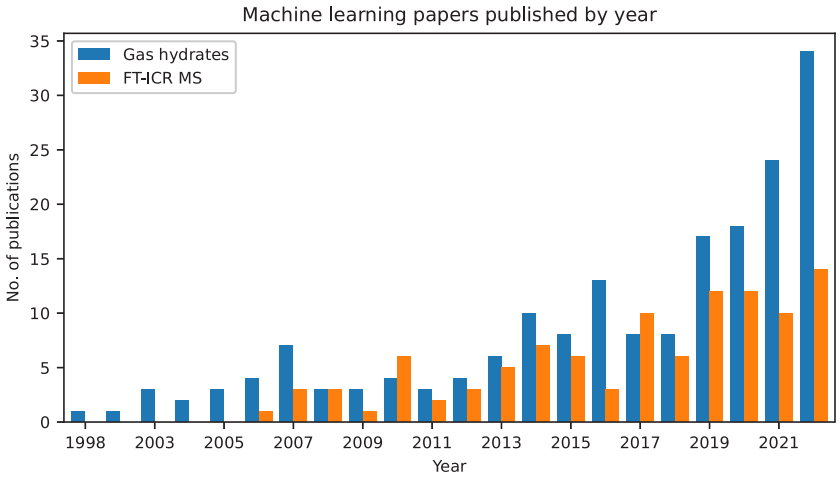


Fig. 2. The retrieved publications published by year, for gas hydrates in blue and FT-ICR MS in orange. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

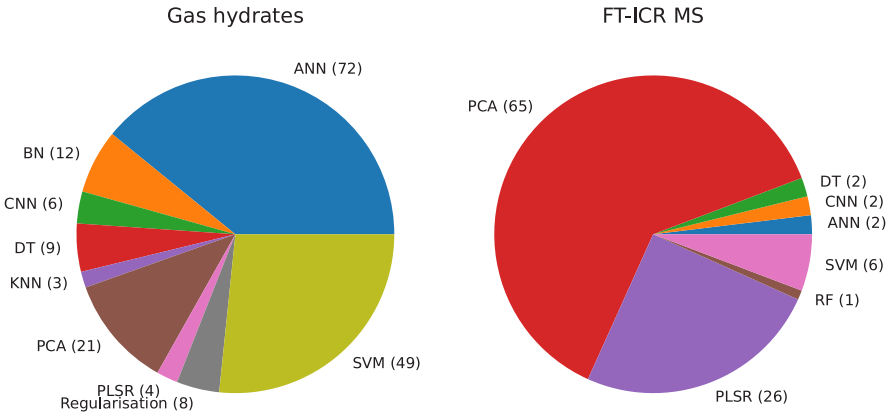


Fig. 3. Pie chart of the methods from Table 1 in combination with gas hydrates to the left and FT-ICR MS to the right with number of publications for each method in parenthesis.

A text analysis for the FT-ICR MS publications was also performed and the t-SNE plot with 3 topics is shown in Fig. 6. As t-SNE models similarities and dissimilarities, it is clear from Fig. 6 that Topic 1 (blue) is very different from Topic 2 (orange) as they are on the opposite sides of the plot, with Topic 3 (green) as a bridge between them. The most common words for each topic are shown in the word clouds in Fig. 7. Topic 1 is associated to oil spectroscopy and contains words from FT-ICR MS, Topic 2 contains words associated with organic matter analysis and Topic 3 contains metabolomic analyses. The machine learning studies performed on crude oils are therefore likely within Topic 1.

4.2. Classification vs. Regression

Machine learning can be used for analysis and visualisation of trends and allows identification of underlying phenomena in a data set. A typical pipeline for machine learning is displayed in Fig. 8. The process starts with collection of data, pre-processing, training of the model, testing of the model and finally deployment of the model through prediction from new data. The reader should seek out general textbooks for an introduction to machine learning [45,46].

Machine learning can be separated into two categories based on the desired response. When the response is continuous, regression analysis is used, while when the response is a discrete class label classification is used. Some algorithms can be used for both classification and regression tasks with only minor modifications. For gas hydrate purposes, both regression and classification methods are of interest. Which method to use is dependent on the type of data and the desired response to be predicted. For instance, when predicting thermodynamic properties of crude oils regression methods are most commonly used, as the desired prediction often is temperature, pressure or other measurements on the continuous scale. Classification methods are commonly used when samples are to be predicted based on their similarities to the defined classes. For instance when classifying oils into different types, properties etc.

In the following section, the methods included in the literature study and relevant references will be discussed to achieve an overview of the use of machine learning for analysis of petroleum related gas hydrates.

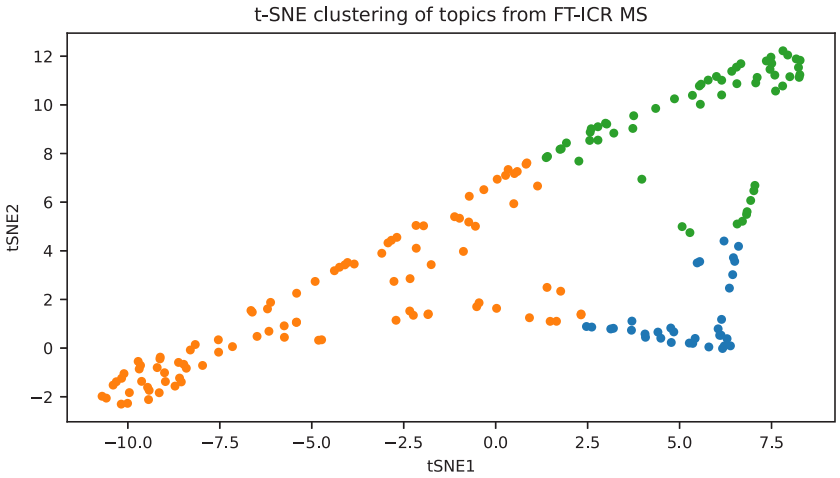


Fig. 4. t-SNE plot with three topics of the text analysis of machine learning publications on gas hydrates.

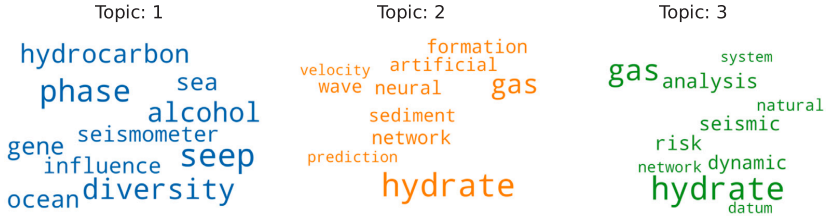


Fig. 5. Word clouds for each of the three topics and their most common words from the gas hydrate publications, with Topic 1 in blue, Topic 2 in orange and Topic 3 in green. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

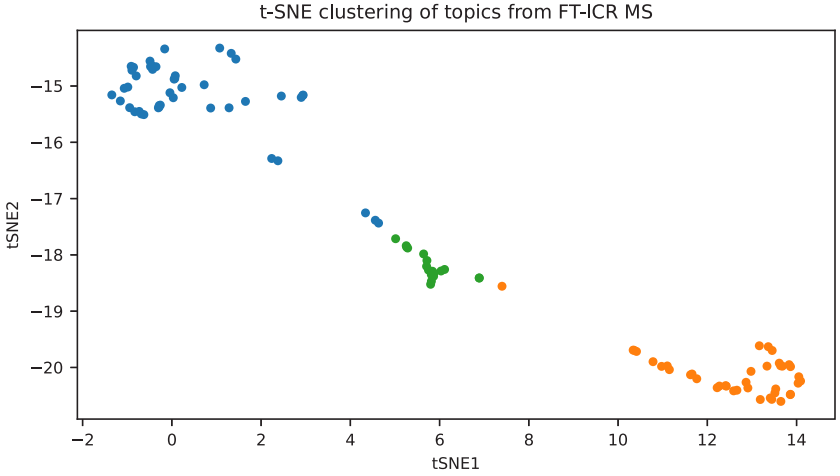


Fig. 6. t-SNE plot with three topics of the text analysis of machine learning publications on FT-ICR MS. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

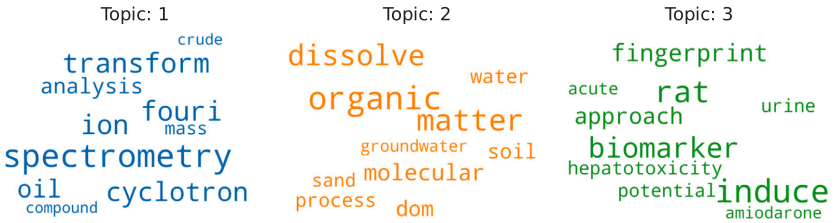


Fig. 7. Word clouds for each of the three topics and their most common words, from the FT-ICR MS publications, with Topic 1 in blue, Topic 2 in orange and Topic 3 in green. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

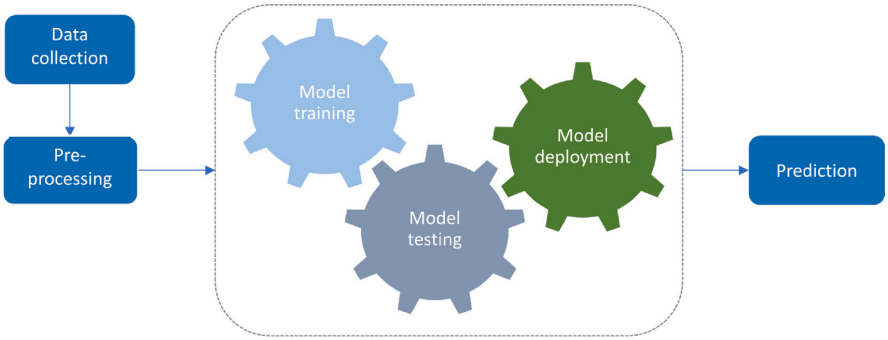


Fig. 8. Schematic illustration of a machine learning pipeline, with data collection, pre-processing, model training, testing, deployment and prediction.

4.3. Ordinary Least Squares (OLS)

OLS is a regression method for estimating the unknown parameters in a linear regression model. OLS minimises the sum of squares of the differences between the observed value and the value predicted by the linear function of the independent variable as shown by Eq. (1).

$$\boldsymbol{y} = \boldsymbol{X}\beta + \epsilon \tag{1}$$

The coefficients ($\hat{\beta}$) can be estimated from Eq. (2).

$$\hat{\beta} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y} \tag{2}$$

A major drawback with OLS regression is that the matrix inversion used in the calculation of the regression coefficients requires the regressors to be linearly independent or uncorrelated. It also requires that the number of samples is larger than the number of variables, which is most often not the case when analysing data from FT-ICR MS. This renders OLS regression unsuitable for many data analysis problems. Two commonly used strategies, outlined below, to overcome this problem are (i) use of latent variables which represent linearly independent phenomena and (ii) regularisation.

4.4. Latent variable-based methods

4.4.1. Principal Component Analysis (PCA)

PCA [47] decomposes a large data set $\boldsymbol{X}$ into a subspace of latent variables representing the main features of variance as shown by Eq. (3).

$$\boldsymbol{X} = \boldsymbol{X}_{In} wgt_X \tag{3}$$

where $\boldsymbol{X}_{\mathrm{In}}$ is the data set with shape (N, K) for N samples and K variables, and wgt_{X} are the statistical weights balancing the sum of squares for the K $\boldsymbol{X}$-variables in $\boldsymbol{X}$, which has the shape (N, K). PCA is an effective dimension reduction technique that gives overview of large data sets and can be used prior to other data analysis methods in order to increase accuracy, overview and interpretation. Eq. (4) shows the PCA model for $\boldsymbol{A}$ Principal Components (PCs).

$$\boldsymbol{X} = \bar{\boldsymbol{x}} + \boldsymbol{T}_A\boldsymbol{P}_A^T + \boldsymbol{E}_A \tag{4}$$

where $\boldsymbol{P}_A$ are the loadings and orthonormal eigenvectors of $(\boldsymbol{X} - \bar{\boldsymbol{x}})^T(\boldsymbol{X} - \bar{\boldsymbol{x}})$ minimising the covariance between the $\boldsymbol{X}$-variables after $\boldsymbol{A}$ PCs. The scores ($\boldsymbol{T}_{\mathrm{A}}$) are orthogonal and calculated by Eq. (5).

$$\boldsymbol{T}_A = (\boldsymbol{X} - \boldsymbol{x})\boldsymbol{P}_A \tag{5}$$

The error term in Eq. (4) is $\boldsymbol{E}_A$ which is calculated by Eq. (6).

$$\boldsymbol{E}_A = \boldsymbol{X} - \bar{\boldsymbol{x}} - \boldsymbol{T}_A\boldsymbol{P}_A^T \tag{6}$$

PCA has commonly been used to identify correlations between analytical data and the properties of crude oils particularly from FT-ICR MS spectra as shown by the text mining study [48–52]. For instance, Hur et al. [49] analysed positive and negative mode APPI-FT-ICR MS spectra from 20 crude oils by PCA and identified differences between the oils based on their chemical composition. Moreover, their study showed a strong relationship between peaks in the mass spectra and the chemical properties of the oils indicating the potential for predicting crude oil properties from mass spectra.

4.4.2. Partial Least Squares Regression (PLSR)

PLSR [53] decomposes large data sets into a subspace of latent variables (scores and loadings) representing the main features of covariance between $\boldsymbol{X}$ (regressors) and $\boldsymbol{Y}$ (response). Both $\boldsymbol{X}$ and $\boldsymbol{Y}$ can be multivariate. $\boldsymbol{X}$ has the same input model as for PCA shown in Eq. (3). As PLSR also takes the response into account, as opposed to PCA, there is an input model for $\boldsymbol{Y}$ which is shown in Eq. (7).

$$\boldsymbol{Y} = \boldsymbol{Y}_{In} wgt_Y \tag{7}$$

where $\boldsymbol{Y}_{\mathrm{In}}$ is the response with shape (N, J) for N samples and J response variables and wgt_{Y} are the statistical weights balancing the sum of squares for the J $\boldsymbol{Y}$-variables in $\boldsymbol{Y}$, which has the shape (N, J).

The decomposition of $\boldsymbol{X}$ and $\boldsymbol{Y}$ is done simultaneously and iteratively, taking co-linearities in $\boldsymbol{Y}$ into account. For $\boldsymbol{X}$ the decomposition is shown in Eq. (8) and for $\boldsymbol{Y}$ in Eq. (9).

$$\boldsymbol{X} = \bar{\boldsymbol{x}} + \boldsymbol{T}_A \boldsymbol{P}_A^T + \boldsymbol{E}_A \tag{8}$$

$$\boldsymbol{Y} = \bar{\boldsymbol{y}} + \boldsymbol{U}_A \boldsymbol{Q}_A^T + \boldsymbol{F}_A \tag{9}$$

where $\boldsymbol{A}$ denotes the number of Principal Components (PCs) used and $\boldsymbol{E}_A$ and $\boldsymbol{F}_A$ are the error terms using $\boldsymbol{A}$ PCs. The loading weight matrix ($\boldsymbol{W}_A$) maximise the covariance between $\boldsymbol{X}$ and $\boldsymbol{Y}$ by maximising the covariance between $\boldsymbol{T}$ and $\boldsymbol{U}$ with $\boldsymbol{A}$ PCs. The scores ($\boldsymbol{T}_A$) are orthogonal as shown by Eq. (10).

$$\boldsymbol{T}_A = (\boldsymbol{X} - \bar{\boldsymbol{x}})\boldsymbol{W}_A \tag{10}$$

The loadings for $\boldsymbol{X}$ ($\boldsymbol{P}_A$) are calculated by Eq. (11) while the loadings for $\boldsymbol{Y}$ ($\boldsymbol{Q}_A$) are calculated by Eq. (12).

$$\boldsymbol{P}_A = (\boldsymbol{T}_A^T \boldsymbol{T}_A^T)^{-1} \boldsymbol{T}_A^T (\boldsymbol{X} - \boldsymbol{x}) \tag{11}$$

$$\boldsymbol{Q}_A = (\boldsymbol{T}_A^T \boldsymbol{T}_A^T)^{-1} \boldsymbol{T}_A^T (\boldsymbol{Y} - \boldsymbol{y}) \tag{12}$$

The error term for $\boldsymbol{X}$ ($\boldsymbol{E}_A$) is calculated as for PCA in Eq. (6) and the error term for $\boldsymbol{Y}$ ($\boldsymbol{F}_A$) is calculated by Eq. (13).

$$\boldsymbol{F}_A = \boldsymbol{Y} - \bar{\boldsymbol{y}} - \boldsymbol{T}_A \boldsymbol{Q}_A^T \tag{13}$$

The regression coefficients ($\boldsymbol{B}_A$), which are measures of the impact of variations in the various regressors on the respective response variables, are calculated by Eq. (14).

$$\boldsymbol{B}_A = \boldsymbol{W}_A \boldsymbol{Q}_A^T \tag{14}$$

Prediction of $\boldsymbol{Y}$ for a new sample ($\boldsymbol{X}_{\text{new}}$) is then obtained by Eq. (15) where $\boldsymbol{b}_0$ is the intercept.

$$\boldsymbol{Y}_{pred} = \boldsymbol{b}_0 + \boldsymbol{X}_{new} \boldsymbol{B}_A + \boldsymbol{F}_A \tag{15}$$

PLSR has been widely used for analysis of mass spectra in a variety of application areas, including for gas hydrates and FT-ICR MS. Vaz et al. [51] correlated the chemical composition of crude oil from FT-ICR MS data with the total acid number (TAN), using PLSR and support vector machines (SVMs) as multivariate calibration methods. In Terra et al. [54] negative-ion mode electrospray ionisation, ESI(-)-FT-ICR MS was coupled to PLSR and variable selection methods to estimate the TAN of Brazilian crude oil samples. They showed that it was possible to relate the selected variables to their corresponding molecular formulas, thus identifying the main chemical species responsible for the TAN values. In Hemmingsen et al. [16] TAN values were also used as a response for PLSR to predict the acidic properties of the crude oils.

Terra et al. [55] predicted basic nitrogen and aromatics contents in crude oil, using positive ion mode laser desorption ionisation (LDI) coupled to FT-ICR MS and PLSR with variable selection based on competitive adaptive reweighted sampling (CARS) in a procedure called CARSPLS regression.

Lozano et al. [56] used PLSR and genetic algorithm variable selection on APPI(+)-FT-ICR MS data for quantitative analysis of crude oils and their fractions. They estimated the API gravity and Conradson Carbon Residue of Colombian crude oil and vacuum residue (VR) samples with high accuracy.

PLSR can also be used for classification problems, for instance in the combination with discriminant analysis (DA), as PLS-DA. Two common DA methods are Linear Discriminant Analysis (LDA) and Quadratic Discriminant Analysis (QDA) which model the class conditional distribution of the data $P(X|y = k)$ for each class k. Predictions are obtained by using Bayes' rule, and the class that maximises this conditional probability is selected. The class priors $P(y = k)$ (the proportion of instances of class k), the class means and the covariance matrices are then estimated from the training data.

In Chua et al. [57] PLS-DA was used in tandem with PCA to analyse crude oil spill data from gas chromatography techniques. The PLS-DA and PCA combination accurately characterised the crude oil spill samples, overcoming the shortcomings of the traditional methods.

Likewise, Melendez-Perez et al. [58] utilised PLS-DA for analysis of ESI(-)-FT-ICR MS spectra of lacustrine oil and marine oil samples aiming towards comparing and classifying the samples. Results show that FT-ICR MS coupled with PLS-DA has potential to reveal oil characteristics more clearly.

Gjelsvik et al. [24] was the only publication from the text mining results regarding natural inhibitors. In this study, machine learning-based variable selection was used to identify components related to gas hydrate formation and PLS-DA emerged as the best performing method. This study showed that it is possible to identify features from FT-ICR MS spectra related to hydrate formation.

Accordingly, PLSR have already been shown to be able to predict chemical properties of crude oils, and PLS-DA has been shown to be able to classify crude oils samples with high accuracy.

4.4.3. Hierarchical Cluster-based Partial Least Squares Regression (HC-PLSR)

One promising extension of PLSR is the HC-PLSR [59] method, which is a locally linear regression method based on separating the observations into clusters and generating local PLSR models within each cluster. A global PLSR model comprising all observations is first made, and the observations are clustered based on the scores from this PLSR model. Local PLSR models are then made within each cluster. New observations are projected into the global model and classified based on their predicted $\boldsymbol{X}$-scores. Prediction of the response is based on either the closest local model or a weighted sum of all local models. HC-PLSR can be used with any clustering and classification method. HC-PLSR allows for local analysis within each cluster, and represents a way to handle highly nonlinear relationships between the regressors and the response.

4.4.4. Artificial Neural Networks (ANNs)

ANNs [60–62] are computing systems consisting of nodes called artificial neurons, between which the connections have numeric weights that are often initialised at random, and adjusted by backpropagation. Backpropagation uses the prediction error to calculate the gradient of the loss function with respect to the weights in the network. The neurons are placed in different layers, typically an input layer, one or more hidden layers, and an output layer. A widely used type of composition is the nonlinear weighted sum given by Eq. (16).

$$f(x) = K\left(\sum_i w_i g_i(x)\right) \tag{16}$$

where K is the activation function (some predefined function, such as the hyperbolic tangent or a sigmoid function), w_i are the weights and g_i are the different functions that are combined in the network. As ANNs use self learning, the network can adjust weights when a new situation is introduced, which leads to more flexible predictions than traditional regression models. ANNs are trained with experimental data where the output is a nonlinear function of the input data after learning a pattern and creating a prediction model [63]. Deep Neural Networks [64] are ANNs with multiple hidden layers between the input and output layers, as shown in Fig. 9. These can contain many layers of nonlinear hidden units.

Elgibaly and Elkamel [65,66] were the first to develop ANNs to predict thermodynamic conditions and suitable inhibitors for gas hydrate systems. Their network performed well compared to previous prediction methods based on traditional statistics and experimental data analysis, but showed signs of overfitting supposedly due to lack of experimental data. Chapoy et al. [67] used feed-forward neural networks (FNNs) to predict hydrate stability zones achieving a reasonable

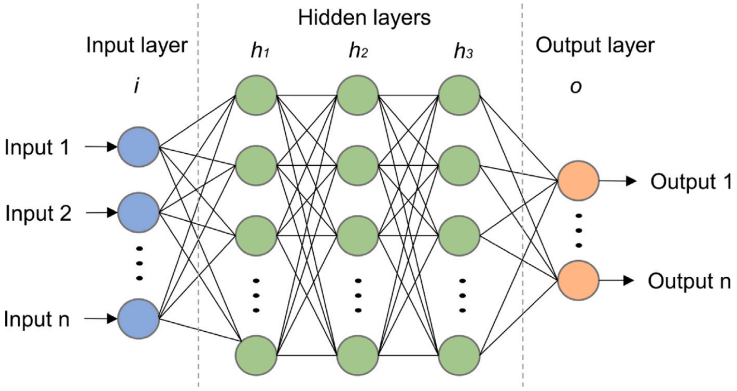


Fig. 9. Schematic example of a neural network with input layer, hidden layers and output layer.

model, but also pointing out deficiencies in the experimental data as a weakness of the study.

Ghavipour et al. [68] constructed an apparatus that measured specific gravity of different gas mixtures and pressure during a hydrate formation process. ANNs were then used to predict the hydrate formation conditions by a network with two hidden layers and 10 neurons in each layer, validated with Leave-One-Out cross validation.

Several studies have in the recent years used ANNs to predict hydrate formation conditions [69–71]. The purpose of these types of predictions is to identify the conditions where gas hydrates are formed and avoid operation within this region.

4.4.5. Support Vector Machines (SVMs)

SVMs [72] are supervised learning methods that analyse data for classification or regression analysis. SVMs are well suited for learning tasks where the number of variables is large compared to the number of observations in the training set.

For classification, SVMs construct a hyperplane or a set of hyperplanes in a high-dimensional space to separate the observations into two groups [73]. The goal is to find the hyperplane that has the largest distance (margin) to the nearest data point belonging to any of the two classes. The margin is defined as the distance between the separating hyperplane (decision boundary) and the training samples that are closest to this hyperplane. Data points that lie on the margin are known as support vector points, and the solution is represented as a linear combination of only these points. Decision boundaries with large margins tend to have a lower generalisation error, while decision boundaries with small margins are more prone to overfitting.

SVMs can be applied to nonlinear classification problems by using the so-called kernel trick, where the original space is mapped into a much higher-dimensional space where the observations can be more easily separated. To achieve this, a mapping function φ is used, as shown in Fig. 10. The hyperplanes in the higher-dimensional space are defined as the set of points whose dot product with a vector in that space is constant.

In Support Vector Regression (SVR), the hyperplane is the line that is used to predict the continuous output, shown in Fig. 11. SVR basically considers the points that are within the decision boundary lines and the regression line is then the hyperplane that has a maximum number of points.

SVMs are the second most commonly used methods for gas hydrates. Cao et al. [74] developed an SVM model for predictions of gas hydrate formation conditions, in combination with selection algorithms to optimise the process parameters for the SVM. Qin et al. [75] used both SVM and ANNs to predict gas hydrate plugging risks from flowloop and field data with SVM outperforming the ANN.

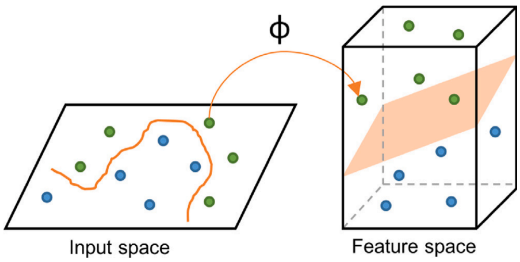


Fig. 10. The kernel trick to handle non-linear problems.

Rashid et al. [76], Mesbah et al. [77], Ghiasi et al. [78] and Yarveicy and Ghiasi [79] created SVM models with a linear modification of the SVM algorithm known as the least squares support vector machine (LSSVM) to predict thermodynamic properties of gas hydrate systems. One drawback with SVMs is the large number of quadratic computations performed to analyse the data, requiring high computational power, but LSSVM overcomes this due to the less complicated calculation methods [80].

As previously mentioned, Vaz et al. [51] predicted the TAN from FT-ICR MS spectra with SVM performing better than both PLSR and univariate methods. SVM is thereby able to both predict thermodynamic properties of hydrates and chemical properties of crude oils.

4.4.6. Decision Trees (DTs)

DTs [81,82] are attractive models when interpretability is important, and consist of a tree root, internal nodes, branches and leaf nodes. DTs ask a series of questions, and generate decision rules based on these. The model seeks to find the smallest set of rules that is consistent with the training data. In general, the rules have the form: *if* $condition_1$ *and* $condition_2$ *and* $condition_3$ *then outcome*. Fig. 12 shows an illustration of a decision tree model.

The rules are chosen to divide observations into segments that have the largest difference with respect to the target variable. Thus the rule selects both the variable and the best break point to separate the resulting subgroups maximally. The break points of variables are found using significance testing (F- or Chi-square with Bonferroni corrections) or reduction in variance criteria. To avoid overfitting, one often has to prune the tree by setting a limit for the maximal depth of the tree. A leaf can no longer be split when there are too few observations, the maximum depth (hierarchy of the tree) has been reached, or

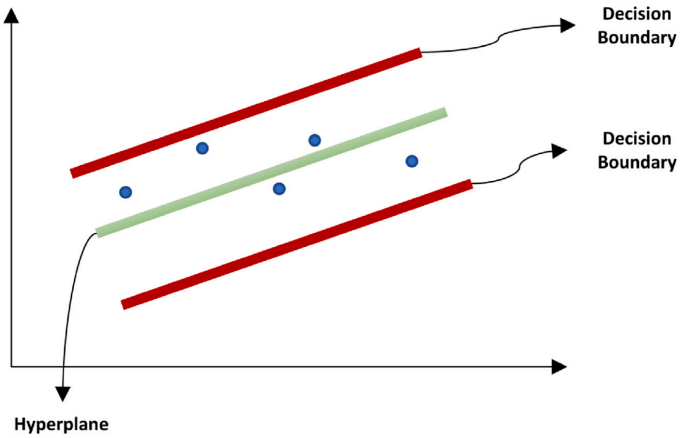


Fig. 11. Illustration of the hyperplane and decision boundaries in SVR.

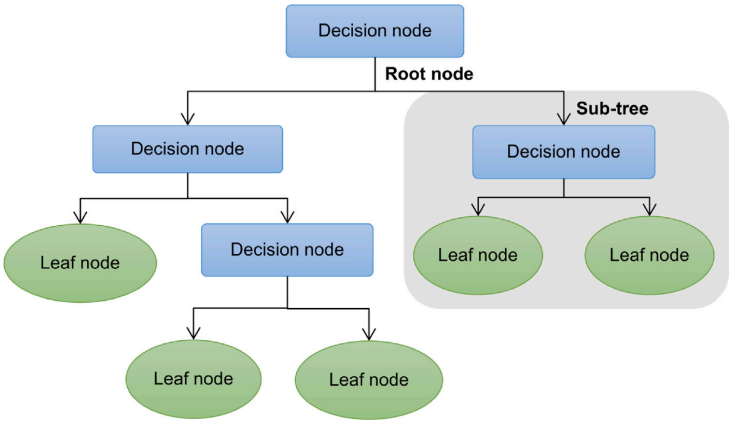


Fig. 12. Illustration of decision trees with the root node, sub-trees, decision nodes, branches and leaf nodes.

no significant split can be identified. It is assumed that observations belonging to different classes have different values in at least one of their variables. DTs are usually univariate, since they use splits based on a single feature at each internal node, but methods are available for constructing multivariate trees [83].

To improve the prediction of the DT, a boosting method can be applied. Boosting is an ensemble method for improving predictions of a weak learning algorithm [84]. The weak learners are trained sequentially, trying to improve upon its predecessor. When boosting is applied to a tree, each tree is dependent on prior trees and the algorithm learns by fitting the residual of the prior trees. One example of a boosting method is XGBoost (eXtreme Gradient Boosting). In XGBoost, trees are built at every iteration, always minimising the prediction error of the classifier while introducing a penalty function to utilise the computational power more efficiently.

4.4.7. Random Forest (RF)

In DTs, the initial selected split affects the optimality of variables considered for subsequent splits. Ensemble tree models grow trees with varying initial splits, and use either a voting or the average of the predictions for each new data point across all trees. The vote distribution can be used to develop a nonparametric probabilistic predictive model. The ensemble is less prone to overfitting and other problems of individual DTs, and generally performs better. RF [85–87] is an example of such an ensemble tree method. For RF, each tree is based on a random subset of the data and variables (selected by bootstrapping). The change in prediction accuracy when the values of a feature are randomly permuted among observations gives estimates of the importance of each feature.

Tree models and boosting are among the most common regression and classification methods, and has been used for gas hydrates and crude oil analysis. Song et al. [88] used a gradient boosted regression tree algorithm to predict hydrate phase equilibrium conditions in the presence of various salts, organic substances or water. The model was compared to an ANN, where the regression tree achieved the best prediction model for gas hydrates' phase equilibrium conditions in the presence of various salts or organics.

In Acharya and Bahadur [89] RF and XGBoost were used to predict gas hydrate dissociation temperatures in the presence of hydrate inhibitors and precursors achieving good predictions.

Lovatti et al. [90] proposed two strategies for the use of RF and data reduction techniques for NMR spectra of petroleum samples. The study compared the NMR spectra to the TAN values of the petroleum, and

the method was able to identify a relationship between the TAN and specific regions in the spectra.

4.4.8. Naïve Bayes (NB) classification

A Bayesian network (BN) is a probabilistic model that represents a set of random variables and their conditional independence via a directed acyclic graph (DAG). Using e.g. Chi-squared and mutual information tests, one can find the conditional independence relationships among the variables and use these relationships as constraints to construct a BN. BNs can take prior knowledge into account, by e.g. setting a certain node as *root node* or *leaf node*, thereby applying knowledge of nodes that are direct causes or effects of other nodes. This results in nodes that are not directly connected to another node, or that two nodes are independent.

The probabilistic parameters are encoded into a set of tables, one for each variable, in the form of local conditional distributions of a variable given its parents. The joint distribution can be reconstructed by multiplying these tables (given the independencies encoded into the network). BNs are DAGs whose nodes represent random variables that may be e.g. observable quantities or latent variables. Edges (connections) represent conditional dependencies, and each node is associated with a probability function.

Naïve Bayesian networks are very simple BNs which are composed of DAGs with only one parent (representing the unobserved node) and several children (corresponding to observed nodes), where the child nodes are assumed to be independent. Naïve Bayes (NB) classification may be impaired by probabilities of 0, but this can be avoided by using a Laplace estimator.

The assumption of independence among child nodes is most often not valid, but this can be corrected for by adding extra edges to include some of the dependencies between the variables. In this case, the network has the limitation that each feature can be related to only one other feature [91]. Selective Bayesian classifiers [92] include a feature selection stage to remove irrelevant variables or one of the two totally correlated variables.

Shi et al. [93] used a variational Bayesian neural network for probabilistic deepwater natural hydrate gas dispersion modelling of simulated data. Combined with a convolutional neural network, the model performed well.

Bayesian networks have been used for risk and safety assessment of storing and transportation of crude and heavy oil. For example, Zhang et al. [94] used BNs to evaluate the leak safety of heavy oil gatherings in pipelines. BNs find the probability for leakage and fuzzy set theory evaluates the consequences of the leakage.

4.4.9. k-nearest neighbours (KNN) classification

KNN [95] locates the k nearest observations to the observation to be classified (e.g. by an exhaustive search algorithm) based on the chosen distance metric, and identifies the most frequent class membership among the neighbours. The number k is specified by the user, and the right choice of k is crucial to find a good balance between overfitting and underfitting. Weights are assigned to the contributions of the neighbours in a majority voting to predict the classes, so that the nearer neighbours contribute more to the average than the more distant ones.

KNN is fundamentally different from the other supervised classifiers described here, in that it is a so-called lazy learner. KNN does not learn a discriminative function from the training data but memorises it instead. The main advantage of such a memory-based approach is that the classifier immediately adapts as we collect new training data. However, the computational complexity for classifying new samples grows with the number of samples in the training data set and storage space can hence become a challenge when working with large data sets.

Only two instances where KNN were used related to gas hydrates were found. Xu et al. [96] used KNN regression, SVM, RF and XGBoost for the prediction of hydrate formation temperatures achieving good predictions with all methods.

Amin et al. [97] used KNN to predict hydrate equilibrium conditions to CO_2 capture. The model was simple but showed good predictions with low errors, indicating that KNN is a valuable method for analysis of gas hydrate thermodynamics.

4.5. Regularisation-based methods

Another group of machine learning methods that we find promising for identification of hydrate active compounds in crude oils is regularisation-based methods, which are very useful for feature selection purposes. The most commonly used regularisation-based methods are Ridge regression [98], LASSO (least absolute shrinkage and selection operator) [99] and Elastic net [100]. Regularisation-based versions of PLSR are also available, which have shown promise in feature selection, such as Sparse-PLS [101]/Soft-Threshold PLS [102] and Powered PLS [103]. These may have advances over other regularisation-based methods in cases where interpretation is important, due to the possibilities to gain overview of complex data sets through decomposition of the data into a lower-dimensional subspace of latent variables.

4.5.1. Ridge regression

Ridge regression is also known as L2-regularisation. In Ridge, the sum of the squares of the regression coefficients (β) is forced to be less than a fixed value, which shrinks the size of the coefficients. Ordinary least squares (OLS) minimises Eq. (17).

$$RSS_{OLS} = \sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2 \tag{17}$$

while Ridge regression minimise Eq. (18).

$$RSS_{Ridge} = \sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2 + \lambda\sum_{i=1}^{p}\beta_j^2 \tag{18}$$

where $\lambda \geq 0$ is a penalty term which is often found by cross-validation. This gives Eqs. (19) and (20).

$$B_{OLS} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y} \tag{19}$$

$$B_{Ridge} = (\boldsymbol{X}^T\boldsymbol{X} + \lambda I)^{-1}\boldsymbol{X}^T\boldsymbol{Y} \tag{20}$$

Hence, Ridge regression handles multicollinearity in the regressor ($\boldsymbol{X}$) matrix, while OLS regression does not.

4.5.2. LASSO

In LASSO, the estimates of the regression coefficients are obtained using L1-constrained least squares. This forces the sum of the absolute values of the regression coefficients to be less than a fixed value, which forces certain coefficients to be set to zero. LASSO is a feature selection method, since variables having zero regression coefficients are omitted from the model. In LASSO Eq. (21) is minimised.

$$RSS_{LASSO} = \sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2 + \lambda\sum_{i=1}^{p}\beta_j \tag{21}$$

4.5.3. Elastic net

Elastic net combines the L1 and L2 penalties of the Ridge and LASSO methods linearly as given by Eq. (22).

$$RSS_{EN} = \sum_{i=1}^{n}\left(y_i - \beta_0\sum_{j=1}^{p}\beta_j x_{ij}\right)^2 + \lambda_1\sum_{i=1}^{p}\beta_j^2 + \lambda_2\sum_{i=1}^{p}\beta_j \tag{22}$$

In Elastic net, highly correlated regressors will tend to have similar regression coefficients, which creates a grouping effect that is desirable in many applications.

Landgrebe and Nkazi [104] used traditional L1/L2 in order to reduce overfitting of the neural network, but dropout regularisation proved more effective.

In Singh et al. [105] Ridge Regression (L2) was used among other methods to estimate gas hydrate saturation in sedimentary systems from well-logs by NMR measurements. L2 achieved good accuracy and was one of the best performing methods. This is an indication that L2 could also perform well with other spectroscopic data of gas hydrate related samples, such as FT-ICR MS spectra.

Similarly, other regularisation methods have been used in combination with spectroscopic data previously. In Fu et al. [106] both Sparse-PLS and Elastic net were used for wavelength selection on data from NIR spectroscopy of corn and gasoline. Both methods select intervals of wavelengths, where Elastic net selects a smaller model, while Sparse-PLS achieves a higher accuracy. Finding the wavelengths closely related to the response could significantly improve a prediction model.

4.5.4. Convolutional Neural Networks (CNNs)

CNNs are deep neural networks which use convolutions to extract information in one or more of the hidden layers [63]. CNNs are regularised versions of fully connected networks. In a convolutional layer, the data is organised in a feature map where the weights are connected to the previous layer. These weights are used to filter for patterns in the data. Commonly used in pattern recognition, CNNs are good feature extractors by learning the most important variables by itself.

CNNs can be a valuable tool for instance for the analysis of mass spectrometry data. Lv et al. [107] used CNNs to analyse peak information in tandem mass spectrometry (MS/MS). This method outperformed others such as SVMs, PCA, deep neural networks and XGBoost. Due to the nature of the convolutional filters, CNNs are able to learn both the peak shape and the m/z values, achieve greater robustness for low signal-to-noise ratios and can allow for a higher-level representation of lower-level features representing patterns [108]. Hence, CNNs could be very useful for analysing FT-ICR MS data.

Kim et al. [109] used CNNs for saturation modelling from X-ray CT images. The 1-dimensional CNNs performed well, but the method shows difficulties in determination of optimal parameters for the CNNs.

Li et al. [110] constructed a neural network based on a variational autoencoder with convolutional layers to predict pore size distributions in subsurface shale reservoirs. The method showed good predictions and although this is not directly related to gas hydrates, gas hydrates are analysed in a similar manner, indicating that CNN could be a valuable method given an optimal parameter search.

4.6. Data used in literature

The data used in many of the machine learning models previously developed in the field of gas hydrates have been sampled from the literature. In this review, a number of the cited articles discussed are based on data sampled from other publications [65–67,69–71,74,77,78,88,89,96,104]. These references are mainly based on thermodynamic data, concerning prediction of gas hydrate formation/dissociation conditions and phase equilibrium measurements. Sloan and Koh [1] present an extensive list of experimental data which are frequently used by the authors sampling experimental data from the literature [65–67,69,71,104]. Consequently, the models from these authors are based on the same data. This can result in shortcomings, as the errors in predictions from these models approximate the errors of the experiments. Additionally, where the data are deficient, extrapolation has to be performed which decreases the accuracy of the predictions [3]. It is therefore clear that there is a need for more experimental data. New experimental data should fill the gaps in already published data, and as many of the models are based on thermodynamic properties, other aspects of gas hydrates could be valuable to examine closer. Better understanding of the mechanisms and the molecular composition related to the inhibition/dissociation of gas hydrates, could lead to strengthened prediction models for the thermodynamic, physical and chemical properties of gas hydrates in the future.

5. Conclusions and future perspectives

In this paper a text mining study was performed to evaluate the use of machine learning methods within the field of gas hydrates with specific focus on the oil chemistry. An evaluation of FT-ICR MS was included in the study to establish a link between aspects of gas hydrates and analysis of crude oils. Several machine learning methods were identified as promising and their use in the literature was evaluated. For studies regarding gas hydrates, predictions of thermodynamic properties were most common, while FT-ICR MS was used for analysis of oil chemistry and chemical properties. Most of the publications on thermodynamic properties of gas hydrates were also created using the same data sources. It could therefore be beneficial to explore other areas of gas hydrate research using machine learning in the future. Although there is little literature describing the use of FT-ICR MS to characterise gas hydrates, the text mining results show that FT-ICR MS has been used to characterise crude oils for some time and with success. Therefore, with the combination of FT-ICR MS and machine learning, it may be possible to identify the hydrate-active compounds responsible for the differences between oils forming plugging hydrates and oils forming transportable hydrates. This can be done by relating the composition of the oil, determined by FT-ICR MS to information regarding hydrate formation. The methods presented in this paper successfully predicted thermodynamic properties in gas hydrates or chemical properties from FT-ICR MS, and the methods could therefore be tested with the aim of predicting chemical properties from gas hydrate related samples. We believe that an approach which is able to predict hydrate behaviour may lead to new knowledge about natural gas hydrate inhibitors. The development of a universal method to identify natural components which inhibit, or work as AAs for gas hydrates would contribute to new understanding and decision making tools in the field of gas hydrate flow assurance and management strategies. This could lead to better decision support tools and better risk evaluations for transportation of crude oils with gas hydrates present.

The text mining study revealed that the amount of research using machine learning to analyse both gas hydrate and FT-ICR MS data is still limited, but research on both topics have increased in recent years. For FT-ICR MS, most publications used PCA for analysis of the data, and several of the publications used the chemical composition data to build machine learning models instead of using the mass spectra directly. Identifying relationships and building models based on the mass spectra requires less pre-processing steps and could therefore be advantageous and could be explored further.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

No data was used for the research described in the article.

Acknowledgements

The authors thank the The Research Council of Norway, Equinor ASA, Norway, OMV (Norge) AS, Norway, Wintershall DEA Norge AS and TotalEnergies, Norway for funding. This work is a part of the Knowledge-Building Project for Industry (PETROMAKS 2), Project number: 294636 “New Hydrate Management: New understanding of hydrate phenomena in oil systems to enable safe operation within the hydrate zone”.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.fuel.2022.126696.**S1 Table. Results from the text mining study**. Table of the publications returned by the text mining study.

References

[1] Sloan ED, Koh CA. Clathrate hydrates of natural gases. Chemical industries series, 3rd ed.. vol. 119, Boca Raton, FL: CRC Press, Taylor & Francis Group; 2008.
[2] Fotland P, Askvik KM. Some aspects of hydrate formation and wetting. J Colloid Interface Sci 2008;321:130–41.
[3] Sloan ED. A changing hydrate paradigm—from apprehensionto avoidance to risk management. Fluid Phase Equilib 2005;228–229:67–74.
[4] Kelland MA. History of the development of low dosage hydrate inhibitors. Energy Fuels 2006;20:825–47.
[5] Nasir Q, Suleman H, Elsheikh YA. A review on the role and impact of various additives as promoters/ inhibitors for gas hydrate formation. J Nat Gas Sci Eng 2020;76:103211.
[6] Sa J-H, Melchuna A, Zhang X, Rivero M, Glénat P, Sum AK. Investigating the effectiveness of anti-agglomerants in gas hydrates and iceformation. Fuel 2019;255:115841.
[7] Ding L, Shi B, Liu Y, Song S, Wang W, Wu H, Gong J. Rheology of natural gas hydrate slurry: Effect of hydrate agglomeration anddeposition. Fuel 2019;239:126–37.
[8] Lederhos J, Longs J, Sum A, Christiansen RL, Sloan Jr ED. Effective kinetic inhibitors for natural gas hydrates. Chem Eng Sci 1995;51:1221–9.
[9] Shahnazar S, Bagheri S, TermehYousefi A, Mehrmashhadi J, Karim MSA, Kadri NA. Structure, mechanism, and performance evaluation of natural gas hydrate kinetic inhibitors. Rev Inorg Chem 2018;38:1–19.
[10] Lingelem MN, Majeed AI, Stange E. Industrial experience in evaluation of hydrate formation, inhibition, and dissociation in pipeline design and operation. Ann New York Acad Sci 1994;715:75–93.
[11] Fadnes FH. Natural hydrate inhibiting components in crude oils. Fluid Phase Equilib 1996;117:186–92.
[12] Borgund AE, Høiland S, Barth T, Fotland P, Askvik KM. Molecular analysis of petroleum derived compounds that adsorbonto gas hydrate surfaces. Appl Geochem 2009;24:777–86.
[13] Høiland S, Askvik KM, Fotland P, Alagic E, Barth T, Fadnes F. Wettability of Freon hydrates in crude oil/brine emulsions. J Colloid Interface Sci 2005;287:217–25.
[14] Høiland S, Borglund AE, Barth T, Fotland P, Askvik KM. Wettability of Freon hydrates in crude oil/brine emulsions: the effects of chemical additives. In: 5th international conference in gas hydrate, Vol. 4. Trondheim; 2005, p. 1151–61.
[15] Borgund AE, Erstad K, Barth T. Fractionation of crude oil acids by HPLC and characterization of their properties and effects on gas hydrate surfaces. Energy Fuels 2007;21:2816–26.
[16] Hemmingsen PV, Kim S, Pettersen HE, Rodgers RP, Sjöblom J, Marshall AG. Structural characterization and interfacial behavior of acidic compounds extracted from a North Sea oil. Energy Fuels 2006;20:1980–7.
[17] Hemmingsen PV, Li X, Peytavy J-L, Sjöblom J. Hydrate plugging potential of original and modified crude oils. J Dispers Sci Technol 2007;28:371–82.
[18] Erstad K, Høiland S, Fotland P, Barth T. Influence of petroleum acids on gas hydrate wettability. Energy Fuels 2009;23:2213–9.
[19] Qiao P, Harbottle D, Tchoukov P, Masliyah J, Sjoblom J, Liu Q, Xu Z. Fractionation of asphaltenes in understanding their role in petroleum emulsion stability and fouling. Energy Fuels 2016;31:3330–7.
[20] Salmin DC. The impact of synthetic and natural surface-active components on hydrate agglomeration (Doctoral thesis), Golden, Colorado: Colorado School of Mines; 2019.
[21] Adams JJ. Asphaltene adsorption, a literature review. Energy Fuels 2014;28:2831–56.
[22] Kilpatrick PK. Water-in-crude oil emulsion stabilization: Review and unanswered questions. Energy Fuels 2012;26:4017–26.
[23] Yang F, Tchoukov P, Dettman H, Teklebrhan RB, Liu L, Dabros T, Czarnecki J, Masliyah J, Xu Z. Asphaltene subfractions responsible for stabilizing water-in-crude oil emulsions. Part 2: Molecular representations and molecular dynamics simulations. Energy Fuels 2015;29:4783–94.
[24] Gjelsvik EL, Fossen M, Brunsvik A, Tøndel K. Using machine learning-based variable selection to identify hydrate related components from FT-ICR MS spectra. PLoS One 2022;17(8):e0273084.
[25] Marshall AG, Rogers RP. Petroleomics: The next grand challenge for chemical analysis. Acc Chem Res 2004;37:53–9.
[26] Hughey CA, Rodgers RP, Marshall AG. Resolution of 11 000 compositionally distinct components in a single electrospray ionization Fourier transform ion cyclotron resonance mass spectrum of crude oil. Anal Chem 2002;74:4145–9.
[27] Cho Y, Ahmed A, Islam A, Kim Sunghwan. Developments in FT-ICR MS instrumentation, ionization techniques, and data interpretation methods for petroleomics. Mass Spectrom Rev 2014;34:248–63.
[28] Emmett MR, White FM, Hendrickson CL, Shi SD-H, Marshall AG. Application of micro-electrospray liquid chromatography techniques to FT-ICR MS to enable high-sensitivity biological analysis. J Am Soc Mass Spectrom 1998;9:333–40.
[29] Hughey CA, Hendrickson CL, Rodgers RP, Marshall AG. Kendrick mass defect spectrum: A compact visual analysis for ultrahigh-resolution broadband mass spectra. Anal Chem 2001;73:4676–81.
[30] Marshall AG, Rodgers RP. Petroleomics: Chemistry of the underworld. Proc Natl Acad Sci USA 2008;105:18090–5.
[31] de Hoffmann E, Stroobant V. Mass spectrometry: Principles and applications. 3rd ed.. West Sussex, England: John Wiley and Sons Ltd.; 2012.
[32] Hur M, Yeo I, Kim E, No M-h, Koh J, Cho YJ, Lee JW, Kim S. Correlation of FT-ICR mass spectra with the chemical and physical properties of associated crude oils. Energy Fuels 2010;24:5524–32.
[33] Klein GC, Kim S, Rodgers RP, Marshall AG, Yen A. Mass spectral analysis of asphaltenes. II. Detailed compositional comparison of asphaltenes deposit to its crude oil counterpart for two geographically different crude oils by ESI FT-ICR MS. Energy Fuels 2006;20:1973–9.
[34] Schaub TM, Jennings DW, Kim S, Rodgers RP, Marshall AG. Heat-exchanger deposits in an inverted steam-assisted gravity drainage operation. Part 2. Organic acid analysis by electrospray ionization Fourier transform ion cyclotron resonance mass spectrometry. Energy Fuels 2007;21:185–94.
[35] Smith DF, Rahimi P, Teclemariam A, Rodger RP, Marshall AG. Characterization of athabasca bitumen heavy vacuum gas oil distillation cuts by negative/positive electrospray ionization and automated liquid injection field desorption ionization Fourier transform ion cyclotron resonance mass spectrometry. Energy Fuels 2008;22:3118–25.
[36] Headley JV, Peru KM, Barrow MP, Derrick PJ. Characterization of naphthenic acids from athabasca oil sands using electrospray ionization: The significant influence of solvents. Anal Chem 2007;79:6222–9.
[37] Barrow MP, Headley JV, Peru KM, Derrick PJ. Data visualization for the characterization of naphthenic acids within petroleum samples. Energy Fuels 2009;23:2592–9.
[38] Fernandez-Lima FA, Becker C, McKenna AM, Rodgers RP, Marshall AG, Russell DH. Petroleum crude oil characterization by IMS-MS and FTICR MS. Anal Chem 2009;81:9941–7.
[39] Qian K, Robbins WK, Hughey CA, Cooper HJ, Rodgers RP, Marshall AG. Resolution and identification of elemental compositions for more than 3000 crude acids in heavy petroleum by negative-ion microelectrospray high-field Fourier transform ion cyclotron resonance mass spectrometry. Energy Fuels 2001;15:1505–11.
[40] Qian K, Rodgers RP, Hendrickson CL, Emmett MR, Marshall AG. Reading chemical fine print: Resolution and identification of 3000 nitrogen-containing aromatic compounds from a single electrospray ionization Fourier transform ion cyclotron resonance mass spectrum of heavy petroleum crude oil. Energy Fuels 2001;15:492–8.
[41] Burnham JF. Scopus database: a review. Biomed Digit Libr 2006;3:8.
[42] Rose ME, Kitchin JR. Pybliometrics: Scriptable bibliometrics using a Python interface to Scopus. SoftwareX 2019;10:100263.
[43] AlRyalat SAS, Malkawi LW, Momani SM. Comparing bibliometric analysis using PubMed, Scopus, and Web of Science Databases. J Vis Exp 2019;152:12.
[44] van der Maaten L, Hinton G. Visualizing data using t-SNE. J Mach Learn Res 2008;9:2579–605.
[45] Bishop CM. Pattern recognition and machine learning. Information science and statistics, 1st ed.. New York, NY: Springer; 2006.
[46] Mitchell TM. Machine learning. McGraw-Hill series in computer science. Artificial intelligence, 1st ed.. McGraw-Hill Education; 1997.
[47] Pearson K. On lines and planes of closest fit to systems of points in space. Phil Mag 1901;2:559–72.
[48] Fossen M, Hemmingsen PV, Hannisdal A, Sjöblom J, Kallevik H. Solubility parameters based on IR and NIR spectra: I. Correlation to polar solutes and binary systems. J Dispers Sci Technol 2004;26:227–41.
[49] Hur M, Yeo I, Park E, Kim YH, Yoo J, Kim E, No M-h, Koh J, Kim S. Combination of statistical methods and Fourier transform ion cyclotron resonance mass spectrometry for more comprehensive, molecular-level interpretations of petroleum samples. Anal Chem 2010;82:211–8.
[50] Chiaberge S, Fiorani T, Savoini A, Bionda A, Ramello S, Pastori M, Cesti P. Classification of crude oil samples through statistical analysis of APPI FTICR mass spectra. Fuel Process Technol 2013;106:181–5.
[51] Vaz BG, Abdelnur PV, Rocha WFC, Gomes AO, Pereira RCL. Predictive petroleomics: Measurement of the total acid number by electrospray Fourier transform mass spectrometry and chemometric analysis. Energy Fuels 2013;27:1873–80.
[52] Sad CM, d. Silva M, d. Santos FD, Pereira LB, Corona RR, Silva SR, Portela NA, Castro EV, Filgueiras PR, Jr VL. Multivariate data analysis applied in the evaluation of crude oil blends. Fuel 2018;239:421–8.
[53] Wold S, Martens H, Wold H. The multivariate calibration problem in chemistry solved by the PLS method. In: Matrix pencils. Lecture notes in mathematics, vol. 973, Berlin, Heidelberg: Springer; 1983, p. 286–93.

[54] Terra LA, Filgueiras PR, Tose LV, Romão W, d. Souza DD, d. Castro EVR, d. Oliveira MSL, Diase JCM, Poppi RJ. Petroleomics by electrospray ionization FT-ICR mass spectrometry coupled to partial least squares with variable selection methods: prediction of the total acid number of crude oils. Analyst 2014;139:4908–16.

[55] Terra LA, Filgueiras PR, Tose LV, Romão W, d. Castro EV, d. Oliveira LM, Dias JC, Vaz BG, Poppi RJ. Laser desorption ionization FT-ICR mass spectrometry and CARSPLS for predicting basic nitrogen and aromatics contents in crude oils. Fuel 2015;160:274–81.

[56] Lozano DCP, Orrego-Ruiz JA, Hernández RC, Guerrero JE, Mejía-Ospino E. APPI(+)-FTICR mass spectrometry coupled to partial least squares with genetic algorithm variable selection for prediction of API gravity and CCR of crude oil and vacuum residues. Fuel 2017;193:39–44.

[57] Chua CC, Brunswick P, Kwok H, Yan J, Cuthbertson D, Aggelen Gv, Helbing CC, Shang D. Enhanced analysis of weathered crude oils by gas chromatography-flame ionization detection, gas chromatography-mass spectrometry diagnostic ratios, and multivariate statistics. J Chromatogr A 2020;1634:461689.

[58] Melendez-Perez JJ, Oliveira LFC, Miranda N, Sussulini A, Eberlin MN, Bastos WL, Rangel MD, d. S. Rocha Y. Lacustrine versus marine oils: Fast and accurate molecular discrimination via electrospray Fourier transform ion cyclotron resonance mass spectrometry and multivariate statistics. Energy Fuels 2020;8:9222–30.

[59] Tøndel K, Indahl UG, Gjuvsland AB, Vik JO, Hunter P, Omholt SW, Martens H. Hierarchical cluster-based partial least squares regression (HC-PLSR) is an efficient tool for metamodelling of nonlinear dynamic models. BMC Syst Biol 2011;5:90.

[60] Bishop CM. Neural networks for pattern recognition. Advanced texts in econometrics, 1st ed.. vol. 198, Madison Ave. New York, NY, United States: Oxford University Press, Inc.; 1995.

[61] Udelhoven T, Naumann D, Schmitt J. Development of a hierarchical classification system with artificial neural networks and FT-IR spectra for the identification of bacteria. Appl Spectrosc 2000;54.

[62] Udelhoven T, Novozhilovx M, Schmitt J. The NeuroDeveloper®: a tool for modular neural classification of spectroscopic data. Chemometr Intell Lab Syst 2003;66:219–26.

[63] LeCun Y, Bengio Y, Hinton G. Deep learning. Nature 2015;521:436–44.

[64] Schmidhuber J. Deep learning in neural networks: An overview. Neural Netw 2015;61:85–117.

[65] Elgibaly A, Elkamel A. A new correlation for predicting hydrate formation conditions forvarious gas mixtures and inhibitors. Fluid Phase Equilib 1998;152:23–42.

[66] Elgibaly A, Elkamel A. Optimal hydrate inhibition policies with the aid of neural networks. Energy Fuels 1998;13:105–13.

[67] Chapoy A, Mohammadi A, Richon D. Predicting the hydrate stability zones of natural gases using artificial neural networks. Oil Gas Sci Technol 2007;62:701–6.

[68] Ghavipour M, Ghavipour M, Chitsazan M, Najibi SH, Ghidary SS. Experimental study of natural gas hydrates and a novel use of neural network to predict hydrate formation conditions. Chem Eng Res Des 2012;91:264–73.

[69] Hesami SM, Dehghani M, Kamali Z, Bakyani AE. Developing a simple-to-use predictive model for prediction of hydrate formation temperature. Int J Ambient Energy 2015;38:380–8.

[70] Soroush E, Mesbah M, Shokrollahi A, Rozyn J, Lee M, Kashiwao T, Bahadori A. Evolving a robust modeling tool for prediction of natural gas hydrateformation conditions. J Unconv Oil Gas Resour 2015;12:45–55.

[71] Ghayyem MA, Nasab AG, Khormizi MZ, Rostami M. Predicting the conditions for gas hydrate formation. Pet Sci Technol 2019;37:1855–60.

[72] Cortes C, Vapnik V. Support-vector networks. Mach Learn 1995;20. 273-229.

[73] Burges CJ. A tutorial on support vector machines for pattern recognition. Data Min Knowl Discov 1998;2:121–67.

[74] Cao J, Zhu S, Li C, Han B. Integrating support vector regression with genetic algorithm for hydrate formation condition prediction. Processes 2020;8:519.

[75] Qin H, Srivastava V, Wang H, Zerpa LE, Koh CA. Machine learning models to predict gas hydrate plugging risks using flowloop and field data. In: Offshore technology conference, conference paper. 2019, p. 12.

[76] Rashid S, Fayazi A, Harimi B, Hamidpour E, Younesi S. Evolving a robust approach for accurate prediction of methane hydrate formation temperature in the presence of salt inhibitor. J Nat Gas Sci Eng 2014;18:194–204.

[77] Mesbah M, Soroush E, Rezakazemi M. Development of a least squares support vector machine model for prediction of natural gas hydrate formation temperature. Chin J Chem Eng 2016;25:1238–48.

[78] Ghiasi MM, Yarveicy H, Arabloo M, Mohammadi AH, Behbahani RM. Modeling of stability conditions of natural gas clathrate hydrates using least squares support vector machine approach. J Mol Liq 2016;223.

[79] Yarveicy H, Ghiasi MM. Modeling of gas hydrate phase equilibria: Extremely randomized trees and LSSVM approaches. J Mol Liq 2017;243:533–41.

[80] Suykens J, Vandewalle J. Least squares support vector machine classifiers. Neural Process Lett 1999;9:293–300.

[81] Quinlan JR. Simplifying decision trees. Int J Man-Mach Stud 1987;27:221–34.

[82] Utgoff PE. Incremental induction of decision trees. Mach Learn 1989;4:161–86.

[83] Brodley CE, Utgoff PE. Multivariate decision trees. Mach Learn 1995;19:45–77.

[84] Breiman L. Arcing classifier (with discussion and a rejoinder by the author). Ann Statist 1998;26(3):801–49.

[85] Breiman L. Bagging predictors. Mach Learn 1996;24:123–40.

[86] Breiman L. Random forests. Mach Learn 2001;45:5–32.

[87] Ho TK. The random subspace method for constructing decision forests. IEEE Trans Pattern Anal Mach Intell 1998;20:832–44.

[88] Song Y, Zhou H, Wang P, Yang M. Prediction of clathrate hydrate phase equilibria using gradient boosted regression trees and deep neural networks. J Chem Thermodyn 2019;135:86–96.

[89] Acharya PV, Bahadur V. Thermodynamic features-driven machine learning-based predictions of clathrate hydrate equilibria in the presence of electrolytes. Fluid Phase Equilib 2021;530:112894.

[90] Lovatti BP, Nascimento MH, Rainha KP, Oliveira EC, Neto AC, Castro EV, Filgueiras PR. Different strategies for the use of random forest in NMR spectra. J Chemometr 2020;34:e3231.

[91] Kotsiantis SB, Zaharakis ID, Pintelas PE. Machine learning: a review of classification and combining techniques. Artif Intell Rev 2007;26:159–90.

[92] a. Ratanamahatana C, Gunopulos D. Feature selection for the naive bayesian classifier using decision trees. Appl Artif Intell 2003;17(5–6):475–87.

[93] Shi J, Li J, Usmani AS, Zhu Y, Chen G, Yang D. Probabilistic real-time deep-water natural gas hydrate dispersion modeling by using a novel hybrid deep learning approach. Energy 2021;219:119572.

[94] Zhang P, Chen X, Fan C. Research on a safety assessment method for leakage in a heavy oil gathering pipeline. Energies 2020;13:1340.

[95] Altman NS. An introduction to kernel and nearest-neighbor nonparametric regression. Amer Statist 1992;46:175–85.

[96] Xu H, Jiao Z, Zhang Z, Huffman M, Wang Q. Prediction of methane hydrate formation conditions in salt water using machine learning algorithms. Comput Chem Eng 2021;151:107358.

[97] Amin JS, Bahadori A, Nia BH, Rafiee S, Kheilnezhad N. Prediction of hydrate equilibrium conditions using k-nearest neighbor algorithm to CO2 capture. Pet Sci Technol 2017;35:1070–7.

[98] Hoerl AE. Application of ridge analysis to regression problems. Chem Eng Prog 1958;58(3):54–9.

[99] Tibshirani R. Regression Shrinkage and selection via the Lasso. J R Stat Soc Ser B Stat Methodol 1996;58(1):267–88.

[100] Zou H, Hastie T. Regularization and variable selection via the elastic net. J R Stat Soc Ser B Stat Methodol 2005;67:301–20.

[101] Cao K-AL, Rossouw D, Robert-Granié C, Besse P. A sparse PLS for variable selection when integrating omics data. Stat Appl Genet Mol Biol 2008;7:35.

[102] Sæbø S, Almøy T, Aarøe J, Aastveit AH. ST-PLS: a multi-directional nearest shrunken centroid type classifier via PLS. J Chemometr 2007;22:54–62.

[103] Liland KH, Indahl U. Powered partial least squares discriminant analysis. J Chemometr 2009;23:7–18.

[104] Landgrebe MKB, Nkazi D. Toward a robust, universal predictor of gas hydrate equilibria by means of a deep learning regression. ACS Omega 2019;4:22399–417.

[105] Singh H, Seol Y, Myshakin EM. Prediction of gas hydrate saturation using machine learning and optimal set of well-logs. Comput Geosci 2020.

[106] Fu G-H, Zong M-J, Wang F-H, Yi L-Z. A comparison of sparse partial least squares and elastic net in wavelength selection on NIR spectroscopy data. Int J Anal Chem 2019;2019:7314916.

[107] Lv J, Wei J, Wang Z, Cao J. Multiple compounds recognition from the tandem mass spectral data using convolutional neural network. Molecules 2019;24:4590.

[108] Skarysz A, Alkhalifah Y, Darnley K, Eddleston M, Hu Y, Mclaren DB, Nailon WH, Salman D, Sykora M, Thomas CLP, Soltoggio A. Convolutional neural networks for automated targeted analysis of raw gas chromatography-mass spectrometry data. In: International joint conference on neural networks (IJCNN 2018). Rio de Janeiro, Brazil; 2018, p. 1–8.

[109] Kim S, Lee K, Lee M, Ahn T, Lee J, Suk H, Ning F. Saturation modeling of gas hydrate using machine learning with X-ray CT images. Energies 2020;13:5032.

[110] Li H, Misra S, He J. Neural network modeling of in situ fluid-filled pore size distributions in subsurface shale reservoirs under data constraints. Neural Comput Appl 2020;32:3873–85.

Paper II

Gjelsvik E.L., Fossen M., Brunsvik A., Tøndel K., Identifying components related to hydrate formation by machine-learning based variable selection, in *Tekna Oil Field Chemistry Symposium 2022* (Geilo), March 2022.

Winner of The Terje Østvold Memory Award for Best Paper 2022.

Identifying Components Related to Hydrate Formation by Machine Learning-based Variable Selection

Elise Lunde Gjelsvik[1], Martin Fossen[2], Anders Brunsvik[2], and Kristin Tøndel[1]

[2]*SINTEF AS, Trondheim, Norway*
[1]*Faculty of Science and Technology, Norwegian University of Life Sciences, Aas, Norway*

Abstract

Gas hydrates represent a major flow assurance challenge in the O&G industry as they can cause plugging or complete blockage of a production pipeline. Experiments have shown that some crude oils form gas hydrates that remain as transportable particles in a slurry. This is commonly believed to be due to naturally occurring polar components in the crude oil rendering the surface of the particle hydrophobic. The composition of these components are still not identified. In this study, Fourier transform ion cyclotron resonance mass spectrometry (FT-ICR MS) was used to analyse crude oil samples. Machine learning-based variable selection was applied to find components in the data related to hydrate formation and the best performing model was determined to be gradient boosting. The FT-ICR MS raw spectra of the spiking levels were compared to non-spiked samples to identify changes in composition during the spiking procedure. Principal component analysis showed a difference between the crude oils and the spiked samples. The five most important variables were selected from decision trees, random forest, gradient boosting, bagging, least absolute shrinkage and and selection operator (LASSO), Ridge regression and permutation feature importance for the best performing model. Molecular formulas, double bond equivalent (DBE) and hydrogen-carbon (H/C) ratios were determined for each of the selected variables and evaluated, in an attempt to identify hydrate related components.

Introduction

Gas hydrates are one of the main flow assurance issues during production of oil and gas, as they can agglomerate into larger masses or deposit on the pipe wall, with the potential of completely blocking the system. Most commonly, gas hydrates are treated with addition of chemicals such as thermodynamic inhibitors and low dosage hydrate inhibitors (LDHIs), or by operating outside the hydrate region, by controlling the pressure and/or temperature. However, operating outside the hydrate region is not always possible or economically

feasible, and an addition of large amounts of chemicals has a negative environmental impact. Previous experiments have shown that some crude oils form gas hydrates not prone to plugging, but remain as transportable particles in a slurry [1, 2]. A commonly accepted explanation is that some crude oils contain polar components with hydrate active properties altering the surface of the particles to be hydrophobic. This type of natural inhibitors has been investigated for a long time, yet their exact structures have not been determined [3, 4, 5]. It has been suggested that these components are present in the acid fraction of the crude oils [6, 7, 8] which has been shown to contain a substantial amount of napthenic compounds [9, 10].

Comparatively, it has been shown that some asphaltene fractions in crude oil have self-agglomerating properties that stabilise the water-in-oil emulsions leading to increased probability of deposition on pipe walls [11]. Consequently, asphaltenes can change the plugging potential of some crude oils [12, 13].

Crude oils are highly complex organic mixtures, and with the high resolution of Fourier Transform Ion Cyclotron Resonance Mass Spectrometry (FT-ICR MS) it is possible to analyse and identify a large number of polar groups including compounds present in low concentrations [14]. FT-ICR MS has previously been used extensively for crude oil characterisation [15, 16, 17, 18, 19, 20, 21, 22]. For example, Qian et al. [23, 24] showed that ESI FT-ICR MS was able to identify more than 3000 chemical formulas of acid containing compounds in negative mode, and more than 3000 unique elemental compositions of nitrogen containing aromatic compounds in positive mode.

In this study machine learning-based variable selection methods were applied to highly detailed FT-ICR MS spectra with the objective of identifying naturally occurring hydrate inhibitors.

Experimental

Successive spiking of the hydrate phase

A successive accumulation procedure (spiking) was performed with the aim of accumulating possible hydrate active components. The procedure is shown schematically in Figure 1. In short, a fresh oil sample was added to a high-pressure sapphire cell from Top Industrie (France), located at SINTEF Multiphase Flow Laboatory in Trondheim, with the water and the gas phase pressurised to 65 bar. The temperature was lowered to 2°C while stirring, to provide the conditions necessary for hydrate formation for the specific fluid systems. The formation was normally allowed to proceed over night to come as close to the maximum hydrate formation as possible. Draining of the bulk phase, the liquid not associated to the hydrates, was done through the bottom of the cell, under pressure retaining the gas hydrate phase using a Hy-Lok FT Micron Tee Filter with a 150 μm sintered stainless steel filter element. Once the bulk phase was drained, the cell was depressurised and the temperature was increased leading to melting (dissociation) of the gas hydrate phase which was drained from the cell and collected. Small samples of the oil and water phase were taken from the bulk phase and the hydrate phase for FT-ICR MS analysis. The remaining liquid was mixed with fresh oil and water at a ratio maintaining the same water cut as the first

test and then the hydrate formation was repeated. By conducting this procedure multiple times for a given oil, generations with possibly increased concentration of hydrate active oil components could be accumulated.

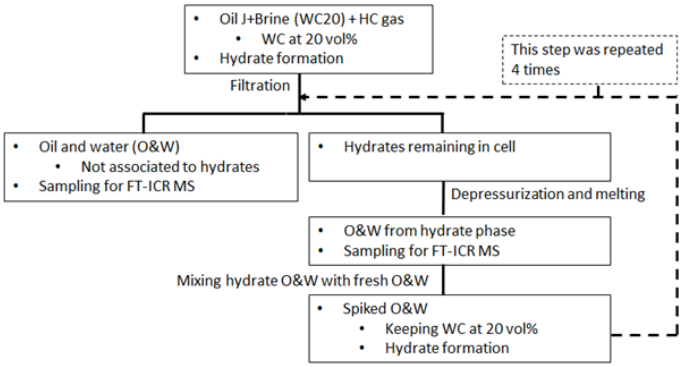


Figure 1: Schematic illustration of the successive accumulation procedure.

FT-ICR MS analysis

For the FT-ICR MS analysis, the oil samples were prepared by dissolving 20 μL sample in 980 μL of dichloromethane and 20 μL diluted sample was added to 980 μL of a 1:1 mixture of toluene and methanol. Then, 100 μL was injected into the FT-ICR MS at a flow of 10 μL/min through a Agilent 1290 Infinty HPLC system. Three parallels were analysed for each sample. The mass spectra were acquired using SINTEF's Bruker Solarix XR Fourier transform ion cyclotron resonance mass spectrometer (FT-ICR MS) (Bruker Daltonik GmbH, Germany) equipped with a 12 Tesla magnet (Bruker Biospin, France) located in Trondheim. Its resolution is 450 000 at m/z 400. The FT-ICR was equipped with an electrospray ion source (ESI) operating in positive mode and the mass range was set to 150-3000 m/z.

PCA

PCA [25] is an unsupervised method for data reduction where a large data set $\boldsymbol{X}$ is decomposed into a subspace of latent variables. Scores ($\boldsymbol{T}$) and loadings ($\boldsymbol{P}$) represent the main features of variance as shown by equation 1.

$$\boldsymbol{X} = \bar{\boldsymbol{x}} + \sum_{a=1}^{A} \boldsymbol{t}_{Xa}\boldsymbol{p}_a^{'} = \bar{\boldsymbol{x}} + \boldsymbol{T}_{XA}\boldsymbol{P}_A^{'} + \boldsymbol{E}_A \tag{1}$$

$\boldsymbol{A}$ denotes the number for Principal components (PCs) used and $\boldsymbol{E}_A$ is the error term using $\boldsymbol{A}$ PCs. The score vectors are orthogonal.

Variable selection methods

Variable selection is the process of selecting a subset of relevant variables for use when constructing a model. When a data set contains a large amount of variables, it is often assumed that the data contains irrelevant or redundant variables that can be removed without loss of information. This can improve the prediction ability of the model and reduce the computational cost during modelling. Variable selection can also be used to identify the features with the highest importance in the model. In this paper, variable selection methods such as Permutation feature importance, Decision trees, Random forest, Boosting, Bagging and regularisation methods such as Ridge regression and LASSO were used to predict whether the samples were related to the hydrate or the bulk phase, and to identify components in the data related to hydrate formation with the hypothesis that there could be systematic differences in these spectra possible to distinguish with the proposed methods.

Permutation feature importance

Permutation feature importance is a model inspection technique that identifies important variables based on changes in the prediction accuracy when a variable is randomly shuffled (permuted) [26]. If the prediction accuracy of the model decreases significantly when a variable is randomly shuffled, this indicates that the variable is important for the model's ability to predict the response. Similarly, if the prediction accuracy is unaffected when a variable is randomly shuffled, the variable is not important for the prediction. The importance of the variables is calculated from equation 2.

$$\boldsymbol{i}_j = \boldsymbol{s} - \frac{1}{\boldsymbol{K}} \sum_{k=1}^{K} \boldsymbol{s}_{k,j} \tag{2}$$

Where $\boldsymbol{s}$ is the reference prediction accuracy of the model with the original features, $\boldsymbol{s}_{k,j}$ is the prediction accuracy of the models with shuffled variables and $\boldsymbol{K}$ is the number of variables.

Decision Trees (DTs)

DTs [27, 28] are models where decisions are made by asking a series of questions and generating decision rules based on these. These models consist of a tree root, internal nodes, branches and leaf nodes. They aim to find the smallest set of rules that is consistent with the training data. In general, the rules have the form: *if* $condition_1$ *and* $condition_2$ *and* $condition_3$ *then outcome* and are chosen to divide observations into segments that have the largest difference with respect to the target variable. Therefore, the rule selects both the variable and the best break point (usually selected by significance testing or reduction in variance criteria) for maximal separation of the resulting subgroups. Figure 2 shows an illustration of a decision tree model.

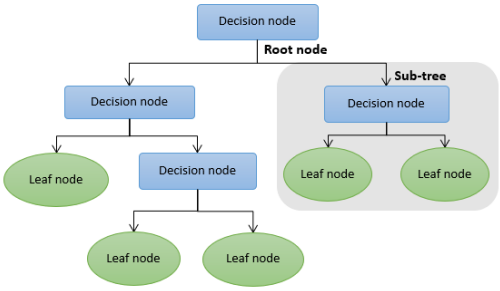


Figure 2: Illustration of decision trees

To avoid overfitting, the trees often have to be pruned by setting a limit for the maximal depth. A leaf can no longer be split when there are too few observations, the maximum depth (the hierarchy of the tree) has been reached, or no significant split can be identified. It is assumed that observations belonging to different classes have different values in at least one of their features. DTs are usually univariate, since they use splits based on a single feature at each internal node.

Random forest (RF)

In DTs, the initial selected split effects the optimality of variables considered for subsequent splits, making these methods prone to overfitting and other problems. This can be handled by introducing RF [29, 26, 30], an ensemble tree method where each tree is based on a random subset of the data and its features (selected by bootstrapping). The advantage of ensemble trees is that the trees are grown with varying initial splits, and either a voting or the average of the predictions for each new data point across all trees is used. The vote distribution can be used to develop a non-parametric probabilistic predictive model. The change in prediction accuracy when the values of a feature are randomly permuted among the observations gives estimates of the importance of each feature.

Ensemble learning

Ensemble learning combines weak classification models with the main idea that many models in combination perform better than one model alone [31].

Bootstrap aggregating (bagging) [29] is an ensemble learner where random subsets of the data set are generated and models are trained individually based on these bootstrapped data sets. The ensembles overall decision is achieved by aggregating the results from the individual models. This leads to a reduction in the risk of overfitting as the method combines different models built from different subsets of the available data.

Boosting [32] is an ensemble learner where weak learners are trained sequentially, trying to improve upon its predecessor. The classifiers emphasise errors made by the previous classifier, aiming at decreasing the model bias. Boosting learners combine underfitting models with low prediction accuracy with the aim of improving the final prediction. Gradient Boosting [33, 34] is one boosting method where trees are built in every iteration, always

minimising the prediction error of the classifier. This combination of several smaller trees forms a stronger learner able to fit larger parts of the data than a simple decision tree can.

Regularisation

Another group of machine learning methods valuable for variable selection purposes are regularisation-based methods, the most common being Ridge regression [35], LASSO (least absolute shrinkage and selection operator) [36] and Elastic net [37].

Ridge regression

Ridge regression is also known as L2-regularisation. In Ridge, the sum of the squares of the regression coefficients is forced to be less than a fixed value, which shrinks the size of the coefficients. Ordinary least squares (OLS) minimises equation 3,

$$RSS_{OLS} = \sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij})^2 \tag{3}$$

while Ridge regression minimises equation 4.

$$RSS_{Ridge} = \sum_{i=1}^{p}(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij})^2 + \lambda\sum_{i=1}^{p}\beta_j^2 = RSS_{OLS} + \lambda\sum_{i=1}^{p}\beta_j^2 \tag{4}$$

where $\lambda \geq 0$ is a penalty term which is often found by cross-validation. This gives equation 5 and 6.

$$B_{OLS} = (X^{'}X)^{-1}X^{'}Y \tag{5}$$

$$B_{Ridge} = (X^{'}X + \lambda I)^{-1}X^{'}Y \tag{6}$$

Hence, Ridge regression handles multicollinearity in the regressor (X) matrix, while OLS regression does not.

LASSO

In LASSO the estimates of the regression coefficients are obtained using L1-constrained least squares. This forces the sum of the absolute values of the regression coefficients to be less than a fixed value, which forces certain coefficients to be set to zero. LASSO is a feature selection method, since features having zero regression coefficients are omitted from the model. LASSO minimises equation 7.

$$RSS_{LASSO} = \sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij})^2 + \lambda\sum_{i=1}^{p} \mid \beta_j \mid = RSS_{OLS} + \lambda\sum_{i=1}^{p} \mid \beta_j \mid \tag{7}$$

Elastic Net

Elastic net combines the L1 and L2 penalties of the Ridge and LASSO methods linearly as given by equation 8.

$$RSS_{EN} = \sum_{i=1}^{n}(y_i - \beta_0 \sum_{j=1}^{p} \beta_j x_{ij})^2 + \lambda_1 \sum_{i=1}^{p} \beta_j^2 + \lambda_2 \sum_{i=1}^{p} \mid \beta_j \mid = RSS_{OLS} + \lambda_1 \sum_{i=1}^{p} \beta_j^2 + \lambda_2 \sum_{i=1}^{p} \mid \beta_j \mid \tag{8}$$

In Elastic net, highly correlated regressors will tend to have similar regression coefficients, which creates a grouping effect that is desirable in many applications.

Variable importance score

For each of the variable selection methods a variable importance score can be computed, which is a measure of the variables' relative importance in the prediction model. These scores therefore reflect which variables are the most relevant for the target and which variables are of least importance.

The variable importance score can also be used to improve the prediction model by including only the variables with high scores in the model.

Data Analysis

Two oil samples (A and J2) from the Norwegian continental shelf were received from the project partners. The wetting index for oil J2 was determined by Fossen et al. in [38] to be +0.44 and the same procedure was used to determine the wetting index for oil A to be 0. The samples underwent the successive accumulation procedure resulting in 26 samples with different spiking levels, 6 spiking levels for oil A and 4 spiking levels for oil J2. The samples were analysed by FT-ICR MS. The response consisted of a vector containing information of the samples origin, i.e. whether it was extracted from the bulk phase or from the hydrate phase during the successive accumulation procedure. Several different variable selection methods were tested to attempt to find features related to hydrate formation.

A bucket table was created of the data set using Bruker Compass ProfileAnalysis 2.1. The statistical methods were implemented using Python 3.8 and its machine learning packages. Molecular formulas were determined using Bruker Compass DataAnalysis 5.0 and their molecular structures were investigated using SciFinder. All models were validated using training and test sets.

Results and discussion

PCA

Each of the oil samples were analysed by PCA and the resulting scoreplot of the first principal component (PC1) and the second principal component (PC2) for oil A is shown in Figure 3. The scoreplot to the left shows the samples from the bulk phase, and the plot to the right shows the samples from the hydrate phase, coloured by their corresponding

spiking levels. Both plots show that the spiking level of the samples increases along PC2, meaning that the spiking level is explained by PC2. The crude oil sample and the spiked samples are separated along PC1. Hence PC1 explains the difference between the crude oil and the samples that have undergone the successive accumulation procedure.

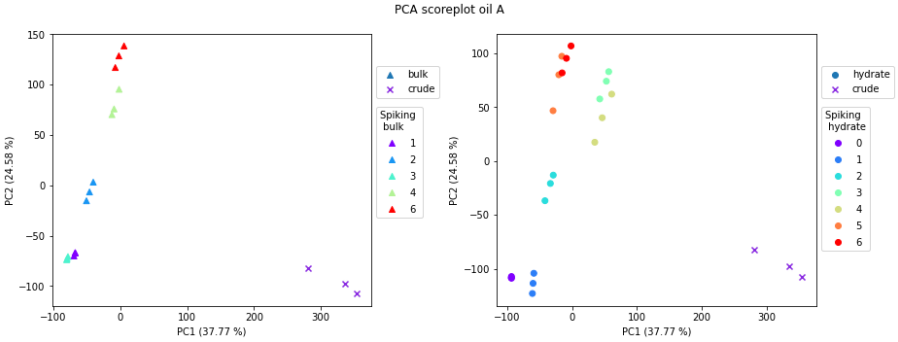


Figure 3: Scoreplot from PCA for oil A where samples from the bulk phase are shown in the left plot and the samples from the hydrate phase are shown in the right plot. The crude oil is marked with x.

The results from the PCA of oil J2 are shown in Figure 4. The scoreplot to the left shows the samples from the bulk phase, and the plot to the right shows the samples from the hydrate phase, coloured by their corresponding spiking levels. The results are to a high degree similar to those for oil A. PC1 explains the difference between the spiked samples and the crude oil samples. PC2 explains the spiking levels, but the groupings for the spiking levels of oil J2 are not as clear as for oil A.

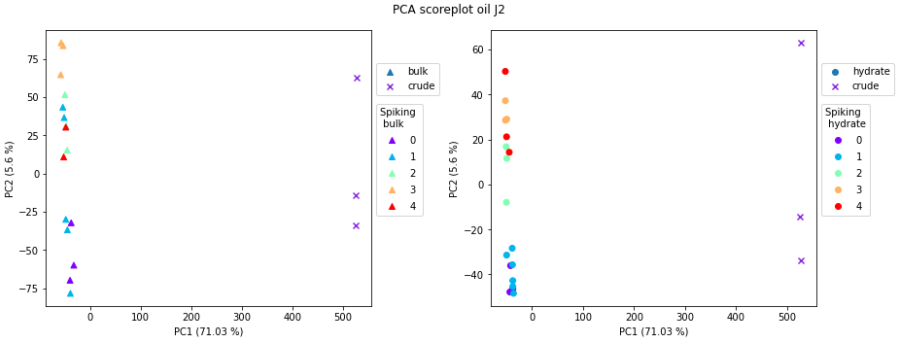


Figure 4: Scoreplot from PCA for oil J2 where samples from the bulk phase are shown in the left plot and the samples from the hydrate phase are shown in the right plot. The crude oil is marked with x.

We see clearly from the scoreplots for both oil A and J2 that the crude oil is distinguishable from the spiking samples both in the bulk phase and in the hydrate phase. The successful accumulation procedure therefore alters the composition of the oil sufficiently to observe this separation between the spiking levels and the crude. This suggests that there is an accumulation of hydrate active components.

Comparing raw spectra

To investigate differences between the spiking fractions, the average raw spectra for each spiking sample were compared to a sample which had not been spiked. This was performed by subtracting the spectra of a sample removed before the successive accumulation from the spectra of the remaining spiked samples to identify a possible increase or decrease in any of the peaks during the accumulation. The results for oil A are shown in Figure 5. Three m/z values appeared to increase as the spiking level increased: 326.38, 469.31 and 229.14. Table 1 shows the molecular formulas for these compounds, their double bond equivalent (DBE) numbers, which is the degree of unsaturation of the molecule, and their hydrogen-carbon (H/C) ratios. Two of the compounds have DBE numbers of 0 (m/z 326.38 and 229.14) meaning that they are saturated. It is therefore likely that they are paraffins.

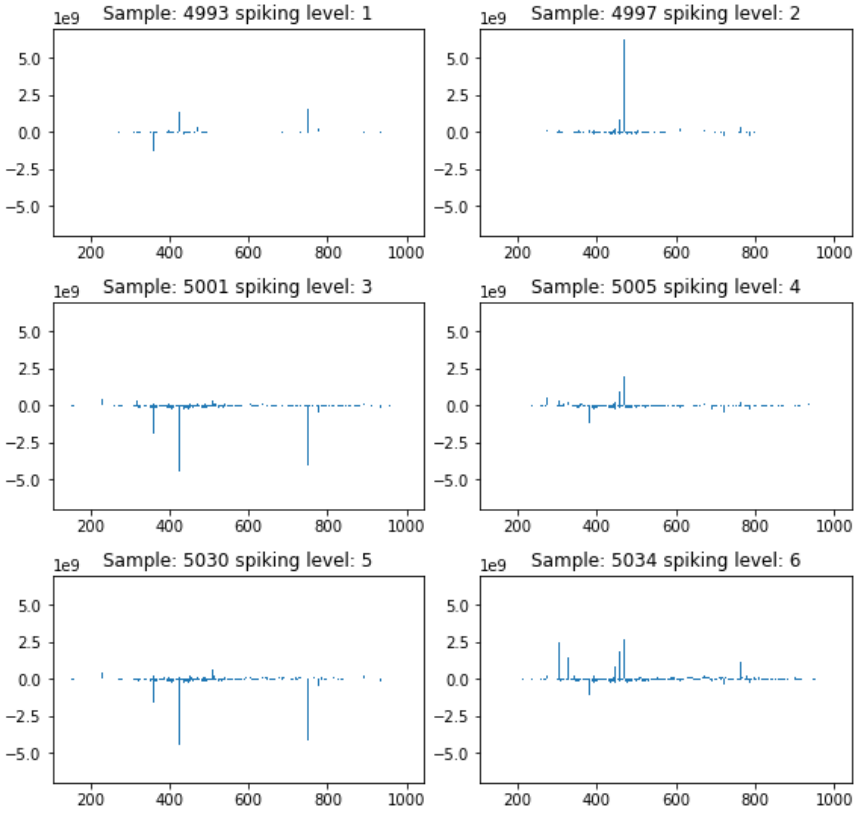


Figure 5: Raw spectra of oil A from hydrate phase compared to spiking level 0

m/z	Formulae	DBE	H/C
326.38	$C_{22}H_{48}N$	0	2.18
469.31	$C_{22}H_{50}N_4OSV$	4	2.72
229.14	$C_{10}H_{22}NaO_4$	0	2.20

Table 1: The m/z values in oil A that increased as spiking level increased, their molecular formula, DBE and hydrogen-carbon ratio.

Spiking level 0 was also compared to the remaining spectra for oil J2 and the results are shown in Figure 6. For oil J2 no distinct m/z values increased with increasing spiking level. However, Figure 6 shows that for spiking level 2, 3 and 4, the area between 400 and 600 m/z increased, indicating that the variables of interest with regards to hydrate formation may lie in this area. Another possible explanation is that this oil is saturated with hydrate active components, and the spiking procedure therefore does not change the composition of the oil to any significant extent.

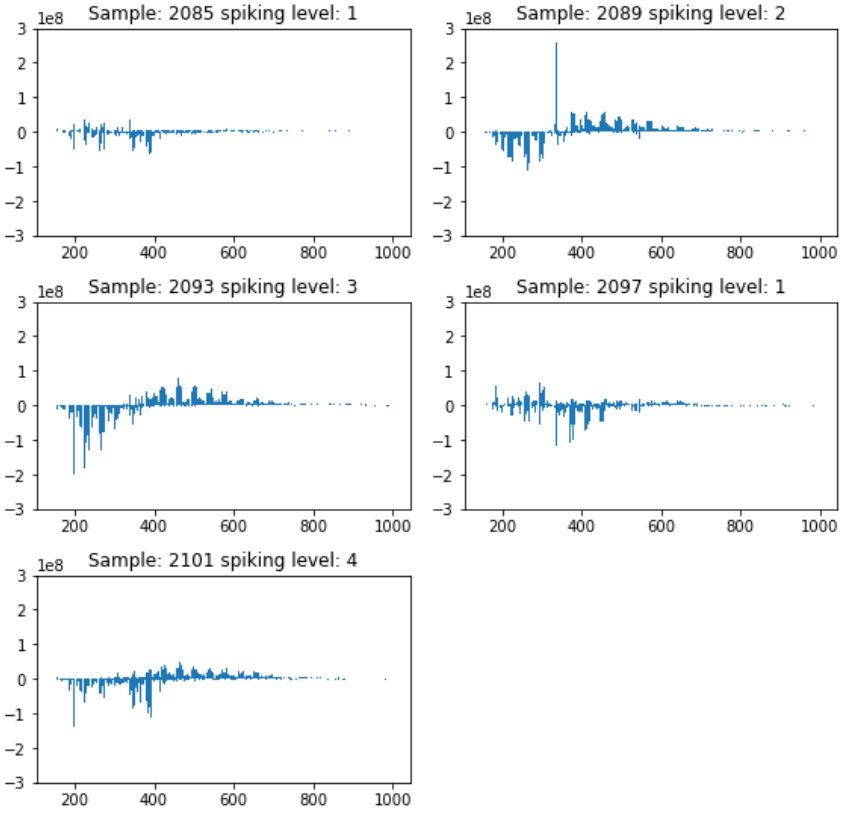


Figure 6: Raw spectra of oil J2 from hydrate phase compared to spiking level 0

Variable selection methods

Several variable selection methods such as decision trees, random forest, bagging, gradient boosting, and regularisation methods such as LASSO and ridge regression, were tested with the aim of finding components related to hydrate formation. The samples were classified as either bulk (0) or hydrate (1). The accuracy scores for each of the classification methods are shown in Figure 7. Figure 7 shows that the best performing prediction model is gradient boosting which achieves the highest prediction accuracy score of 0.7. Figure 7 shows that two of the other tree based methods, decision trees and random forest, also perform well. Tree based models often outperform linear models when the variables in the data are highly correlated to each other. This can be remedied by using regularisation and for this data set, ridge regression improves the model considerably compared to LASSO regularisation. The gradient boosting model seems to be able to create a model well fitted to the differences, e.g different spiking levels and two different oils that this data contains.

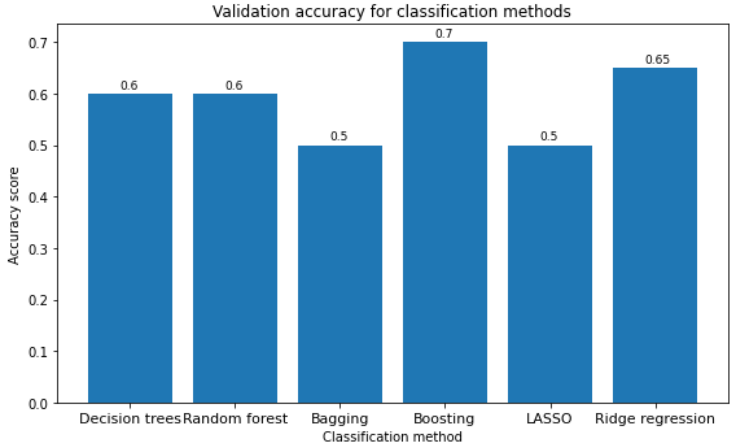


Figure 7: Accuracy scores for the different classification methods

For each of the variable selection methods the five m/z-values that received the highest variable importance scores were extracted and are shown in table 2. Permutation feature importance was applied to the best performing model, gradient boosting. Molecular formulas corresponding to these m/z-values were determined and their molecular structures were investigated to identify commonalities in the selected variables. Table 2 shows a large range in the m/z-values selected as important by the variable selection methods, stretching from 243.20 to 781.56 and the carbon chains ranges from C_{10} to C_{46}. The DBE numbers show that two of the selected variables are saturated while the remaining variables have DBE numbers between 1 and 25. Inspection of the suggested molecular structures revealed mostly long chained hydrocarbons with increasing number of rings as the DBE number increases. Seawater is present in the samples and therefore, some of the variables selected contain sodium complexes. Some of the oils contain sulphur and a three consist of a metal complex with vanadium.

The average weight of asphaltenes is ∼750 Da, and some of the selected m/z-values lie in the range of 709-782 indicating that these could be asphaltenes, apart from the two (m/z 716.62 and 781.56) with DBE numbers of 0. The asphaltene fractions able to stabilise water-in-oil emulsions are often more polar, with higher oxygen content and thereby higher acidity and lower DBEs [39]. Agglomeration have also been related to the presence of sulfoxides (O_xS_y) [40], which can be seen in all of the m/z ratios between 709-782 where all are detected in oil A. This corresponds well with the wetting index for oil A of 0, which means that it can be a plugging oil.

The mass spectra were inspected to identify which samples contained the selected m/z-values. From table 2 it is clear that most of the selected variables are in either oil A or oil J2. Which sample each m/z-value is detected in is described in Appendix A. As the response used in the machine learning methods is whether the sample is related to hydrate formation or not, the variables selected by the methods should be related to hydrate formation. However, if the oil samples are too different, the differences will overshadow the underlying effects between the response and the variables, and variables separating the samples will be selected. As this data set only contains two oils, it is difficult to determine which of these effects are evident in these variables.

It was not possible to determine the molecular formulas for the last five variables, m/z 436.57, 545.84, 545.72, 436,67 and 545.83. The spectra showed that these peaks were only detected in four of the samples from oil J2, and are most likely peaks stemming from background contamination or instrument noise.

m/z	Formula	DBE	H/C	Detected in	Method
388.19	$C_{18}H_{30}NO_8$	5	1.67	A	
709.47	$C_{41}H_{73}O_3S_3$	12	1.78	A J2	Permutation
460.24	$C_{31}H_{30}N_3O$	19	0.97	J2	feature importance
422.22	$C_{16}H_{32}N_5O_8$	4	2.0	J2	
716.62	$C_{39}H_{91}N_3O_2SV$	0	2.33	J2	
608.41	$C_{29}H_{58}N_3O_{10}$	3	2.0	A	
618.56	$C_{40}H_{76}NOS$	6	1.90	A	Random forest
320.21	$C_{21}H_{26}N_3$	11	1.24	J2	
781.56	$C_{33}H_{90}N_6O_4S_3V$	0	2.73	A	
766.69	$C_{46}H_{97}NNaOS_2$	3	2.11	A	
709.47	$C_{41}H_{73}O_3S_3$	12	1.78	A J2	
351.16	$C_{11}H_{31}N_2O_6S_2$	2	2.81	J2	Boosting
243.20	$C_{13}H_{27}N_2O_2$	2	2.08	J2	
357.17	$C_{10}H_{25}N_6O_8$	2	2.5	J2	
571.47	$C_{32}H_{69}N_2NaOV$	1	2.16	A J2	
641.37	$C_{30}H_{58}N_4NaO_5S_2$	8	1.93	J2	
588.40	$C_{39}H_{50}N_5$	18	1.28	A	Bagging
392.31	$C_{22}H_{43}NNaO_3$	6	1.95	A	
655.24	$C_{40}H_{35}N_2O_7$	25	0.88	A	
340.11	$C_{15}H_{14}N_7O_3$	13	0.93	A	
463.25	$C_{18}H_{35}N_6O_8$	5	1.94	J2	
753.50	$C_{40}H_{70}N_6NaO_4S$	11	1.75	A	LASSO
382.22	$C_{18}H_{33}NNaO_5$	3	1.83	A	
436.57	ND			J2	
436.57	ND			J2	
545.84	ND			J2	
545.72	ND			J2	Ridge Regression
436.67	ND			J2	
545.83	ND			J2	

Table 2: Table of the five m/z values selected as most important from each method, their molecular formulas, DBE numbers, hydrate-carbon ratio and in which oil the m/z values are detected. Molecular formulas that could not be identified are labelled ND.

Future work

In this study, several m/z-values were determined by the machine learning-based variable selection methods to have a higher relevance to the differences between samples from the hydrate phase and samples from the bulk phase. To try to pinpoint the exact molecular structures of the selected m/z-values, the peaks will be isolated by FT-ICR MS and fragmented. This will make it easier to identify the structures of complicated molecules. When the compounds are found, they can be tested with the oils to evaluate how their presence changes the characteristics of the oils and the formation of hydrates.

In this study only two oil samples were examined. The successive accumulation procedure will be performed for additional oils and the methods presented in this paper will be used to identify important variables. The results will be compared between all oils with the aim of identifying component groups that are related to hydrate formation. The possible presence of one common compound between the oils will be investigated. The ultimate aim of the work is to conclude on an optimal method which identifies components related to hydrate formation based on the spiking procedure and FT-ICR MS data.

Conclusion

Several machine learning-based variable selection methods were tested with the aim of finding components related to hydrate formation. The best performing prediction model was gradient boosting with an accuracy score of 0.7. The m/z-values of highest importance for the response were identified from all models and their molecular formulas were determined to attempt to identify a group of molecules related to hydrate formation. Some of the variables were identified as possible asphaltenic structures and believed to contribute to the wetting index of 0 for oil A.

Identifying the variables in the oils that are important for the formation of hydrates takes us one step closer to identifying the nature and molecular structure of naturally occurring hydrate active components.

Acknowledgement

The authors thank The Research Council of Norway, Equinor ASA, OMV (Norge) AS, Wintershall DEA Norge AS and TotalEnergies for funding. This work is a part of the Knowledge-Building Project for Industry (PETROMAKS 2), Project number: 294636 “New Hydrate Management: New understanding of hydrate phenomena in oil systems to enable safe operation within the hydrate zone”.

References

[1] M. N. Lingelem, A. I. Majeed, and E. Stange, “Industrial Experience in Evaluation of Hydrate Formation, Inhibition, and Dissociation in Pipeline Design and Operation,” *Annals of the New York Academy of Sciences*, vol. 715, pp. 75–93, Apr. 1994.

[2] F. H. Fadnes, “Natural hydrate inhibiting components in crude oils,” *Fluid Phase Equilibria*, vol. 117, pp. 186–192, Mar. 1996.

[3] P. Fotland and K. M. Askvik, “Some aspects of hydrate formation and wetting,” *Journal of Colloid and Interface Science*, vol. 321, pp. 130–141, May 2008.

[4] A. E. Borgund, S. Høiland, T. Barth, P. Fotland, and K. M. Askvik, “Molecular analysis of petroleum derived compounds that adsorbonto gas hydrate surfaces,” *Applied Geochemistry*, vol. 24, pp. 777–786, Jan. 2009.

[5] S. Høiland, A. E. Borglund, T. Barth, P. Fotland, and K. M. Askvik, "Wettability of Freon hydrates in crude oil/brine emulsions: the effects of chemical additives.," in *5th International Conference in Gas Hydrate*, vol. 4, (Trondheim), pp. 1151–1161, June 2005.

[6] K. Erstad, S. Høiland, P. Fotland, and T. Barth, "Influence of Petroleum Acids on Gas Hydrate Wettability," *Energy & Fuels*, vol. 23, pp. 2213–2219, Feb. 2009.

[7] P. V. Hemmingsen, X. Li, J.-L. Peytavy, and J. Sjöblom, "Hydrate Plugging Potential of Original and Modified Crude Oils," *Journal of Dispersion Science and Technology*, vol. 28, pp. 371–382, Apr. 2007.

[8] P. V. Hemmingsen, S. Kim, H. E. Pettersen, R. P. Rodgers, J. Sjöblom, and A. G. Marshall, "Structural Characterization and Interfacial Behavior of Acidic Compounds Extracted from a North Sea Oil," *Energy & Fuels*, vol. 20, pp. 1980–1987, July 2006.

[9] A. E. Borgund, K. Erstad, and T. Barth, "Fractionation of Crude Oil Acids by HPLC and Characterization of Their Properties and Effects on Gas Hydrate Surfaces," *Energy & Fuels*, vol. 21, pp. 2816–2826, July 2007.

[10] J. S. Clemente and P. M. Fedorak, "A review of the occurrence, analyses, toxicity, and biodegradation of naphthenic acids," *Chemosphere*, vol. 60, pp. 585–600, July 2005.

[11] P. Qiao, D. Harbottle, P. Tchoukov, J. Masliyah, J. Sjoblom, Q. Liu, and Z. Xu, "Fractionation of Asphaltenes in Understanding Their Role in Petroleum Emulsion Stability and Fouling," *Energy Fuels*, vol. 31, pp. 3330–3337, Dec. 2016.

[12] D. C. Salmin, *The Impact of Synthetic and Natural Surface-Active Components on Hydrate Agglomeration*. Doctoral thesis, Colorado School of Mines, Golden, Colorado, 2019.

[13] J. J. Adams, "Asphaltene Adsorption, a Literature Review," *Energy Fuels*, vol. 28, pp. 2831–2856, Mar. 2014.

[14] M. R. Emmett, F. M. White, C. L. Hendrickson, S. D.-H. Shi, and A. G. Marshall, "Application of micro-electrospray liquid chromatography techniques to FT-ICR MS to enable high-sensitivity biological analysis," *Journal of the American Society for Mass Spectrometry*, vol. 9, pp. 333–340, Apr. 1998.

[15] Y. Cho, A. Ahmed, A. Islam, and Sunghwan Kim, "Developments in FT-ICR MS instrumentation, ionization techniques, and data interpretation methods for petroleomics," *Mass Spectrometry Reviews*, vol. 34, pp. 248–263, Mar. 2014.

[16] M. Hur, I. Yeo, E. Kim, M.-h. No, J. Koh, Y. J. Cho, J. W. Lee, and S. Kim, "Correlation of FT-ICR Mass Spectra with the Chemical and Physical Properties of Associated Crude Oils," *Energy & Fuels*, vol. 24, pp. 5524–5532, Aug. 2010.

[17] G. C. Klein, S. Kim, R. P. Rodgers, A. G. Marshall, and A. Yen, "Mass Spectral Analysis of Asphaltenes. II. Detailed Compositional Comparison of Asphaltenes Deposit to Its Crude Oil Counterpart for Two Geographically Different Crude Oils by ESI FT-ICR MS," *Energy & Fuels*, vol. 20, pp. 1973–1979, Sept. 2006.

[18] T. M. Schaub, D. W. Jennings, S. Kim, R. P. Rodgers, and A. G. Marshall, "Heat-Exchanger Deposits in an Inverted Steam-Assisted Gravity Drainage Operation. Part 2. Organic Acid Analysis by Electrospray Ionization Fourier Transform Ion Cyclotron Resonance Mass Spectrometry," *Energy & Fuels*, vol. 21, pp. 185–194, Jan. 2007.

[19] D. F. Smith, P. Rahimi, A. Teclemariam, R. P. Rodger, and A. G. Marshall, "Characterization of Athabasca Bitumen Heavy Vacuum Gas Oil Distillation Cuts by Negative/Positive Electrospray Ionization and Automated Liquid Injection Field Desorption Ionization Fourier Transform Ion Cyclotron Resonance Mass Spectrometry," *Energy & Fuels*, vol. 22, pp. 3118–3125, Sept. 2008.

[20] J. V. Headley, K. M. Peru, M. P. Barrow, and P. J. Derrick, "Characterization of Naphthenic Acids from Athabasca Oil Sands Using Electrospray Ionization: The Significant Influence of Solvents," *Analytical Chemistry*, vol. 79, pp. 6222–6229, Aug. 2007.

[21] M. P. Barrow, J. V. Headley, K. M. Peru, and P. J. Derrick, "Data Visualization for the Characterization of Naphthenic Acids within Petroleum Samples," *Energy & Fuels*, vol. 23, pp. 2592–2599, Mar. 2009.

[22] F. A. Fernandez-Lima, C. Becker, A. M. McKenna, R. P. Rodgers, A. G. Marshall, and D. H. Russell, "Petroleum Crude Oil Characterization by IMS-MS and FTICR MS," *Analytical Chemistry*, vol. 81, pp. 9941–9947, Dec. 2009.

[23] K. Qian, R. P. Rodgers, C. L. Hendrickson, M. R. Emmett, and A. G. Marshall, "Reading Chemical Fine Print: Resolution and Identification of 3000 Nitrogen-Containing Aromatic Compounds from a Single Electrospray Ionization Fourier Transform Ion Cyclotron Resonance Mass Spectrum of Heavy Petroleum Crude Oil," *Energy & Fuels*, vol. 15, pp. 492–498, Jan. 2001.

[24] K. Qian, W. K. Robbins, C. A. Hughey, H. J. Cooper, R. P. Rodgers, and A. G. Marshall, "Resolution and Identification of Elemental Compositions for More than 3000 Crude Acids in Heavy Petroleum by Negative-Ion Microelectrospray High-Field Fourier Transform Ion Cyclotron Resonance Mass Spectrometry," *Energy and Fuels*, vol. 15, pp. 1505–1511, July 2001.

[25] K. Pearson, "On lines and planes of closest fit to systems of points in space," *Philosophical Magazine*, vol. 2, pp. 559–572, 1901.

[26] L. Breiman, "Random Forests," *Machine Learning*, vol. 45, pp. 5–32, Oct. 2001.

[27] J. R. Quinlan, "Simplifying decision trees," *International Journal of Man-Machine Studies*, vol. 27, pp. 221–234, Sept. 1987.

[28] P. E. Utgoff, "Incremental Induction of Decision Trees," *Machine Learning*, vol. 4, pp. 161–186, Nov. 1989.

[29] L. Breiman, "Bagging predictors," *Machine Learning*, vol. 24, pp. 123–140, Aug. 1996.

[30] T. K. Ho, "The random subspace method for constructing decision forests," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, pp. 832 – 844, Aug. 1998.

[31] L. Breiman, "Arcing classifier (with discussion and a rejoinder by the author)," *Annals of Statistics*, vol. 26, no. 3, pp. 801–849, 1998.

[32] R. E. Schapire, "The strength of weak learnability," *Machine Learning*, vol. 5, pp. 197–227, June 1990.

[33] J. H. Friedman, "Greedy Function Approximation: A Gradient Boosting Machine," *The Annals of Statistics*, vol. 29, pp. 1189–1232, Oct. 2001.

[34] J. H. Friedman, "Stochastic gradient boosting," *Computational Statistics & Data Analysis*, vol. 38, pp. 367–378, Feb. 2002.

[35] A. E. Hoerl, "Application of Ridge Analysis to Regression Problems," *Chemical Engineering Progress*, vol. 58, no. 3, pp. 54–59, 1958.

[36] R. Tibshirani, "Regression Shrinkage and Selection Via the Lasso," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 58, no. 1, pp. 267–288, 1996.

[37] H. Zou and T. Hastie, "Regularization and variable selection via the elastic net," *The Journal of the Royal Statistical Society, Series B (Statistical Methodology)*, vol. 67, pp. 301–320, Apr. 2005.

[38] M. Fossen, M. Wolden, and A. Brunsvik, "Successive accumulation of naturally occurring hydrate active components and the effect on the wetting properties," in *32nd Oil Field Chemistry Symposium 2021*, p. 16, TEKNA, May 2021.

[39] P. K. Kilpatrick, "Water-in-Crude Oil Emulsion Stabilization: Review and Unanswered Questions," *Energy Fuels*, vol. 26, pp. 4017–4026, May 2012.

[40] F. Yang, P. Tchoukov, H. Dettman, R. B. Teklebrhan, L. Liu, T. Dabros, J. Czarnecki, J. Masliyah, and Z. Xu, "Asphaltene Subfractions Responsible for Stabilizing Water-in-Crude Oil Emulsions. Part 2: Molecular Representations and Molecular Dynamics Simulations," *Energy Fuels*, vol. 29, pp. 4783–4794, July 2015.

		VIP				Random forest					Boosting					Bagging					LASSO					Ridge regression				
Oil	Sample	388,19	709,47	460,24	422,22	716,62	608,41	618,56	320,21	781,56	766,69	709,47	351,16	243,2	357,17	571,47	641,37	588,4	392,31	655,24	340,11	463,25	753,5	382,22	436,57	436,57	545,84	545,72	436,67	545,83
	2058																													
	2080		x	x	x	x			x			x	x	x	x	x	x					x								
	2082		x	x	x	x			x			x	x	x	x	x	x					x								
	2085			x	x	x			x				x	x	x	x	x					x								
	2087			x	x	x			x				x	x	x	x	x					x								
	2089		x	x	x	x			x			x	x	x	x	x	x					x					x		x	x
J2	2091		x	x	x	x			x			x	x	x	x	x	x					x					x		x	x
	2093		x	x	x	x			x			x	x	x	x	x	x					x			x	x		x		
	2095				x	x			x				x	x	x	x	x					x			x	x		x		
	2097			x	x	x			x				x	x	x	x	x					x								
	2099		x	x	x	x			x			x	x	x	x	x	x					x								
	2101			x	x	x			x				x	x	x	x	x					x								
	2103		x	x	x	x			x			x	x	x	x	x	x					x								
	2057																													
	4988		x					x			x	x				x			x					x						
	4993	x	x				x	x		x	x	x				x		x	x				x	x						
	4995		x					x		x	x	x				x		x	x				x	x						
	4997	x	x					x		x	x	x				x			x				x	x						
	4999		x					x		x	x	x				x			x					x						
A	5001	x					x	x		x	x					x		x	x				x	x						
	5003						x	x		x	x					x		x	x		x		x	x						
	5005	x	x				x	x		x	x	x				x		x	x	x			x	x						
	5030	x					x	x		x	x					x		x	x				x	x						
	5031	x					x	x		x	x					x		x	x		x		x	x						
	5034		x				x	x		x	x	x				x		x	x		x		x	x						
	5053	x					x	x		x	x					x		x	x				x	x						

Crude
Hydrate
Bulk

Appendix A
Overview of m/z-values selected by each variable selection method and in which samples the m/z-values were found.

Paper III

Gjelsvik E.L., Fossen M., Brunsvik A., Tøndel K. (2022), Using machine-learning based variable selection to identify hydrate related components from FT-ICR MS spectra *PLoS ONE*, 17(8): e0273084, doi: 10.1371/journal.pone.0273084

PLOS ONE

RESEARCH ARTICLE

Using machine learning-based variable selection to identify hydrate related components from FT-ICR MS spectra

Elise Lunde Gjelsvik[1☯]*, **Martin Fossen**[2☯], **Anders Brunsvik**[2☯], **Kristin Tøndel**[1☯]

1 Norwegian University of Life Sciences, Faculty of Science and Technology, Aas, Norway, **2** SINTEF AS, Trondheim, Norway

☯ These authors contributed equally to this work.
* elise.lunde.gjelsvik@nmbu.no

Citation: Gjelsvik EL, Fossen M, Brunsvik A, Tøndel K (2022) Using machine learning-based variable selection to identify hydrate related components from FT-ICR MS spectra. PLoS ONE 17(8): e0273084. https://doi.org/10.1371/journal.pone.0273084

Editor: Joseph Banoub, Fisheries and Oceans Canada, CANADA

Received: May 6, 2022

Accepted: August 2, 2022

Published: August 17, 2022

Peer Review History: PLOS recognizes the benefits of transparency in the peer review process; therefore, we enable the publication of all of the content of peer review and author responses alongside final, published articles. The editorial history of this article is available here: https://doi.org/10.1371/journal.pone.0273084

Data Availability Statement: All 1 files are available from the Zenodo database (doi: 10.5281/zenodo.6524710).

Abstract

The blockages of pipelines caused by agglomeration of gas hydrates is a major flow assurance issue in the oil and gas industry. Some crude oils form gas hydrates that remain as transportable particles in a slurry. It is commonly believed that naturally occurring components in those crude oils alter the surface properties of gas hydrate particles when formed. The exact structure of the crude oil components responsible for this surface modification remains unknown. In this study, a successive accumulation and spiking of hydrate-active crude oil fractions was performed to increase the concentration of hydrate related compounds. Fourier Transform Ion Cyclotron Resonance Mass Spectrometry (FT-ICR MS) was then utilised to analyse extracted oil samples for each spiking generation. Machine learning-based variable selection was used on the FT-ICR MS spectra to identify the components related to hydrate formation. Among six different methods, Partial Least Squares Discriminant Analysis (PLS-DA) was selected as the best performing model and the 23 most important variables were determined. The FT-ICR MS mass spectra for each spiking level was compared to samples extracted before the successive accumulation, to identify changes in the composition. Principal Component Analysis (PCA) exhibited differences between the oils and spiking levels, indicating an accumulation of hydrate active components. Molecular formulas, double bond equivalents (DBE) and hydrogen-carbon (H/C) ratios were determined for each of the selected variables and evaluated. Some variables were identified as possibly asphaltenes and naphthenic acids which could be related to the positive wetting index (WI) for the oils.

Introduction

One of the major flow assurance challenges in the oil and gas industry is the formation of gas hydrates and their agglomeration, causing complete blockage of pipelines [1]. Gas hydrates are formed under low temperatures and high pressures, as guest molecules are trapped inside and help stabilise crystalline cages consisting of water molecules held together by hydrogen bonds.

Funding: Funding from the Norwegian Research Council, Equinor ASA, OMV (Norge) AS, Wintershall DEA Norge AS and TotalEnergies. This work is a part of the Knowledge-Building Project for Industry (PETROMAKS 2), Project number: 294636 "New Hydrate Management: New understanding of hydrate phenomena in oil systems to enable safe operation within the hydrate zone".

Competing interests: The authors have declared that no competing interests exist.

Remediation methods consists of thermodynamic inhibitors (methanol, ethanol or glycols), low dosage hydrate inhibitors (LDHIs), or by ensuring operation outside the hydrate region by controlling the pressure and/or temperature [2]. However, operating outside the hydrate region is not always possible or economically feasible and chemicals have negative environmental impacts and should be avoided if possible. Previous experiments have shown that some crude oils form gas hydrates that do not agglomerate or deposit, but remain as transportable particles [3–5]. This can be explained by the existence of naturally occurring components in the crude oils with hydrate active properties that can interact with and alter the surface wetting properties of the hydrate particles from being hydrophilic to becoming hydrophobic, thus preventing agglomeration [6]. Despite a lot of research on the topic, the nature and structure of the hydrate active components in crude oils have not yet been determined in detail.

To prevent agglomeration of the hydrate particles, their wettability state must be controlled. Oil-wet particles are hydrophobic and associated with non-aggregating and thus flowable dispersions, while water-wet particles are hydrophilic and associated with aggregating hydrate particles with a higher potential for plugging [7]. The particles' wettability can be affected by the crude oil composition by adsorption or inclusion of components naturally occurring in crude oil to the hydrate surface.

Petroleum acids have shown surface activity towards hydrate surfaces. It has therefore been suggested that naturally occurring hydrate inhibiting components are present in the acid fractions of crude oils [8–11]. Furthermore, the acid fractions have been shown to contain large amounts of naphthenic acid compounds [12]. They consist of a complex mixture of alkyl-substituted acyclic and cycloaliphatic carboxylic acids with the general formula $C_nH_{2n+z}O_2$ where n corresponds to the number of carbon atoms and z specifies the hydrogen deficiency from ring formation [13]. Comparatively, asphaltene fractions are known to possess self-agglomerating properties and can stabilise oil-wetted systems [14]. It has been shown that the asphaltene fractions able to stabilise oil-wetted systems often are more polar, with higher oxygen content, higher acidity and lower DBEs [15]. Other studies have suggested that the possible hydrate activity of asphaltenes is related to their sulfoxide content [16]. Accordingly, some asphaltenes can alter the plugging potential of crude oils [17, 18].

The complex mixture and relatively high masses of the components in crude oils make it difficult to identify single components with most mass spectrometers. However, with the high mass accuracy of Fourier Transform Ion Cyclotron Resonance Mass Spectrometry (FT-ICR MS) more detailed analysis of crude oils with the ability to identify a large number of polar groups, including compounds present in low concentrations, is possible [19]. FT-ICR MS has previously been used extensively for crude oil characterisation [20–27]. Qian et al. [28, 29] showed that electrospray ionisation (ESI) FT-ICR MS was able to identify more than 3000 chemical formulas of nitrogen containing aromatic compounds in positive mode. Additionally, studies have shown that asphaltenes can be characterised by positive mode ESI FT-ICR MS [30–32].

With the highly detailed spectra derived from FT-ICR MS, there is a need for powerful data analysis methods to efficiently extract valuable information and disregard unimportant information. The present work describes the use of machine learning-based variable selection for the identification of naturally occurring hydrate inhibitors from ESI positive FT-ICR MS spectra and relating the selected variables to the wettability state of the respective crude oils.

Materials and methods

Fluid system

The crude oils used originated from the Norwegian continental shelf and were used as received unless specifically mentioned. The water phase consisted of 3.5 wt% NaCl in tap water, thus

containing only monovalent ions in the water-phase, which simplifies the water chemistry avioding possible unwanted reactions by bivalent ions such as Ca^{2+} [33]. The gas phase was a mixture of 86/8/6 mol% of methane, ethane and propane respectively (Linde Gas AS) with a mixture tolerance of 10% and an analysis uncertainty of 2%.

Experimental set-up

The autoclave used in the experiments was a 200 mL high-pressure sapphire cell (Top Industrie) owned by SINTEF AS, placed inside a temperature controlled chamber. The temperature was measured using a PT-100 element positioned at the bottom of the cell. A connected stirrer mixed the phases to create a fully dispersed system. The cell was fitted with a Hy-Lok FT Micron Tee Filter with a 150 μm sintered stainless steel filter element. A probe inserted from the top was used to measure the conductivity in the liquid phase. Gas filling was controlled using an IN-FLOW HI-Press MFC mass flow controller (Bronkhorst).

Successive accumulation of hydrate active components

A successive accumulation procedure (spiking) was performed with the aim of accumulating possible hydrate active components. A schematic illustration of the procedure is shown in Fig 1. The method developed by Fossen et al. [34] was based on Borgund et al. [6] which presented the same procedure, but with a non-pressurised system using tetrahydrofuran as hydrate former. The procedure started with a fresh oil sample which was added to the cell with the water phase at a given water cut and pressurised with a hydrocarbon gas phase. The pressure used for the current study was 65 bar. The water cut is the ratio of water compared to the total volume of the system. The temperature was lowered to 2˚C while stirring the liquid to ensure a homogeneous dispersion. By cooling the system at high pressure, the hydrate formation region will eventually be reached, and given enough sub-cooling, the system will form hydrates. For the current tests, the system was kept at low temperature over night, to ensure hydrate formation. When hydrates had formed, and the reaction allowed to reach equilibrium,

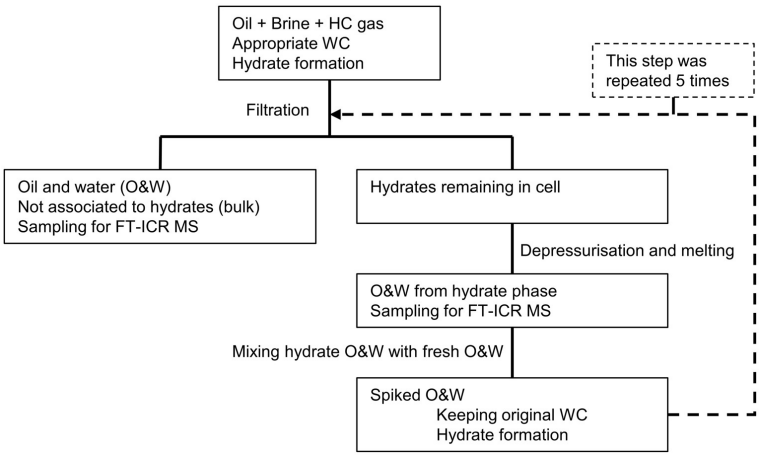


Fig 1. Schematic illustration of the successive accumulation experiment for spiking of the hydrate phase.

https://doi.org/10.1371/journal.pone.0273084.g001

the phase not associated with hydrates, called the bulk phase, was drained through the bottom of the cell. The driving force for draining was the pressure difference of the cell and the ambient pressure conditions outside the cell. The hydrate phase was retained by the filter, so only water and oil not associated to hydrates were drained. Once the bulk phase was drained, the cell was depressurised and the temperature was increased, leading to dissociation of the hydrate phase which was drained and collected, resulting in an oil and a water phase that had been associated to the gas hydrates. The now liquid hydrate phase was then mixed with fresh oil and water at a ratio ensuring the same water cut as the previous run, before repeating the hydrate formation and draining procedure. Small samples were taken from both the bulk phase and the hydrate phase at each step for analysis by FT-ICR MS.

Wetting index experiments

A wetting index (WI) procedure for determining the emulsion inversion point was developed by Høiland et al. [35] and advanced by Fossen et al. [34]. In short, the WI is obtained from determination of the inversion point of the emulsions with and without hydrates present. When the emulsion inversion point shifts towards higher water cuts after hydrate formation, the hydrates are oil-wetted, and when the shift is towards lower water cuts, the hydrates are water-wetted. This is in accordance with the principles of Bancroft [36]. The WI is defined as the normalised difference in inversion point with, and without hydrates present, represented by a number between -1 and +1. Positive values indicate oil-wetted systems with little or no potential of plugging, while negative values indicate water-wetted systems with a high potential of plugging. The absolute value of the WI number is expected to be of importance, and a higher positive or negative value indicates higher degrees of oil-wetted or water-wetted hydrate particles.

FT-ICR MS analysis

For the FT-ICR MS analysis, the samples were prepared by dissolving 20 μL sample in 980 μL dichloromethane. 20 μL of the diluted sample was then added to 980 μL of a 1:1 mixture of toluene and methanol. 100 μL were injected onto the FT-ICR MS using a Aglient 1290 Infinity HPLC system as the introduction device. The 100 μL were injected over a period of 10 minutes with a flow of 10 μL per minute. The mass spectra were acquired using a Bruker Solarix XR FT-ICR MS (Bruker Daltonik GmbH, Germany) equipped with a 12 Tesla magnet (Bruker Biospin, France) owned by SINTEF and located in Trondheim (resolution: 450 000 at m/z 400). The FT-ICR was equipped with an electrospray ion source (ESI) operating in positive mode with the mass range set to 150–3000 m/z.

3 oil samples (anonymised to A, J2 and I) underwent the successive accumulation procedure resulting in 41 samples of different spiking levels. 6 spiking levels for oil A and 5 spiking levels for oil J2 and I. The samples were analysed by FT-ICR MS in three parallels each. For each sample, 220 spectra were collected.

Data treatment

A bucket table was created of the data using Bruker Compass ProfileAnalysis 2.1. The settings in ProfileAnalysis was as follows: the average peak list was calculated, normalisation was set to the sum of bucket values in analysis, no baseline or smoothing, S/N threshold of 4, relative intensity threshold of 0.01 and absolute intensity threshold of 100. The resulting data set consisted of 123 samples and 27600 variables between m/z 148.44 and 1001.66.

Principal Component Analysis (PCA)

PCA [37] is an unsupervised method for data reduction where a large data set $\boldsymbol{X}$ is decomposed into a subspace containing linear combinations of the original variables as shown in Eq 1.

$$\boldsymbol{X} = \boldsymbol{X}_{In} wgt_X \tag{1}$$

Where $\boldsymbol{X}_{\mathrm{In}}$ has the shape (N, K) and is the mass spectra for N oil samples with K $\boldsymbol{X}$ -variables and $\boldsymbol{X}$ has the shape (N, K) which are the balanced spectra for N oil samples with K variables. Eq 2 shows the PCA model for A Principal Components (PCs).

$$\boldsymbol{X} = \bar{\boldsymbol{x}} + \boldsymbol{T}_A \boldsymbol{P}_A^T + \boldsymbol{E}_A \tag{2}$$

Where $\boldsymbol{P}_{\mathrm{A}}$ are the loadings and orthonormal eigenvectors of $(\boldsymbol{X} - \bar{\boldsymbol{x}})^T (\boldsymbol{X} - \bar{\boldsymbol{x}})$ with shape (K, A), minimising the covariance between the $\boldsymbol{X}$ -variables after A PCs. The scores $(\boldsymbol{T}_{\mathrm{A}})$ are orthogonal as shown by Eq 3 andwil have shape (N, A).

$$\boldsymbol{T}_A = (\boldsymbol{X} - \boldsymbol{x})\boldsymbol{P}_A \tag{3}$$

The error term in 2 is $\boldsymbol{E}_{\mathrm{A}}$ which is calculated by Eq 4.

$$\boldsymbol{E}_A = \boldsymbol{X} - \bar{\boldsymbol{x}} - \boldsymbol{T}_A \boldsymbol{P}_A^T \tag{4}$$

Variable selection methods

Variable selection is the process of selecting a subset of relevant variables to use when constructing a model. When a data set contains a large number of variables, it is often assumed that the data contains irrelevant or redundant variables that can be removed without loss of information. Removing them can improve the prediction ability of the model and reduce the computational cost during modelling. Variable selection can also be used to identify the features with the highest correlation to the response, i.e. the most important variables.

In this paper, variable selection methods such as Partial Least Squares Discriminant Analysis, Decision Trees, Random Forest, Boosting, and LASSO (Least Absolute Shrinkage and Selection Operator) regularisation were compared with the aim of predicting whether the samples were related to the hydrate or the bulk phase. An attempt was made to identify components in the data related to hydrate formation with the hypothesis that there could be systematic differences in the spectra which the proposed methods could be able to distinguish.

Partial Least Squares Discriminant Analysis (PLS-DA). PLS-DA [38] decomposes large data sets into a subspace of latent variables consisting of scores and loadings which represent the main features of covariance in the data. The latent variables are found by a maximisation of the covariance between the features, $\boldsymbol{X}$ and the response, $\boldsymbol{Y}$. $\boldsymbol{X}$ has the same input model as for PCA, shown in Eq 1. As PLS-DA also takes the response into account as opposed to PCA, the input model for $\boldsymbol{Y}$ is shown in Eq 5.

$$\boldsymbol{Y} = \boldsymbol{Y}_{In} wgt_Y \tag{5}$$

Where $\boldsymbol{Y}_{\mathrm{In}}$ has the shape (N, J) and is the input categorical variables (0 or 1) for N oil samples with J categorical variables, wgt_{X} are the statistical weights for balancing the sum of squares for the $\boldsymbol{Y}$ variables and $\boldsymbol{Y}$ is the balanced data with shape (N, J) for N oil samples with J $\boldsymbol{Y}$ -variables. The decomposition of $\boldsymbol{X}$ is taken into account, resulting in $\boldsymbol{Y}$ relevant latent variables. This is shown by Eqs 6 and 7.

$$\boldsymbol{X} = \bar{\boldsymbol{x}} + \boldsymbol{T}_A \boldsymbol{P}_A^T + \boldsymbol{E}_A \tag{6}$$

$$\boldsymbol{Y} = \bar{\boldsymbol{y}} + \boldsymbol{U}_A \boldsymbol{Q}_A^T + \boldsymbol{F}_A \tag{7}$$

Where $\boldsymbol{A}$ denotes the number of PCs used and $\boldsymbol{E}_A$ and $\boldsymbol{F}_A$ are the error terms using $\boldsymbol{A}$ PCs. Plotting of these latent variables provides overview of co-variations both within and between model inputs and outputs. The loading weight matrix ($\boldsymbol{W}_A$) maximises the covariance between $\boldsymbol{X}$ and $\boldsymbol{Y}$ by maximising the covariance between $\boldsymbol{T}$ and $\boldsymbol{U}$ after A components. The scores ($\boldsymbol{T}_A$) are orthogonal as shown by Eq 8.

$$\boldsymbol{T}_A = (\boldsymbol{X} - \bar{\boldsymbol{x}}) \times \boldsymbol{W}_A \tag{8}$$

The loadings for $\boldsymbol{X}$ ($\boldsymbol{P}_A$) are calculated by Eq 9 while the loadings for $\boldsymbol{Y}$ ($\boldsymbol{Q}_A$) are calculated by Eq 10.

$$\boldsymbol{P}_A = (\boldsymbol{T}_A^T \boldsymbol{T}_A^T)^{-1} \boldsymbol{T}_A^T (\boldsymbol{X} - \boldsymbol{x}) \tag{9}$$

$$\boldsymbol{Q}_A = (\boldsymbol{T}_A^T \boldsymbol{T}_A^T)^{-1} \boldsymbol{T}_A^T (\boldsymbol{Y} - \boldsymbol{y}) \tag{10}$$

The error term for $\boldsymbol{X}$ ($\boldsymbol{E}_A$) is calculated as for PCA in Eq 4 and the error term for $\boldsymbol{Y}$ ($\boldsymbol{F}_A$) is calculated by Eq 11.

$$\boldsymbol{F}_A = \boldsymbol{Y} - \boldsymbol{y}\boldsymbol{T}_A \boldsymbol{Q}_A^T \tag{11}$$

The regression coefficients ($\boldsymbol{B}_A$), which are measures of the impact of variations in the various features on the respective response variables, are calculated by Eq 12.

$$\boldsymbol{B}_A = \boldsymbol{W}_A \boldsymbol{Q}_A^T \tag{12}$$

Prediction of $\boldsymbol{Y}$ is then obtained by Eq 13 where $\boldsymbol{b}_0$ is the intercept.

$$\boldsymbol{Y}_{pred} = \boldsymbol{b}_0 + \boldsymbol{X}_{new} \boldsymbol{B}_A + \boldsymbol{F}_A \tag{13}$$

When $\boldsymbol{Y}$ is categorical and the problem is classification, Linear Discriminant Analysis (LDA) is used to predict the class membership of the samples from the PLS-DA component construction by encoding the class membership of the observed variables in $\boldsymbol{X}$ into 0 or 1 [39].

PLS-DA can be used for variable selection by calculation of the Variable Importance in Projection (VIP) for each $\boldsymbol{X}$ variable in the PLS-DA model. The VIP score summarises the influence of the individual $\boldsymbol{X}$ variables on the PLS-DA model and are calculated as the weighted sum of squares for the PLS-DA weights $\boldsymbol{w}_j$ which takes the amount of explained variance in $\boldsymbol{Y}$ into account for each extracted latent variable. VIP therefore gives a measure that can be used to select variables which contribute the most to the explanation of the variance in $\boldsymbol{Y}$. The VIP score for variable K can be calculated from Eq 14.

$$VIP_K = \sqrt{n \frac{\sum_{j=1}^{A} \boldsymbol{B}_j^2 \boldsymbol{t}_j^T \boldsymbol{t}_j \left(\frac{\boldsymbol{w}_{kj}}{\| \boldsymbol{w}_j \|} \right)}{\sum_{j=1}^{A} \boldsymbol{B}_j^2 \boldsymbol{t}_j^T \boldsymbol{t}_j}} \tag{14}$$

Where $\boldsymbol{B}$ is the regression coefficient matrix, $\boldsymbol{w}_j$ is the weight vector, $\boldsymbol{w}_{kj}$ is the kth element of $\boldsymbol{w}_j$ and $\boldsymbol{t}_j$ the score vector from the PLS-DA model with $\boldsymbol{A}$ PCs. A variable with a VIP score greater then 1 are generally considered as important, however this limit is sensitive to non-relevant information in $\boldsymbol{X}$ [40]. In this study, the threshold for selecting variables were determined as the point where the VIP-values flatten out, which was found to be 5.

Decision Trees (DTs). DTs [41, 42] are models where decisions are made by asking a series of questions and generating decision rules based on them. These models consist of a tree root, decision nodes, branches and leaf nodes. They aim to find the smallest set of rules that is consistent with the training data. In general, the rules have the form: *if* $condition_1$ *and* $condition_2$ *and* $condition_3$ *then outcome* and are chosen to divide observations into segments that have the largest difference with respect to the target variable. Therefore, the rule selects both the variable and the best break point (usually selected by significance testing or reduction in variance criteria) for maximal separation of the resulting subgroups.

To avoid overfitting, the trees often have to be pruned by setting a limit for the maximal depth. A leaf can no longer be split when there are too few observations, the maximum depth (the hierarchy of the tree) has been reached, or no significant split can be identified. It is assumed that observations belonging to different classes have different values in at least one of their features. DTs are usually univariate, since they use splits based on a single feature at each internal node.

Random forest (RF). In DTs, the initial selected split effects the optimality of variables considered for subsequent splits, making these methods prone to overfitting and other problems. This can be handled by introducing RF [43–45], an ensemble tree method where each tree is based on a random subset of the data and its features (selected by bootstrapping). The advantage of ensemble trees is that the trees are grown with varying initial splits, and either a voting or the average of the predictions for each new data point across all trees is used. The vote distribution can be used to develop a non-parametric probabilistic predictive model. The change in prediction accuracy when the values of a feature are randomly permuted among the observations gives estimates of the importance of each feature.

Ensemble learning. Ensemble learning combines weak classification models with the main idea that many models in combination perform better than one model alone [46].

Boosting [47] is an ensemble learner where weak learners are trained sequentially, trying to improve upon its predecessor. The classifiers emphasise errors made by the previous classifier, aiming at decreasing the model bias. Boosting learners combine underfitting models with low prediction accuracy with the aim of improving the final prediction. Gradient Boosting [48, 49] is a boosting method where trees are built in every iteration, always minimising the prediction error of the classifier. This combination of several smaller trees forms a stronger learner able to fit larger parts of the data than a simple decision tree can. XGBoost (eXtreme Gradient Boosting) [50] is another boosting method based on gradient boosting, which introduces a penalty function in the boosting algorithm and utilise the computational power more efficiently, reducing the computation times.

Regularisation. Another type of machine learning method valuable for variable selection purposes is the regularisation-based method LASSO (least absolute shrinkage and selection operator) [51].

LASSO. In LASSO the estimates of the regression coefficients are obtained using L1-constrained least squares. This forces the sum of the absolute values of the regression coefficients to be less than a fixed value, which forces certain coefficients (β_j) to be set to zero. The variables which have their regression coefficients set to zero, are omitted from the model. LASSO minimises Eq 15 where the ordinary least squares (OLS) problem is the first term with β_0 as the intercept, and the second term $\lambda \sum_{i=1}^{p} |\beta_j|$ is the regularisation term.

$$RSS_{LASSO} = \sum_{i=1}^{n} (\boldsymbol{y}_i - \beta_0 - \sum_{j=1}^{p} \beta_j \boldsymbol{X}_{ij})^2 + \lambda \sum_{i=1}^{p} |\beta_j| \tag{15}$$

Variable importance score. For each of the variable selection methods a variable importance score can be computed, which is a measure of the variables' relative importance in the prediction model. These scores therefore reflect which variables are the most relevant for the target and which variables are of least importance.

The variable importance score can also be used to improve the prediction model by including only the variables with high scores in the model.

Data analysis

All statistical methods were implemented using Python 3.8 and its machine learning packages. The response consisted of a vector containing information of the samples origin, either extracted from the hydrate phase or from the bulk phase. For the linear models, PCA, PLS-DA and LASSO, the data set was standardised (standard deviation = 1) and mean centered (mean = 0). For PLS-DA, the optimal number of components were selected by splitting the training set into two, 70% for calibration and 30% for validation, and finding the most commonly selected number of components by calculating the accuracy over 25 splits. All methods were validated using 25 different training and test set splits with 70% in the training set and 30% in the test set. Molecular formulas were determined using Bruker Compass DataAnalysis 5.0. From the peak corresponding to the m/z of the variables selected, the formula best fitting to the peak was chosen.

Results

Wetting index experiments

The three oils underwent the WI experiment and their WIs were calculated. The WI for oil A was shown to be 0, indicating that it has no clear plugging tendency. Oil J2 and I were determined to have positive WIs of 0.44 for oil J2 and 0.31 for oil I, indicating that they have low or no tendency of plugging. The resolution of the measurements in terms of water cut were 10 volume%. This gives an accuracy of the measurement of ±0.05 volume% and thus a corresponding uncertainty in the measured WIs. Evaluation of the sensitivity of the water cut resolution on the WIs was not performed in this study.

PCA

Each of the oil samples were analysed by PCA and the resulting scoreplot of the first Principal Component (PC1) and the second Principal Component (PC2) for the data set is shown in Fig 2 where the samples are identified by the oil they originated from. The same scoreplot is shown in Fig 3 with the samples distinguishing the individual spiking levels. In both figures, the samples from the bulk phase are shown in the plot to the left and the samples from the hydrate phase are shown in the plot to the right. Fig 2 shows the differences between the three oils, and PC1 shows the difference between the samples from oil J2 and the samples from the two other oils, A and I. Additionally, PC1 shows a separation between the samples that have undergone the spiking experiment, and the crude oil samples which are clustered around 0. PC2 shows differences in the spiking samples from oil A and I. The spiking samples for oil J2 are clustered at 0 for PC2.

Fig 3 shows the differences between the spiking levels along PC2.

Comparing mass spectra

To investigate differences between the spiking levels of the hydrate phase in each of the oils, the mass spectra from each spiking level were compared to a sample which had not been

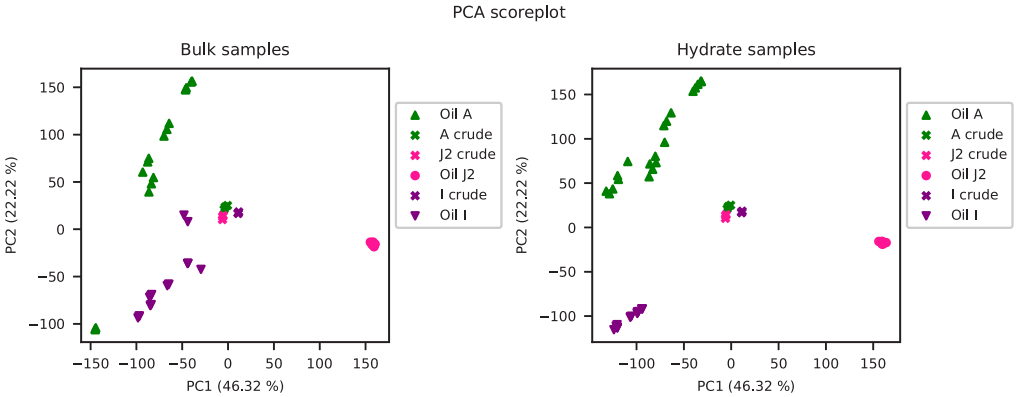


Fig 2. PCA scoreplots with samples from the bulk phase shown in the left plot and samples from the hydrate phase shown in the right plot. Samples are coloured according to which crude oil they originated from and the crude oils have the symbol x.

https://doi.org/10.1371/journal.pone.0273084.g002

spiked. This was only performed for the hydrate phase as it was assumed that the hydrate active components would be present in this phase. An average spectrum was calculated from the tree parallels for each spiking level and for samples removed before the spiking experiment. The sample removed before spiking is referred to as spiking level 0. The mass spectra for spiking level 0 was subtracted from the spectra for the remaining spiking levels for each of the oils. The results for oil A are shown in Fig 4. From Fig 4, four m/z values appeared to have an increasing trend as the spiking levels increased for oil A. They are shown in Table 1 with the molecular formula, double bond equivalent (DBE), the degree of unsaturation of the molecule, the hydrogen-carbon (H/C) ratios, which adduct the molecule has, either sodium (Na) or hydrogen (H^+) and the molecular weight.

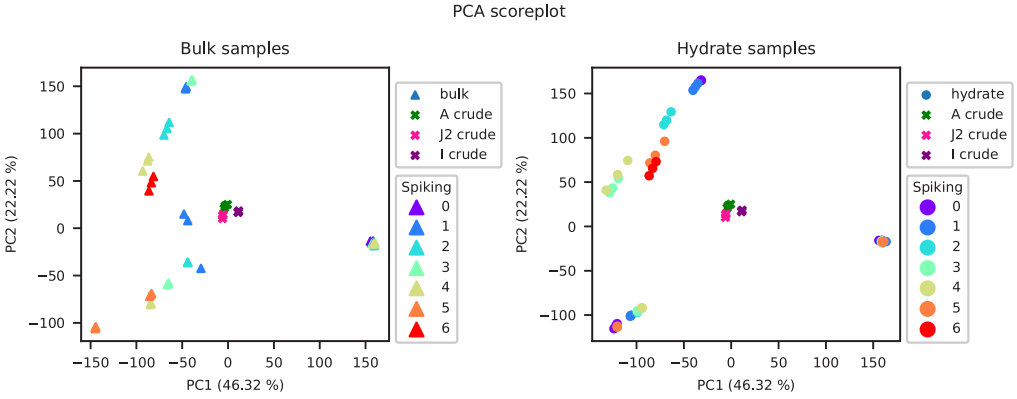


Fig 3. PCA scoreplots with samples from the bulk phase shown in the left plot and samples from the hydrate phase shown in the right plot. Samples are coloured by spiking level and the crude oils have the symbol x.

https://doi.org/10.1371/journal.pone.0273084.g003

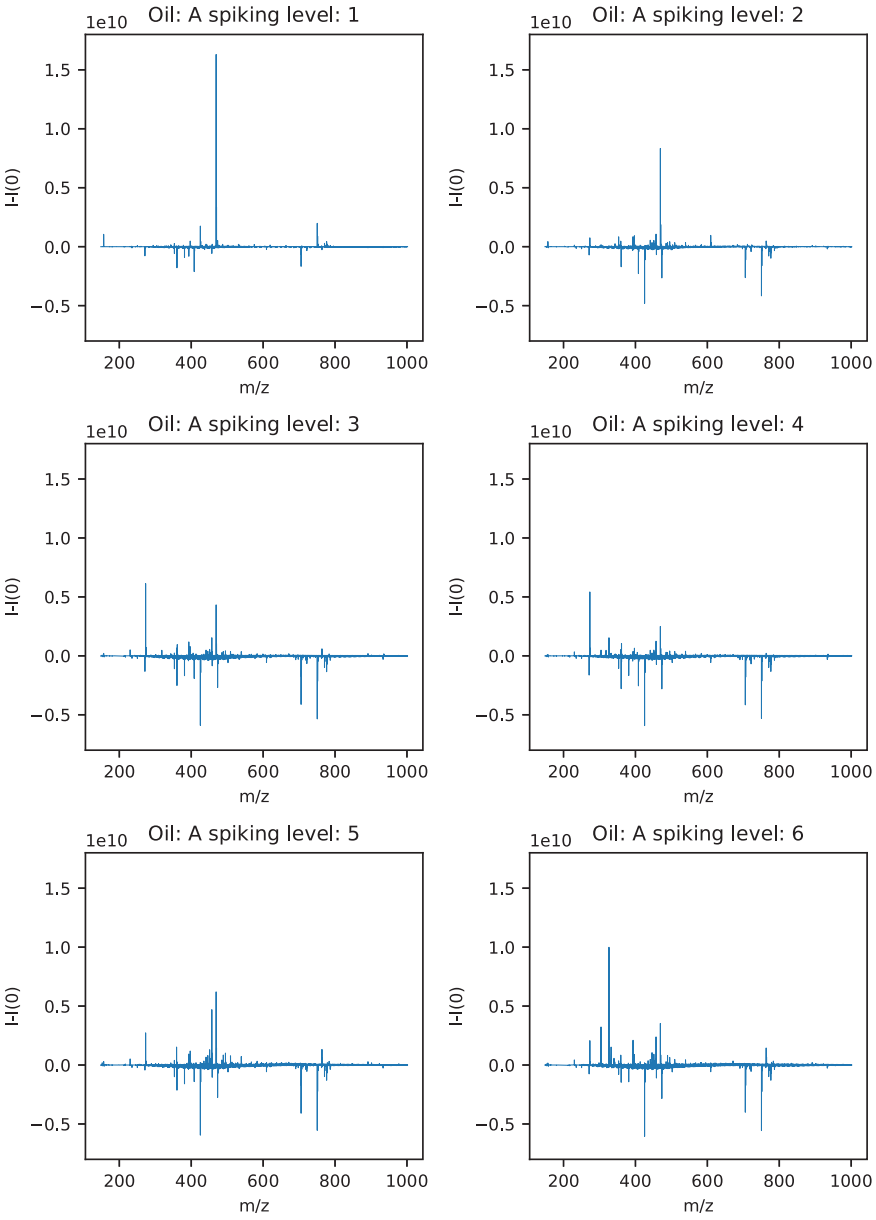


Fig 4. Mass spectra of samples from the hydrate phase for oil A with spiking level 0 subtracted from each of the spiking levels 1–6.

https://doi.org/10.1371/journal.pone.0273084.g004

Table 1. Peaks increasing for oil A.

m/z	Formula	DBE	H/C	Adduct	Molecular formula
273.17	$C_{12}H_{26}O_5$	0	2.17	Na	250.1780
397.18	$C_{18}H_{30}O_8$	4	1.67	Na	374.1941
457.28	$C_{22}H_{42}O_7$	2	1.91	Na	434.2880
469.31	$C_{28}H_{46}O_4$	10	1.64	Na	446.3244

The m/z values with increasing trend as spiking levels increased for oil A, their molecular formula, DBE, H/C-ratio, which adduct the molecule has, Na or H^+, and the molecular weight

https://doi.org/10.1371/journal.pone.0273084.t001

The results with spiking level 0 subtracted from the remaining spiking levels for oil J2 are shown in Fig 5. For oil J2 no distinct m/z values increased with increasing spiking levels.

The results with spiking level 0 subtracted from the remaining spiking levels for oil I are shown in Fig 6. From Fig 6, two m/z values appeared to have an increasing trend as the spiking level increased for oil I. They are shown in Table 2 with molecular formula, DBE, H/C-ratio, which adduct the molecule has and the molecular weight. Additionally, the variable m/z 156.44 increased, but this is an ion with charge three from the m/z 469.32 peak and is therefore not reported.

Variable selection

Several variable selection methods such as Decision Trees, Random Forest, Gradient Boosting, XGBoost, LASSO regularisation and PLS-DA through VIP were tested with the aim of finding components related to hydrate formation. The samples were classified by their origin, whether they were sampled from the bulk phase (0) or from the hydrate phase (1). During the data analysis, it was discovered that the accuracy of the models depended on the composition of the training and test sets. This is an indication that the samples have such large variation between them that some compositions of the training set are not able to predict the test set. This was overcome by running 25 different training and test set combinations. The variable selection methods were tested on all variables to evaluate which method predicted the samples most accurately. The accuracy scores of the test set for each of the six methods are shown in Fig 7, where the accuracy is defined as the fraction of correctly classified samples. The distributions in accuracy for each method is shown by the bars in Fig 7. The best performing model was PLS-DA with an accuracy of 0.62 ± 0.12.

The performance for each of the variable selection methods is shown in Table 3. Each time a model was fitted to a new training and test set, the variables selected by the model were extracted. Variables that were selected by several different training/test sets are more likely to be related to hydrate formation. For the best performing variable selection method, PLS-DA, 26 variables were selected as important by all of the 25 models out of the total 27600 variables in the data set. However, during inspection of the m/z-values, it became apparent that two of the variables referred to the same peak. Additionally, two variables were the corresponding isotope peak, for m/z 393.30 (isotope peak: 394.30) and 469.32 (isotope peak: 470.32). The variables were combined, resulting in 23 unique selected variables which are shown in Table 4 with molecular formula, DBE, H/C-ratio, which adduct the molecule has and molecular weight.

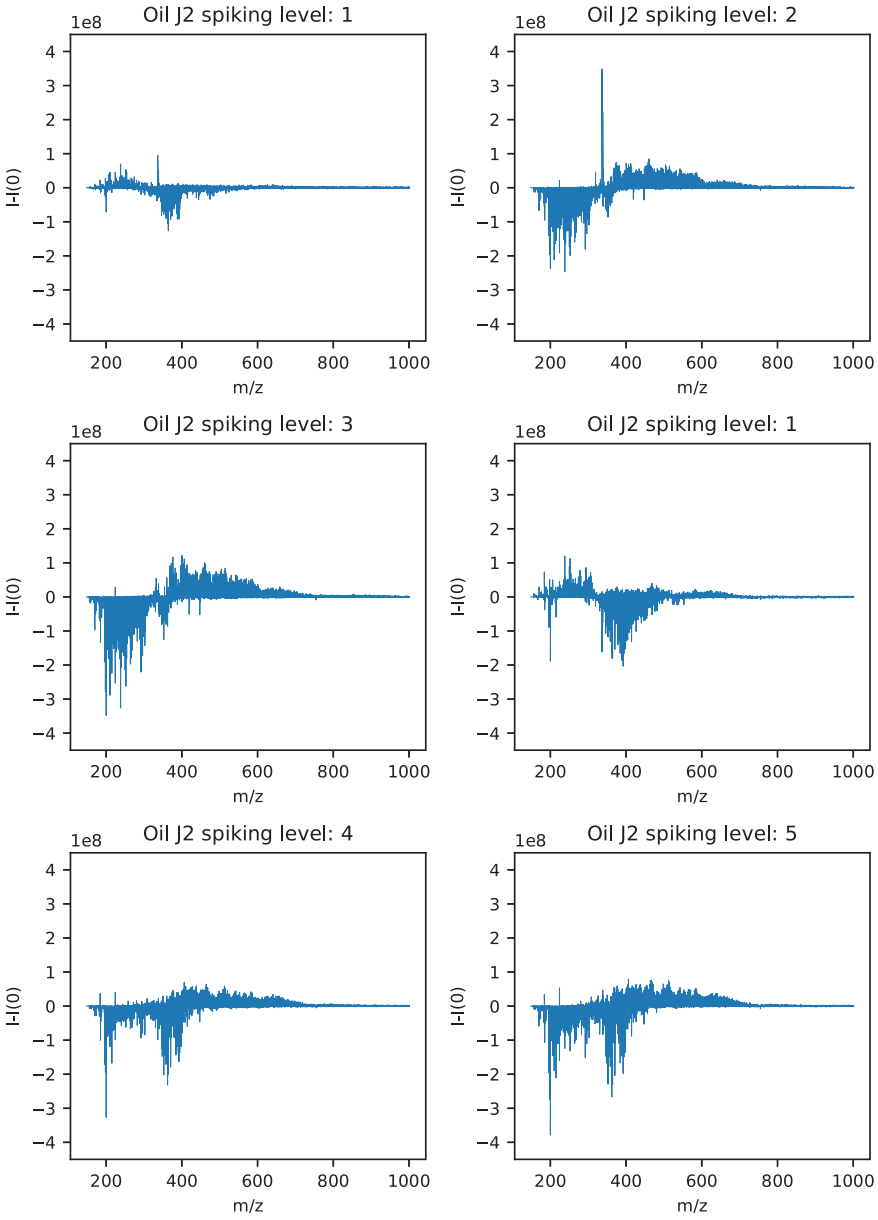


Fig 5. Mass spectra of samples from the hydrate phases for oil J2 with spiking level 0 subtracted from each of the spiking levels 1–5.

https://doi.org/10.1371/journal.pone.0273084.g005

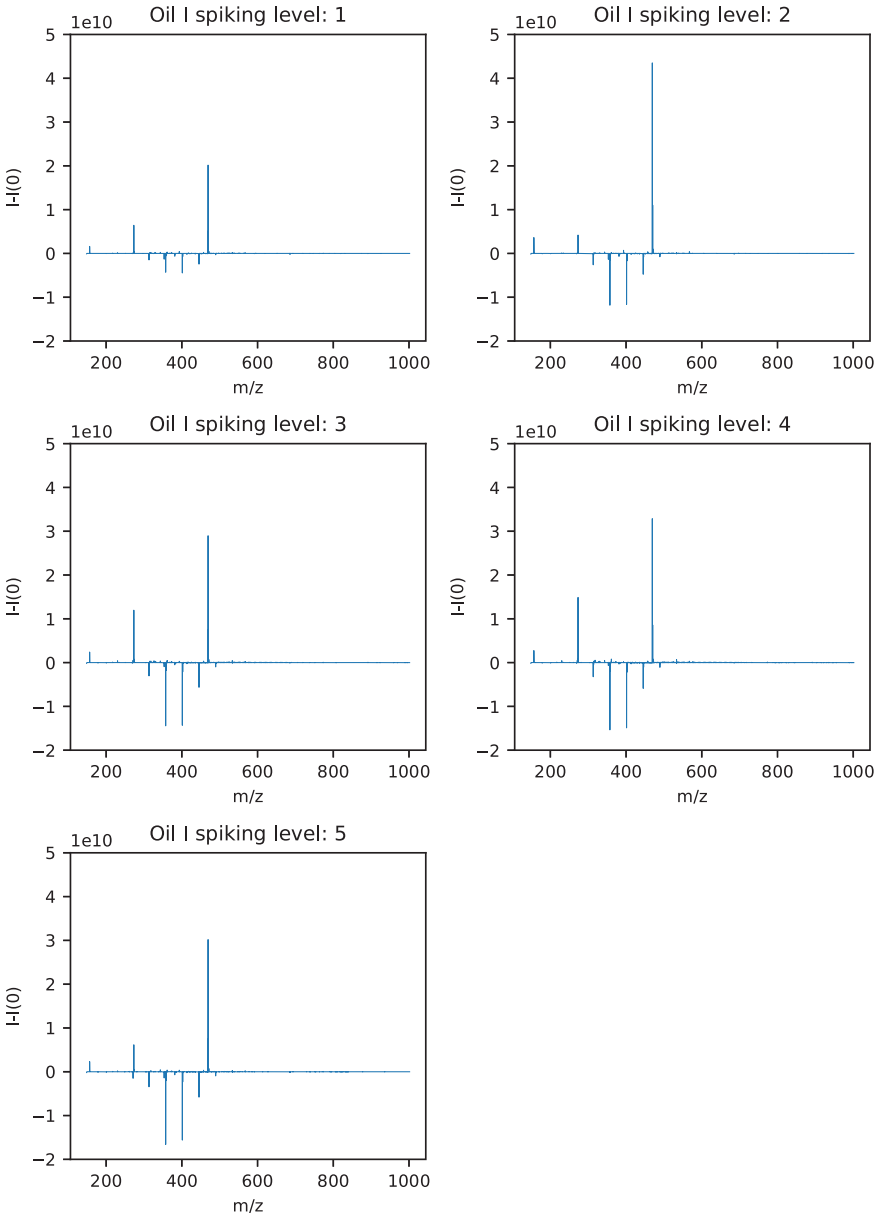


Fig 6. Mass spectra of samples from the hydrate phase for oil I with spiking level 0 subtracted from each of the spiking levels 1–5.

https://doi.org/10.1371/journal.pone.0273084.g006

Table 2. Peaks incresing for oil I.

m/z	Formula	DBE	H/C	Adduct	Molecular formula
273.17	$C_{12}H_{26}O_5$	0	2.17	Na	250.1780
469.32	$C_{28}H_{46}O_4$	10	1.64	Na	446.3396

The m/z values with increasing trend as spiking levels increased for oil I and their molecular formula, DBE, H/C-ratio, which adduct the molecule has, Na or H^+, and the molecular weight.

https://doi.org/10.1371/journal.pone.0273084.t002

Discussion

The results from this work indicated that using machine learning-based variable selection, it is possible to identify components related to hydrate formation. Several methods were tested, and PLS-DA was determined as the best performing method with an accuracy of 0.62 ± 0.12 over 25 different training and test set splits. To determine a representative range, 25, 50, 75 and 100 training and test set splits were run. This sensitivity evaluation indicated that increasing the amount of splits above 25 would not affect the standard deviation significantly. Variable selection models can be prone to overfitting as they consume degrees of freedom, but when using an independent test set, overfitting of the models are counteracted. For each of the 25 times a new model was fitted, the variables selected as important by the model, based on their variable importance score, were extracted. The variables were extracted from the model with the highest accuracy score as that is the model that most accurately predicts the differences between the bulk samples and the hydrate samples, and therefore selects the variables with the highest probability of being related to hydrate formation.

From PLS-DA, 23 variables were selected as important by all of the 25 models and they were identified with their molecular formula, DBE and H/C-ratio. The variables selected ranged from m/z 271.19 to 763.61 and the carbon chains from C_9 to C_{49}. The DBE numbers show

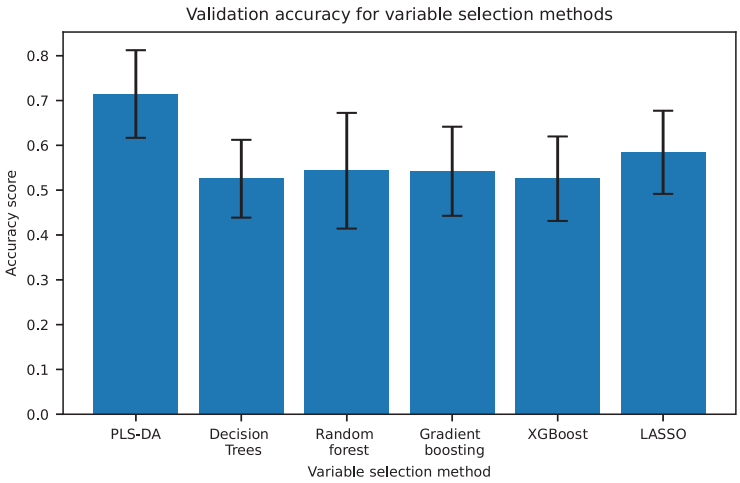


Fig 7. Accuracy scores for the variable selection methods with error bars showing the standard deviation over 25 training/test set splits.

https://doi.org/10.1371/journal.pone.0273084.g007

Table 3. Performance of the variable selection methods.

Method	Accuracy	Standard deviation	No. of variables selected	No. of variables selected in every model
PLS-DA	0.62	0.12	132	26
Decision Trees	0.53	0.13	98	0
Random Forest	0.54	0.09	24929	44
Gradient Boosting	0.54	0.10	12364	0
XGBoost	0.53	0.09	786	0
LASSO	0.58	0.09	334	0

Performance for each variable selection method, their average accuracy, standard deviation, the number of variables selected during the 25 models and the number of variables that were selected in every model.

https://doi.org/10.1371/journal.pone.0273084.t003

6 saturated variables, with DBE of 0, and the highest DBE was 10. The average weight of asphaltenes is ∼750 Da [52], and some of the selected m/z-values were in the range 705–763, indicating that these could be asphaltenes. The asphaltenes with hydrate inhibiting properties often have higher oxygen and sulfoxide content, higher acidity and lower DBEs [15, 53]. All of the possible asphaltenic structures exhibit these properties, and could thereby be related to the positive WI for oil J2 and I. Further studies on the oil samples by extracting and analysing the

Table 4. The variables selected by PLS-DA through VIP.

m/z	Formula	DBE	H/C	Adduct	Molecular weight
271.19	$C_{13}H_{28}O_4$	5	2.15	Na	248.1988
273.17	$C_{12}H_{26}O_5$	0	2.17	Na	250.1780
313.24	$C_{17}H_{32}N_2O_3$	3	1.88	H^+	312.2491
326.38	$C_{22}H_{47}N$	0	2.14	H^+	325.3709
353.27	$C_{19}H_{38}O_4$	1	2.00	Na	330.2770
357.26	$C_{18}H_{38}O_5$	0	2.11	Na	334.2719
359.24	$C_{17}H_{36}O_6$	0	2.12	Na	336.2512
360.32	$C_{22}H_{43}NO$	2	1.95	Na	337.3345
361.22	$C_{16}H_{34}O_7$	0	2.13	Na	338.2305
381.30	$C_{21}H_{42}O_4$	1	2	Na	358.3083
393.30	$C_{22}H_{42}O_4$	2	1.91	Na	370.3083
397.18	$C_{18}H_{30}O_8$	4	1.67	Na	374.1941
401.29	$C_{26}H_{40}OS$	9	1.53	H^+	400.2710
408.31	$C_{22}H_{43}NO_4$	2	1.95	Na	385.3291
425.41	$C_{26}H_{52}N_2O_2$	2	2.00	H^+	424.4029
445.31	$C_{22}H_{46}O_7$	0	2.09	Na	422.4029
451.19	$C_{21}H_{29}O_9$	6	1.52	Na	328.2046
457.28	$C_{22}H_{42}O_7$	2	1.91	Na	434.2880
469.31	$C_{24}H_{46}O_7$	2	1.92	Na	446.3244
469.32	$C_{28}H_{46}O_4$	10	1.64	Na	446.3396
705.58	$C_{42}H_{82}O_4S$	4	1.95	Na	682.5934
750.52	$C_{36}H_{81}N_3O_7SV$	2	2.25	H^+	749.5157
763.61	$C_{45}H_{80}N_4O_3$	8	1.78	Na	740.6180

Table of the 23 m/z values selected in every of the 25 PLS-DA models, their molecular formulas, DBE numbers, hydrate-carbon ratio, which adduct the molecule has, sodium or (Na) or hydrogen (H^+), and the molecular weight.

https://doi.org/10.1371/journal.pone.0273084.t004

asphaltenes may confirm this. One variable, m/z 469.32, follows the general molecular formula for naphthenic acids. Other m/z values appear to have properties that could possibly define them as naphthenic acids, two or more oxygen molecules, DBEs indicating unsaturation and H/C-ratios below 2. As naphthenic acids are suggested to be related to hydrate active components [8], it is therefore likely that they contribute to the positive wetting index for oil J2 and I. Several of the selected variables have molecular formulas corresponding to C_nH_{2n+2} and have a DBE of zero. They have carbon chains between C_{12} and C_{22} and contain either large amounts of oxygen (O_5 or more) or nitrogen. It is therefore probable that these are polyethylene glycol (PEG) molecules stemming from production chemicals used to treat flow assurance issues during extraction and processing of the crude oil [54].

By conducting the successive accumulation procedure for a given oil, generations with possibly increased concentration of hydrate active components could be accumulated. The oils with positive WI, likely to exhibit non-plugging properties, should thereby achieve an increase in the components related to anti-agglomeration, making their identification easier. The PCA scoreplots in Fig 2 show that the crude oils are distinguishable from the spiking samples in both the bulk phase and the hydrate phase. Additionally, the PCA scoreplots in Fig 3 show that the different spiking levels are separated, indicating that there were differences between the samples extracted from each spiking level. The spiking procedure therefore altered the composition of the oils. The variables selected by PLS-DA were also identified as increasing in the hydrate phase spiking fractions for oil A and I supporting the theory of accumulation.

The mass spectra for oil J2 in Fig 5 showed that no distinct m/z values increased as the spiking levels increased. However, for spiking level 2, 3, 4 and 5, the area between m/z 400 and 600 increased, indicating that the variables relevant for hydrate formation could lie in this m/z region. Another possible explanation could be that this oil is saturated with hydrate active components, and the spiking procedure therefore would not change the composition of the oil. This fits well with the WI of +0.44 for oil J2, indicating little or no plugging. It is therefore likely that oil J2 contains more hydrate active components than the oils with lower WI.

The results from the variable selection methods showed that the two linear methods, PLS-DA and LASSO, achieved higher accuracy scores than the tree-based methods. Linear methods are more robust and less susceptible to changes in the data. As there were variations in the accuracy for the models using different training and test set splits, the tree-based methods were likely affected negatively.

The molecular formulas presented in this paper are only suggestions of the most likely molecular formulas from the DataAnalysis software. As the mass of the molecule increases, the amount of possible structures and formulas also increases. Accordingly, the uncertainty of the suggested formulas increases with the mass of the molecule. Nonetheless, the structures give an indication of the nature of the molecules related to hydrate formation, and can be used to indicate whether they are i.e. asphaltenes, acids or alkanes.

For any complex data matrix, there are often assumptions that some of the data is noise and unrelated to the desired prediction. With the methodology presented in this paper, we show that it is possible to extract relevant information from complex data and relate it to the chemical composition of the samples. Thus, the proposed methods can be used in any application where there is a need for extracting, identifying and evaluating important variables.

Further studies

When the m/z values of components related to hydrate formation are identified, the next step will be to determine the molecular structures with higher certainty. This can be done by isolation and fragmentation by FT-ICR MS, making it easier to identify the structures of

complicated molecules. When the compounds are found, they can be tested with the oils to evaluate how their presence changes the characteristics of the oils and the formation of hydrates.

Conclusion

In this study, machine learning-based variable selection was used to identify components related to hydrate formation. A successive accumulation procedure was performed to increase the concentration of the hydrate active components. PCA demonstrated the difference between the spiking levels and the crude oils, establishing that the spiking procedure alters the sample composition significantly, suggesting that hydrate active components have been accumulated. Variable selection methods such as Decision Trees, Random Forest, Gradient Boosting, XGBoost, LASSO regularisation and PLS-DA through VIP were tested to identify the hydrate active components. The best performing prediction model was obtained using PLS-DA which gave an average accuracy of 0.62±0.12 over 25 different training and test set combinations. From the 25 models, 23 variables were selected as important in every model, and their molecular formulas were determined in an attempt to identify molecules related to hydrate formation. Some of the variables were identified as possible asphaltenic structures which could be related to the positive WI for the oils.

Identifying variables in the oil related to hydrate formation takes us one step closer to identifying the naturally occurring hydrate active components.

Supporting information

S1 Fig. Experimental set-up. Picture of the autoclave used for the hydrate formation and spiking experiments. It consists of a sapphire cell between two titanium grad II flanges. Pressure, temperature and conductance is measured inside the sapphire cell. A motor is mounted above the cell driving a stirrer through a magnetic connection.
(PDF)

S2 Fig. Determining the threshold for VIP Plotting of the VIP values for the 25 PLS-DA models with 20 components. The curve flattens around 5 which was selected as the threshold.
(EPS)

S3 Fig. Determining the optimal training/test set split. Accuracy scores for the classification methods showing the mean accuracy over 25, 50, 75 and 100 training/test set splits and standard deviations.
(EPS)

Author Contributions

Conceptualization: Elise Lunde Gjelsvik, Kristin Tøndel.

Data curation: Elise Lunde Gjelsvik.

Formal analysis: Elise Lunde Gjelsvik.

Funding acquisition: Martin Fossen, Kristin Tøndel.

Methodology: Elise Lunde Gjelsvik, Martin Fossen, Anders Brunsvik.

Project administration: Martin Fossen.

Supervision: Martin Fossen, Anders Brunsvik, Kristin Tøndel.

Writing – original draft: Elise Lunde Gjelsvik.

Writing – review & editing: Martin Fossen, Anders Brunsvik, Kristin Tøndel.

References

1. Sloan ED, Koh CA. Clathrate Hydrates of Natural Gases. 3rd ed. No. 119 in Chemical Industries series. Boca Raton, FL: CRC Press, Taylor & Francis Group; 2008.
2. Zhang Q, Kelland MA, Lu H. Non-amide kinetic hydrate inhibitors: A review. Fuel. 2022; 315:123179. https://doi.org/10.1016/j.fuel.2022.123179
3. Fadnes FH. Natural hydrate inhibiting components in crude oils. Fluid Phase Equilibria. 1996; 117(1-2):186–192. https://doi.org/10.1016/0378-3812(95)02952-4
4. Lingelem MN, Majeed AI, Stange E. Industrial Experience in Evaluation of Hydrate Formation, Inhibition, and Dissociation in Pipeline Design and Operation. Annals of the New York Academy of Sciences. 1994; 715(1):75–93. https://doi.org/10.1111/j.1749-6632.1994.tb38825.x
5. Aman ZM, Syddall WGT, Haber A, Qin Y, Graham B, May EF, et al. Characterization of Crude Oils That Naturally Resist Hydrate Plug Formation. Energy & fuels. 2017; 31(6):5806–5816. https://doi.org/10.1021/acs.energyfuels.6b02943
6. Borgund AE, Høiland S, Barth T, Fotland P, Askvik KM. Molecular analysis of petroleum derived compounds that adsorbonto gas hydrate surfaces. Applied Geochemistry. 2009; 24(5):777–786. https://doi.org/10.1016/j.apgeochem.2009.01.004
7. Høiland S, Askvik KM, Fotland P, Alagic E, Barth T, Fadnes F. Wettability of Freon hydrates in crude oil/brine emulsions. Journal of Colloid and Interface Science. 2005; 287(1):217–225. https://doi.org/10.1016/j.jcis.2005.01.080 PMID: 15914170
8. Erstad K, Høiland S, Fotland P, Barth T. Influence of Petroleum Acids on Gas Hydrate Wettability. Energy & Fuels. 2009; 23(4):2213–2219. https://doi.org/10.1021/ef8009603
9. Høiland S, Borglund AE, Barth T, Fotland P, Askvik KM. Wettability of Freon hydrates in crude oil/brine emulsions: the effects of chemical additives. In: 5th International Conference in Gas Hydrate. vol. 4. Trondheim; 2005. p. 1151–1161.
10. Hemmingsen PV, Kim S, Pettersen HE, Rodgers RP, Sjöblom J, Marshall AG. Structural Characterization and Interfacial Behavior of Acidic Compounds Extracted from a North Sea Oil. Energy & Fuels. 2006; 20(5):1980–1987. https://doi.org/10.1021/ef0504321
11. Hemmingsen PV, Li X, Peytavy JL, Sjöblom J. Hydrate Plugging Potential of Original and Modified Crude Oils. Journal of Dispersion Science and Technology. 2007; 28(3):371–382. https://doi.org/10.1080/01932690601107716
12. Borgund AE, Erstad K, Barth T. Fractionation of Crude Oil Acids by HPLC and Characterization of Their Properties and Effects on Gas Hydrate Surfaces. Energy & Fuels. 2007; 21(5):2816–2826. https://doi.org/10.1021/ef070100r
13. Clemente JS, Fedorak PM. A review of the occurrence, analyses, toxicity, and biodegradation of naphthenic acids. Chemosphere. 2005; 60(5):585–600. https://doi.org/10.1016/j.chemosphere.2005.02.065 PMID: 15963797
14. Qiao P, Harbottle D, Tchoukov P, Masliyah J, Sjoblom J, Liu Q, et al. Fractionation of Asphaltenes in Understanding Their Role in Petroleum Emulsion Stability and Fouling. Energy Fuels. 2016; 31(4):3330–3337. https://doi.org/10.1021/acs.energyfuels.6b02401
15. Kilpatrick PK. Water-in-Crude Oil Emulsion Stabilization: Review and Unanswered Questions. Energy Fuels. 2012; 26(7):4017–4026. https://doi.org/10.1021/ef3003262
16. Yang F, Tchoukov P, Dettman H, Teklebrhan RB, Liu L, Dabros T, et al. Asphaltene Subfractions Responsible for Stabilizing Water-in-Crude Oil Emulsions. Part 2: Molecular Representations and Molecular Dynamics Simulations. Energy Fuels. 2015; 29(8):4783–4794. https://doi.org/10.1021/acs.energyfuels.5b00657
17. Salmin DC. The Impact of Synthetic and Natural Surface-Active Components on Hydrate Agglomeration [Doctoral thesis]. Colorado School of Mines. Golden, Colorado; 2019. Available from: https://mountainscholar.org/bitstream/handle/11124/173291/CostaSalmin_mines_0052E_11821.pdf?sequence=1.
18. Adams JJ. Asphaltene Adsorption, a Literature Review. Energy Fuels. 2014; 28(5):2831–2856. https://doi.org/10.1021/ef500282p
19. Emmett MR, White FM, Hendrickson CL, Shi SDH, Marshall AG. Application of micro-electrospray liquid chromatography techniques to FT-ICR MS to enable high-sensitivity biological analysis. Journal of

the American Society for Mass Spectrometry. 1998; 9(4):333–340. https://doi.org/10.1016/S1044-0305(97)00287-0 PMID: 9879363

20. Cho Y, Ahmed A, Islam A, Sunghwan Kim. Developments in FT-ICR MS instrumentation, ionization techniques, and data interpretation methods for petroleomics. Mass Spectrometry Reviews. 2014; 34 (2):248–263. https://doi.org/10.1002/mas.21438 PMID: 24942384

21. Hur M, Yeo I, Kim E, No Mh, Koh J, Cho YJ, et al. Correlation of FT-ICR Mass Spectra with the Chemical and Physical Properties of Associated Crude Oils. Energy & Fuels. 2010; 24:5524–5532. https://doi.org/10.1021/ef1007165

22. Klein GC, Kim S, Rodgers RP, Marshall AG, Yen A. Mass Spectral Analysis of Asphaltenes. II. Detailed Compositional Comparison of Asphaltenes Deposit to Its Crude Oil Counterpart for Two Geographically Different Crude Oils by ESI FT-ICR MS. Energy & Fuels. 2006; 20(5):1973–1979. https://doi.org/10.1021/ef0600208

23. Schaub TM, Jennings DW, Kim S, Rodgers RP, Marshall AG. Heat-Exchanger Deposits in an Inverted Steam-Assisted Gravity Drainage Operation. Part 2. Organic Acid Analysis by Electrospray Ionization Fourier Transform Ion Cyclotron Resonance Mass Spectrometry. Energy & Fuels. 2007; 21(1):185–194. https://doi.org/10.1021/ef0601115

24. Smith DF, Rahimi P, Teclemariam A, Rodger RP, Marshall AG. Characterization of Athabasca Bitumen Heavy Vacuum Gas Oil Distillation Cuts by Negative/Positive Electrospray Ionization and Automated Liquid Injection Field Desorption Ionization Fourier Transform Ion Cyclotron Resonance Mass Spectrometry. Energy & Fuels. 2008; 22(5):3118–3125. https://doi.org/10.1021/ef8000357

25. Headley JV, Peru KM, Barrow MP, Derrick PJ. Characterization of Naphthenic Acids from Athabasca Oil Sands Using Electrospray Ionization: The Significant Influence of Solvents. Analytical Chemistry. 2007; 79(16):6222–6229. https://doi.org/10.1021/ac070905w PMID: 17602673

26. Barrow MP, Headley JV, Peru KM, Derrick PJ. Data Visualization for the Characterization of Naphthenic Acids within Petroleum Samples. Energy & Fuels. 2009; 23(5):2592–2599. https://doi.org/10.1021/ef800985z

27. Fernandez-Lima FA, Becker C, McKenna AM, Rodgers RP, Marshall AG, Russell DH. Petroleum Crude Oil Characterization by IMS-MS and FTICR MS. Analytical Chemistry. 2009; 81(24):9941–9947. https://doi.org/10.1021/ac901594f PMID: 19904990

28. Qian K, Rodgers RP, Hendrickson CL, Emmett MR, Marshall AG. Reading Chemical Fine Print: Resolution and Identification of 3000 Nitrogen-Containing Aromatic Compounds from a Single Electrospray Ionization Fourier Transform Ion Cyclotron Resonance Mass Spectrum of Heavy Petroleum Crude Oil. Energy & Fuels. 2001; 15(2):492–498. https://doi.org/10.1021/ef000255y

29. Qian K, Robbins WK, Hughey CA, Cooper HJ, Rodgers RP, Marshall AG. Resolution and Identification of Elemental Compositions for More than 3000 Crude Acids in Heavy Petroleum by Negative-Ion Microelectrospray High-Field Fourier Transform Ion Cyclotron Resonance Mass Spectrometry. Energy and Fuels. 2001; 15:1505–1511. https://doi.org/10.1021/ef010111z

30. Klein GC, Angström A, Rodgers RP, Marshall AG. Use of Saturates/Aromatics/Resins/Asphaltenes (SARA) Fractionation To Determine Matrix Effects in Crude Oil Analysis by Electrospray Ionization Fourier Transform Ion Cyclotron Resonance Mass Spectrometry. Energy & Fuels. 2006; 20(2):668–672. https://doi.org/10.1021/ef050353p

31. Zhang L, Zhang Y, Zhao S, Xu C, Chung KH, Shi Q. Characterization of heavy petroleum fraction by positive-ion electrospray ionization FT-ICR mass spectrometry and collision induced dissociation: Bond dissociation behavior and aromatic ring architecture of basic nitrogen compounds. Science China Chemistry. 2013; 56:874–882. https://doi.org/10.1007/s11426-013-4899-4

32. Pinto FE, Barros EV, Tose LV, Souza LM, Terra LA, Poppi RJ, et al. Fractionation of asphaltenes in n-hexane and on adsorption onto CaCO3 and characterization by ESI(+)FT-ICR MS: Part I. Fuel. 2017; 210:790–802. https://doi.org/10.1016/j.fuel.2017.09.028

33. Magnusson H, Hanneseth AD, Sjöblom J. Characterization of C80 Naphthenic Acid and Its Calcium Naphthenate. Journal of Dispersion Science and Technology. 2008; 29(3):464–473. https://doi.org/10.1080/01932690701718966

34. Fossen M, Wolden M, Brunsvik A. Successive accumulation of naturally occurring hydrate active components and the effect on the wetting properties. In: 32nd Oil Field Chemistry Symposium 2021. TEKNA; 2021. p. 16.

35. Høiland S, Glénat P, Askvik KM. Wetting Index: A Quantitative Measure Of Indigenous Hydrate Plugging Tendency; Flow Test Validations. In: ICGH7. Edinburgh, UK; 2011. Available from: https://www.semanticscholar.org/paper/THE-WETTING-INDEX-%3A-A-QUANTITATIVE-MEASURE-OF-%3B-H%C3%B8iland-Gl%C3%A9nat/34c1749c108eda3afcda326a73c0e99b24c37d4b.

36. Bancroft WD. The Theory of Emulsification, V. Journal of Physical Chemistry. 1913; 17(6):501–519.

37. Pearson K. On lines and planes of closest fit to systems of points in space. Philosophical Magazine. 1901; 2:559–572.

38. Wold S, Martens H, Wold H. The multivariate calibration problem in chemistry solved by the PLS method. In: Matrix Pencils. No. 973 in Lecture Notes in Mathematics. Berlin, Heidelberg: Springer; 1983. p. 286–293. Available from: https://link.springer.com/chapter/10.1007/BFb0062108.

39. Barker M, Rayens W. Partial least squares for discrimination. Journal of Chemometrics. 2003; 17(3): 166–173. https://doi.org/10.1002/cem.785

40. Tran TN, Afanador NL, Buydens LMC, Blanchet L. Interpretation of variable importance in Partial Least Squares with Significance Multivariate Correlation (sMC). Chemometrics and Intelligent Laboratory Systems. 2014; 138:153–160. https://doi.org/10.1016/j.chemolab.2014.08.005

41. Quinlan JR. Simplifying decision trees. International Journal of Man-Machine Studies. 1987; 27(3):221–234. https://doi.org/10.1016/S0020-7373(87)80053-6

42. Utgoff PE. Incremental Induction of Decision Trees. Machine Learning. 1989; 4:161–186. https://doi.org/10.1023/A:1022699900025

43. Breiman L. Bagging predictors. Machine Learning. 1996; 24:123–140.

44. Breiman L. Random Forests. Machine Learning. 2001; 45:5–32.

45. Ho TK. The random subspace method for constructing decision forests. IEEE Transactions on Pattern Analysis and Machine Intelligence. 1998; 20(8):832–844. https://doi.org/10.1109/34.709601

46. Breiman L. Arcing classifier (with discussion and a rejoinder by the author). Annals of Statistics. 1998; 26(3):801–849. https://doi.org/10.1214/aos/1024691079

47. Schapire RE. The strength of weak learnability. Machine Learning. 1990; 5:197–227. https://doi.org/10.1023/A:1022648800760

48. Friedman JH. Greedy Function Approximation: A Gradient Boosting Machine. The Annals of Statistics. 2001; 29(5):1189–1232. https://doi.org/10.1214/aos/1013203451

49. Friedman JH. Stochastic gradient boosting. Computational Statistics & Data Analysis. 2002; 38(4):367–378. https://doi.org/10.1016/S0167-9473(01)00065-2

50. Chen T, Guestrin C. XGBoost: A Scalable Tree Boosting System. In: KDD'16: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. San Francisco, USA; 2016. p. 785–794. Available from: https://dl.acm.org/doi/10.1145/2939672.2939785.

51. Tibshirani R. Regression Shrinkage and Selection Via the Lasso. Journal of the Royal Statistical Society: Series B (Methodological). 1996; 58(1):267–288.

52. Mullins OC, Sheu EY, Hammami A, Marshall AG. Asphaltenes, Heavy Oils, and Petroleomics. 1st ed. Springer-Verlag New York; 2007. Available from: 10.1007/0-387-68903-6.

53. Fossen M, Kallevik H, Knudsen KD, Sjöblom J. Asphaltenes Precipitated by a Two-Step Precipitation Procedure. 2. Physical and Chemical Characteristics. Energy & Fuels. 2011; 25(8):3552–3567. https://doi.org/10.1021/ef200373v

54. Fink JK. Petroleum Engineer's Guide to Oil Field Chemicals and Fluids. 1st ed. Saint Louis: Elsevier Science & Technology; 2011. Available from: https://www.sciencedirect.com/book/9780123838445/petroleum-engineers-guide-to-oil-field-chemicals-and-fluids.

Paper IV

Gjelsvik E. L., Fossen M., Brunsvik A., Liland K. H., Tøndel K., Multiblock analysis combining data from FT-ICR MS, FTIR and NIR spectroscopy improves prediction of the density of crude oils, *Submitted to Applied Spectroscopy*

Multiblock analysis combining data from FT-ICR MS, FTIR and NIR spectroscopy improves prediction of the density of crude oils

Elise Lunde Gjelsvik[1], Martin Fossen[2], Anders Brunsvik[2], Kristian Hovde Liland[1], and Kristin Tøndel[1]

[1]*Faculty of Science and Technology, Norwegian University of Life Sciences, Aas, Norway*
[2]*SINTEF AS, Trondheim, Norway*

Abstract

Crude oils are among the world's most complex organic mixtures containing a large number of unique components which many analytical techniques lack resolving power to characterise. Fourier Transform Ion Cyclotron Resonance Mass Spectrometry (FT-ICR MS) offers a high mass accuracy, making detailed analysis of crude oils possible. Infrared (IR) spectroscopic methods such as Fourier Transform Infrared (FTIR) and Near-infrared (NIR), can also be used for crude oil characterisation. The three methods measure different properties of the samples, and different data sources can often be combined to improve prediction accuracy of models. In this study, Partial Least Squares Regression (PLSR) models for each of the three methods (single-block PLSR) were compared to Multiblock PLSR (MB-PLSR) and Sequential and Orthogonalised PLSR (SO-PLSR), with the aim of predicting the density of crude oils. Variable importance in projection (VIP) was used to identify the important variables for each method, as spectroscopic data often contains irrelevant variation. The variables were interpreted to evaluate their underlying chemistry and to check whether consistency could be found between the variables selected from the spectroscopic techniques for the single-block and multiblock methods. Combining the different blocks of data increased the prediction abilities of the models both before and after variable selection, and SO-PLSR using a reduced data set resulted in the best prediction model.

Keywords

Multiblock; Partial Least Squares Regression; Fourier Transform Ion Cyclotron Resonance Mass Spectrometry; Infrared Spectroscopy; Crude oil; Petroleomics

1 Introduction

Crude oils are among the world's most complex mixtures, which makes detailed characterisation difficult [1, 2]. They are made up of saturated and unsaturated hydrocarbons with small amounts of heteroatoms (nitrogen, oxygen and sulphur) and metallic constituents. Crude oil properties such as density, viscosity or boiling point are often used as measures for rapid assessments of crude oils, as they are determined by the chemical composition of the oil, and thereby related to the crude oil components [3]. Chemometric methods have been used for identification of crude oil properties for decades [4, 5, 6], and the field of Petroleomics [2] arose from identification of crude oil components using high-resolution mass spectrometry.

Fourier Transform Ion Cyclotron Resonance Mass Spectrometry (FT-ICR MS) offers a high mass accuracy and has previously been used extensively for crude oil characterisation, as the connected resolution power makes it capable of more detailed analysis compared to traditional mass spectrometers [2, 7, 8, 9, 10, 11, 12, 13, 14]. The mass accuracy of FT-ICR MS is sub-ppm, and the mass spectral resolution can be above 10 million (at m/z = 400), which allows for better separation and enables identification of a large number of different chemical groups [15, 16]. Accordingly, FT-ICR MS is regarded as the most efficient technique for crude oil analysis. Atmospheric pressure photoionisation (APPI) is the ionisation source shown to yield the most detailed mass spectra compared to other commonly used ionisation sources for FT-ICR MS, as APPI is efficient at characterising both non-polar molecules, which constitutes approximately 90% of crude oil composition, and polar molecules [17].

However, infrared spectroscopy (IR) has also been used extensively for prediction of crude oil properties with good results [18, 5, 19]. Near infrared (NIR) and mid infrared (FTIR) spectroscopy use the infrared region of the electromagnetic spectrum. IR methods measure the amount of light a sample absorbs at each of the selected wavelengths based on molecular vibrations [20]. For crude oils, the FTIR spectra are usually dominated by the absorption bands from C-H bonds and groups containing aromatics, sulphur, oxygen and nitrogen [21]. In NIR spectra, functional groups such as methylenic, oleifinic and aromatic C-H bonds are usually more prominent, and the bonds involved are C-H, O-H and N-H. However, NIR is limited when it comes to identification of chemical groups and is more suited for quantitative analysis [22].

Several studies have compared various ionisation techniques (ESI/APPI etc.) or ionisation modes (positive/negative) for FT-ICR MS, demonstrating that the most comprehensive characterisation of crude oils is achieved when using more than one ionisation technique/mode [23, 24, 25]. This raises the question of what can be gained by using more than one spectroscopic technique. The various spectroscopic methods measure different parts of a sample and can reveal different properties of the oils. It has been shown that combining data sources measuring complementary information can improve the predictive accuracy or interpretability of a model [26], and multiblock analysis is one way to utilise several different data sources to gain a deeper understanding of the samples. For instance, Dearing et al. [27] used data fusion to combine three spectroscopic techniques for characterisation of crude oils, and illustrated the benefits of combining data from different sources, but did not

consider mass spectrometric techniques.

Another important aspect of spectroscopic techniques is that the data generally contains variation not related to the response, which can diminish the predictive abilities of a model. It is therefore of interest to remove non-relevant variables and keep only those that have an effect on the response. These can be viewed as the important variables, and interpreting them and identifying their corresponding chemistry, is of great importance when trying to understand the mechanisms of a system. Finding the important variables is referred to as variable selection, and has been frequently used for spectroscopic data [28, 29].

In this study, two multiblock methods based on two different data fusion techniques, Multiblock Partial Least Squares Regression (MB-PLSR) and Sequential and Orthogonalised Partial Least Squares Regression (SO-PLSR) were compared to single-block Partial Least Squares (PLSR), with the aim of predicting the density of crude oils from FT-ICR MS and IR data. An additional goal was to evaluate the gain of using multi-block methods to add information from IR to FT-ICR MS data, and whether this could improve the characterisation of crude oil properties. Variable Importance in Projection (VIP) was used to remove irrelevant variables and identify important peaks which were interpreted to evaluate whether some of the methods identified the same chemical structures.

2 Materials and methods

2.1 Density measurements

Density is closely related to the chemical structure of a sample and is a measure often used by the petroleum industry to evaluate the quality of the oil through the American Petroleum Institute (API) gravity. API gravity is related to density via specific gravity and categorises oils into density levels of light, medium and heavy. Density is fairly easy to measure, it is closely related to the chemical composition of the sample, and has been shown to correlate well with both IR [5] and FT-ICR MS spectra [30]. Density was therefore used as the response in this study.

The density of 42 crude oils was measured by a Sigma 703D instrument from Biolin Scientific using the associated density probe. All samples were measured at room temperature within a period of 1 week, under the same ambient conditions. The instrument power was turned on at least 30 minutes before the first sample was measured to reduce variation caused by instrument heating. The measurements were conducted as follows; first, the density ball was placed on the hook before taring and the density should read $1.2e^{-3}$ g/ml, i.e., the density of air. The beaker was then filled with the liquid oil sample and then moved upwards until the stagnant density probe was fully immersed in the sample and the density value on the display was recorded. The density was measured three times per sample and the average of the three was reported including the standard deviation. Between each measurement, the density probe was cleaned with toluene and acetone and dried with an air pistol.

The measured densities were in the range 0.759-0.960 g/mL.

2.2 IR analysis

FTIR and NIR analyses were performed by applying 20 μL of crude oil onto the detection window of a PerkinElmer Frontier FTIR/NIR Spectrometer. In this instrument, NIR mode operates in the range 15 800-2000 cm^{-1} and FTIR mode operates from 8300-400 cm^{-1} [31]. For both FTIR and NIR mode, 16 scans were acquired with resolution 4 cm^{-1}; for FTIR in the range 4000 cm^{-1} to 800 cm^{-1} and for NIR in the range of 12800 cm^{-1} to 4000 cm^{-1}. After preprocessing (as described below), the NIR spectra were cut to contain only the region between 9550-4000 cm^{-1}, in order to remove noise observed between 12800-9550 cm^{-1}.

2.2.1 Preprocessing of IR spectra

Both the FTIR and NIR spectra were preprocessed with three different procedures to identify the most optimal method for data analysis. The different preprocessing procedure were:

- Raw data (no preprocessing)
- Extended Multiplicative Signal Correction (EMSC) [32, 33] with 2^{nd} order polynomial correction, using the mean spectrum as the reference
- EMSC (same settings as above) and Savitzky-Golay (SG) [34] 2^{nd} derivative smoothing (with window width 15 pt and 3^{rd} order polynomial smoothing)

For each procedure, the preprocessed data were either standardised and mean-centred or just mean-centred. The settings for EMSC and SG were chosen based on a preliminary exploration of the data, where the above settings achieved the best degree of smoothing and baseline correction (with respect to predictive accuracy in the following analysis).

2.3 FT-ICR MS analysis

For the FT-ICR MS analysis, the samples were prepared by dissolving 20 μL sample in 980 μL dichloromethane, and 20 μL of this diluted sample were then added to 980 μL of a 1:1 mixture of toluene and methanol. 100 μL were injected onto the FT-ICR MS using an Agilent 1290 Infinity High-performance Liquid Chromatography (HPLC) system as the introduction device over a period of 10 minutes with a flow of 10 μL per minute. The mass spectra were acquired using a Bruker Solarix XR FT-ICR MS (Bruker Daltonic GmbH, Germany) equipped with a 12 Tesla magnet (Bruker Biospin, France) owned by SINTEF AS and located in Trondheim, Norway (resolution 450 000 at m/z 400). The FT-ICR mass spectrometer was equipped with an atmospheric pressure photoionisation ion source (APPI) operating in positive mode, with the mass range set to m/z 150-3000. All samples were measured in three replicates, and for each sample, 220 spectra were collected and the final spectrum reported as the average over the 220.

2.3.1 Data treatment

Before analysis of the FT-ICR MS data, the spectra were combined into a bucket table. Bucketing is the process of removing variations in peak positions due to changes in shifts during analysis [35]. Bruker Compass ProfileAnalysis 2.1 [36] was used for the bucketing with the following settings: normalisation was set to the sum of bucket values in the analysis, no baseline or smoothing was performed, the signal-to-noise (S/N) threshold was 4, the relative intensity threshold was set to 0.01 and the absolute intensity threshold was 100. The average was calculated over the parallels for each sample, resulting in a data set with 42 samples and 23800 variables between m/z 147.51 and m/z 1008.92.

Molecular formulas were determined for each spectrum using Bruker Compass DataAnalysis 5.0 [37]. The settings were as follows: as APPI was set to positive mode, the ions of interest were the molecular ions (M and M+H), the lower limit of atom detection was set to C_6H_6 and the upper limit of atom detection to $O_3S_3N_3$. This means that molecular formulas containing less than 6 carbons or 6 hydrogens would not be suggested, and neither would formulas with more than 3 oxygens, 3 sulphur atoms or 3 nitrogens. These limits were based on previous knowledge of commonly observed and plausible molecular formulas in similar samples. Electron configuration was set to both, with an isotopic fit factor (mSigma) of 100, determining how well the peaks fit the isotopic pattern of the suggested molecular formulas. The resulting molecular formulas were combined into one list for comparison to the variables selected by each method.

2.4 Data analysis

2.4.1 Partial Least Squares Regression (PLSR)

PLSR [38, 39, 40] decomposes large data sets into a subspace of latent variables (scores and loadings) representing the main features of co-variance between the regressors ($\boldsymbol{X}$) and the response ($\boldsymbol{Y}$). PLSR is a commonly used method in chemometrics, and particularly valuable for data with few samples and many variables. The decomposition of $\boldsymbol{X}$ and $\boldsymbol{Y}$ is done jointly and iteratively, taking co-linearities in $\boldsymbol{Y}$ into account. For $\boldsymbol{X}$ the decomposition is shown in Equation 1 and for $\boldsymbol{Y}$ in Equation 2.

$$\boldsymbol{X} = \bar{\boldsymbol{x}} + \boldsymbol{T}_A \boldsymbol{P}_A^T + \boldsymbol{E}_A \tag{1}$$

$$\boldsymbol{Y} = \bar{\boldsymbol{y}} + \boldsymbol{T}_A \boldsymbol{Q}_A^T + \boldsymbol{F}_A \tag{2}$$

where A denotes the number of PLS components used and $\boldsymbol{E}_\mathrm{A}$ and $\boldsymbol{F}_\mathrm{A}$ are the residual terms using A components. The loading weight matrix ($\boldsymbol{W}_\mathrm{A}$) maximises the covariance between $\boldsymbol{X}$ and $\boldsymbol{Y}$ by maximising the covariance between $\boldsymbol{T}$ and $\boldsymbol{U}$ with A components. The scores ($\boldsymbol{T}_\mathrm{A}$) are orthogonal, and are calculated by Equation 3.

$$\boldsymbol{T}_A = \boldsymbol{X}\,\boldsymbol{W}_A(\boldsymbol{P}_A^T\boldsymbol{W}_A)^{-1} \tag{3}$$

The loadings for $\boldsymbol{X}$ ($\boldsymbol{P}_\mathrm{A}$) are calculated by Equation 4, while the loadings for $\boldsymbol{Y}$ ($\boldsymbol{Q}_\mathrm{A}$) are calculated by Equation 5.

$$\boldsymbol{P}_A^T = (\boldsymbol{T}_A^T\boldsymbol{T}_A^T)^{-1}\boldsymbol{T}_A^T(\boldsymbol{X} - \bar{\boldsymbol{x}}) \tag{4}$$

$$\boldsymbol{Q}_A^T = (\boldsymbol{T}_A^T\boldsymbol{T}_A^T)^{-1}\boldsymbol{T}_A^T(\boldsymbol{Y} - \bar{\boldsymbol{y}}) \tag{5}$$

The error terms for $\boldsymbol{X}$ ($\boldsymbol{E}_A$) and $\boldsymbol{Y}$ ($\boldsymbol{F}_A$) are calculated by Equation 6 and Equation 7, respectively.

$$\boldsymbol{E}_A = \boldsymbol{X} - \bar{\boldsymbol{x}} - \boldsymbol{T}_A\boldsymbol{P}_A^T \tag{6}$$

$$\boldsymbol{F}_A = \boldsymbol{Y} - \bar{\boldsymbol{y}} - \boldsymbol{T}_A\boldsymbol{Q}_A^T \tag{7}$$

The regression coefficients ($\boldsymbol{B}_\mathrm{A}$), which are measures of the impact of variations in the various regressors on the respective response variables, are calculated by Equation 8.

$$\boldsymbol{B}_A = \boldsymbol{W}_A(\boldsymbol{P}_A^T\boldsymbol{W}_A)^{-1}\boldsymbol{Q}_A^T \tag{8}$$

Prediction of $\boldsymbol{Y}$ for a new sample ($\boldsymbol{X}_\mathrm{new}$) is then obtained by Equation 9, where $\boldsymbol{b}_0$ is the intercept.

$$\boldsymbol{Y}_{pred} = \boldsymbol{b}_0 + \boldsymbol{X}_{new}\boldsymbol{B}_A \tag{9}$$

2.4.2 Multiblock analysis

Several different techniques exist for combining data for multiblock analysis. The data can for instance be concatenated according to a shared mode, usually with the sample mode acting as the shared mode, or the data can be analysed sequentially, extracting important information from one block before moving to the next block [41]. In the current study, MB-PLSR and SO-PLSR were tested as representatives of the two mentioned strategies. Both methods are based on linear coefficients for the variables with the general formula given by Equation 10.

$$\boldsymbol{Y} = \boldsymbol{X}_1\boldsymbol{B}_1 + \boldsymbol{X}_2\boldsymbol{B}_2 + \cdots + \boldsymbol{X}_N\boldsymbol{B}_N + \boldsymbol{F} \tag{10}$$

where $\boldsymbol{X}$ are the predictor blocks, $\boldsymbol{Y}$ is the response, $\boldsymbol{B}$ are the regression coefficients and $\boldsymbol{F}$ the residuals. The difference between the two methods lies in how the constraints are applied during the decomposition, leading to different orthogonality properties and thereby different independence of the common and distinctive parts. In addition to accomplishing the tasks of the regular single-block techniques, multiblock methods have the advantage of finding common and distinct information present, originating from the different sources of data [42]. Common variation can be

comprehended as the variation associated between data sets, while distinct variation can be regarded as the variation which is unique for each data set.

2.4.3 MB-PLSR

In MB-PLSR, global scores are extracted by maximising the covariance with the response variables, and the extracted global scores are then used in ordinary least squares regression to obtain the predictive models [43, 44]. The data sets are fused by concatenating the individual blocks, after dividing by the square root of the number of variables in each block ($\sqrt{J_m}$). MB-PLSR with super-score deflation of the response starts with an ordinary PLSR on the concatenated blocks, followed by a block-wise extraction of block-weights, block-scores and block-loadings [45]. The prediction is obtained from the PLSR model on the concatenated blocks along with the super-weights, -scores, -loadings and $\boldsymbol{Y}$-scores and -loadings. The block-loading weights ($\boldsymbol{w}_{\mathrm{m}}$) are then obtained by Equation 11 from the original block data ($\boldsymbol{X}_{\mathrm{m}}$).

$$\boldsymbol{w}_m = \frac{\boldsymbol{X}_m^T \boldsymbol{u}}{(\boldsymbol{u}^T \boldsymbol{u})} \tag{11}$$

where $\boldsymbol{u}$ are the $\boldsymbol{y}$ scores. The block-scores ($\boldsymbol{t}_{\mathrm{m}}$) are obtained by Equation 12.

$$\boldsymbol{t}_m = \frac{X_m}{\sqrt{J_m}} \boldsymbol{w}_m^* \tag{12}$$

where J_{m} are the variables for block $\boldsymbol{m}$ and $\boldsymbol{w}^*{}_{\mathrm{m}}$ are the normalised weights ($\boldsymbol{w}_m^* = \boldsymbol{w}_m / \parallel \boldsymbol{w}_m \parallel$). Finally, the block-loadings ($\boldsymbol{p}_{\mathrm{m}}$) are obtained by Equation 13.

$$\boldsymbol{p}_m = \frac{X_m^T}{\sqrt{J_m}} \cdot \frac{\boldsymbol{t}_m}{\boldsymbol{t}_m^T \boldsymbol{t}_m} \tag{13}$$

Figure 1 shows a schematic illustration of how the super-weights, -scores, -loadings and block-weights and -scores are calculated in MB-PLSR.

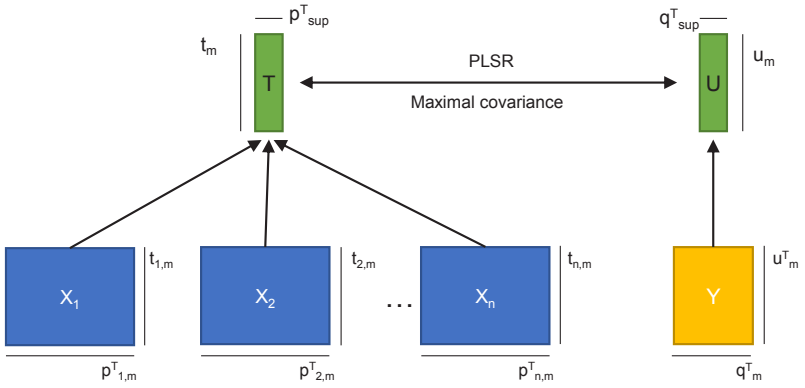


Figure 1: Schematic illustration of MB-PLSR, where the data are concatenated by a shared sample mode, and super-scores and -loadings are calculated from the blocks to achieve maximum covariance.

MB-PLSR applies the same number of components for all the blocks. In cases where the dimensionalities of the blocks are very different, the number of components may not be optimal for all the blocks. However, this allows for simpler models predicting only on one set of components, thus decreasing the susceptibility for overfitting.

2.4.4 SO-PLSR

In SO-PLSR, the blocks of data are incorporated one at a time to evaluate their incremental contribution by letting the method sequentially search for improvements of predictions using additional and orthogonal information provided by the subsequent blocks [46, 47]. This is done by first applying PLSR to the first block and extracting the scores ($\boldsymbol{T}_1$) and loadings ($\boldsymbol{P}_1$), followed by an orthogonalisation of the second block ($\boldsymbol{X}_2$), as shown in Equation 14 for $\boldsymbol{X}_2$ and for $\boldsymbol{Y}$ in Equation 15.

$$X_{2,orth} = (\boldsymbol{I} - \boldsymbol{T}_1(\boldsymbol{T}_1^T\boldsymbol{T}_1)^{-1}\boldsymbol{T}_1^T)\boldsymbol{X}_2 \tag{14}$$

$$\boldsymbol{Y}_{orth} = (\boldsymbol{I} - \boldsymbol{T}_1(\boldsymbol{T}_1^T\boldsymbol{T}_1)^{-1}\boldsymbol{T}_1^T)\,\boldsymbol{Y} \tag{15}$$

In the next step, a new PLSR is fitted to the $\boldsymbol{Y}$-residuals from the first PLSR and the orthogonalised $\boldsymbol{X}_{2,\,\mathrm{orth}}$. This step is repeated for all additional blocks, with all previous blocks included in the orthogonalisation step. Block one and two are concatenated for this purpose, $T_{12} = [T_1\ T_2]$ and T_{12} used for orthogonalisation of block three ($\boldsymbol{X}_3$) by Equation 16.

$$\boldsymbol{X}_{3,orth} = (\boldsymbol{I} - \boldsymbol{T}_{12}(\boldsymbol{T}_{12}^T\boldsymbol{T}_{12})^{-1}\boldsymbol{T}_{12}^T)\boldsymbol{X}_3 \tag{16}$$

$$\boldsymbol{Y}_{orth*} = \boldsymbol{Y}_{orth} - (\boldsymbol{I} - \boldsymbol{T}_2(\boldsymbol{T}_2^T\boldsymbol{T}_2)^{-1}\boldsymbol{T}_2^T)\,\boldsymbol{Y} \tag{17}$$

$\boldsymbol{Y}$ is computed by summing up the predictions of the individual regressions by Equation 18.

$$\boldsymbol{Y}_{pred} = \boldsymbol{T}_1\boldsymbol{Q}_1^T + \boldsymbol{T}_2\boldsymbol{Q}_2^T + \boldsymbol{T}_3\boldsymbol{Q}_3^T + \boldsymbol{F} \tag{18}$$

Figure 2 shows all the steps in the SO-PLSR algorithm where PLSR models are fitted, and how the orthogonalisation is implemented.

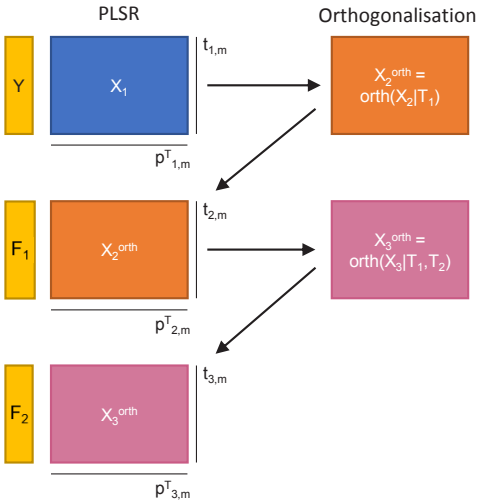


Figure 2: Schematic illustration of the SO-PLSR algorithm, starting with a PLSR model from which the scores are used to orthogonalise the second block, which is then fitted to a new PLSR model. The scores from the first and second PLSR model are used to orthogonalise the third block before it is fitted to a new PLSR model.

SO-PLSR is designed to handle blocks of different complexity and type, including varying numbers of variables and varying dimensionality. Additionally, SO-PLSR is invariant to block-scaling. The order of the blocks is of importance in SO-PLSR, contrary to in MB-PLSR, and changing the order will have an impact on the solution. In this study, the APPI(+)-FT-ICR MS data was used as the first block, since the aim was to uncover whether additional techniques to FT-ICR MS could increase the information that can be extracted from the data and the resulting predictive ability.

2.4.5 Variable selection

Variable Importance in Projection (VIP) is a commonly used method for variable selection in PLSR [48, 49, 50], and is a measure of the influence of the individual $\boldsymbol{X}$ variables on the PLSR model. VIP has also been shown to be efficient for multiblock methods. For example, Biancolillo et al. [51] showed that VIP outperformed other variable selection methods when combined with both SO-PLSR and MB-PLSR.

VIP scores are calculated as the weighted sum of squares for the PLSR weights, which take the amount of explained variance in $\boldsymbol{Y}$ into account for each extracted latent variable. VIP scores therefore select the variables that contribute the most to the explanation of the variance in $\boldsymbol{Y}$. Since the variance explained by each component can be calculated by $\boldsymbol{q}_j^2 \boldsymbol{t}_j^T \boldsymbol{t}_j$, the VIP score for a variable K can be calculated from Equation 19.

$$VIP_K = \sqrt{n \frac{\sum_{j=1}^{A} \boldsymbol{q}_j^2 \boldsymbol{t}_j^T \boldsymbol{t}_j \left(\frac{\boldsymbol{w}_{kj}}{\|\boldsymbol{w}_j\|} \right)^2}{\sum_{j=1}^{A} \boldsymbol{q}_j^2 \boldsymbol{t}_j^T \boldsymbol{t}_j}} \tag{19}$$

where $(\boldsymbol{w}_{kj} / \parallel \boldsymbol{w}_j \parallel)^2$ represents the importance of the k-th variable, wherein $\boldsymbol{w}_\mathrm{j}$ is the weight vector, $\boldsymbol{w}_\mathrm{kj}$ is the k-th element of $\boldsymbol{w}_\mathrm{j}$. Additionally, $\boldsymbol{q}_\mathrm{j}$ are the loadings and $\boldsymbol{t}_\mathrm{j}$ is the score vector from PLSR with A components.

2.4.6 Selecting the VIP threshold

If a variable has a VIP score greater than 1, it is generally considered as important. However, this threshold is sensitive to non-relevant information in $\boldsymbol{X}$, and may have to be altered depending on the data [52]. The absorption in the IR spectra and the intensities of the peaks in FT-ICR MS have different values even after preprocessing and standardisation. Therefore, it is also natural that the methods have different VIP threshold values.

The threshold for each method was selected by fitting a model with the optimal number of components, selected by leave-one-out cross-validation (LOOCV), and extracting the VIP values. A search was performed over the VIP values, including 10 different thresholds between the lowest VIP value, where all variables were included, and the highest VIP value, where only one variable was included. For each of the thresholds, the selected variables were fitted in a reduced model and the optimal threshold was determined from the highest proportion of explained variance R^2. If two thresholds were separated by a small change in R^2, the threshold giving the lowest amount of variables was selected.

2.4.7 Software

Data preprocessing was done using Python 3.8, while all statistical methods were implemented using R 4.2.2 [53] with the *pls* [54], *plsVarSel* [48] and *multiblock* [55] packages. All methods were validated using LOOCV, and the optimal number of components for each method was selected from the highest R^2 after LOOCV.

3 Results and discussion

This study was separated into three parts; in the first part the different preprocessing methods were compared to evaluate their effects on the prediction models, and to select the optimal method for each block. The single-block predictions for FT-ICR MS, FTIR and NIR were then compared to MB-PLSR and SO-PLSR, before variable

selection was performed using VIP based on the single-block PLSR models. The reduced single-block models were then compared to MB-PLSR and SO-PLSR on the variables selected by single-block PLSR. In the last part, VIP was applied to MB-PLSR and SO-PLSR to evaluate how this would change the predictions, and which variables that were selected. All selected variables for the single-block and multiblock methods were interpreted to evaluate whether any commonalities or chemical groups related to the response could be identified.

Figure 3 shows the raw data for the three blocks, FT-ICR MS, FI-IR and NIR, before preprocessing. The scales were different for the three blocks, even after preprocessing, and therefore each block was scaled using the Frobenius norm prior to the analysis.

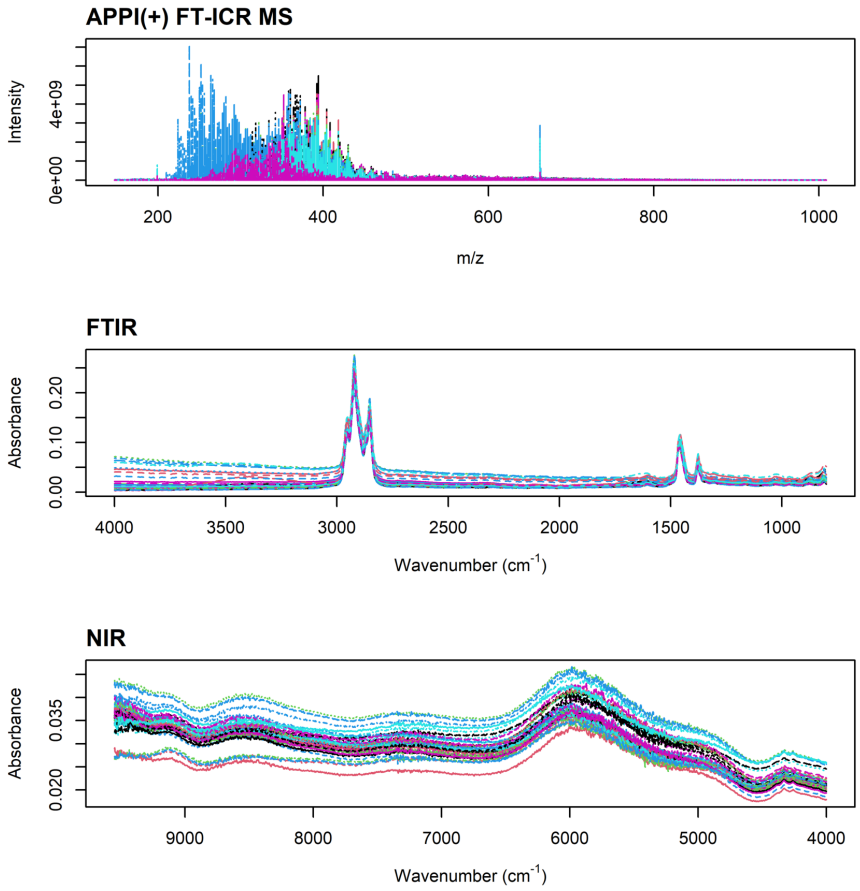


Figure 3: Raw data for the three data sources, FT-ICR MS, FTIR and NIR.

3.1 Comparison of preprocessing methods

It is not given that the same preprocessing technique is optimal for all three types of data (FT-ICR MS, FTIR and NIR). Therefore, to evaluate the optimal preprocessing technique for each block, various combinations were tested. The APPI(+) FT-ICR MS data was tested with and without standardisation, while for FTIR and NIR, the raw data, EMSC preprocessed data and EMSC plus SG preprocessed data were tested with and without standardisation. The preprocessing methods were evaluated based on the highest R^2 value after PLSR validation with LOOCV. MB-PLSR is very sensitive to block-scaling, and to perform optimally, the blocks have to be on similar scales. The search was therefore split into two variants, one where each block was mean-centred and standardised, and a second where the blocks were only mean-centred. Each search was tested on single-block PLSR, MB-PLSR and SO-PLSR, where the optimal preprocessing was selected as the combination which yielded the highest R^2 value, averaged over single-block PLSR, MB-PLSR and SO-PLSR.

For FT-ICR MS and NIR, the inclusion of standardisation was optimal. However for FTIR, standardisation gave a slight decrease in R^2, but this was relatively small compared to the differences in R^2 for FT-ICR MS and NIR with and without standardisation. For FT-ICR MS and NIR, standardisation significantly improved the predictions. The combination that yielded the highest mean R^2 was therefore standardisation of all blocks, in addition to EMSC preprocessing for NIR. These preprocessings were used for the remainder of the analysis. The detailed results from the searches are included in the supporting information.

3.2 Comparison of single-block PLSR to MB-PLSR and SO-PLSR

In the next part of the study, single-block PLSR models on the data from FT-ICR MS, FTIR and NIR were compared to MB-PLSR and SO-PLSR models. PLSR was first performed for the individual blocks, before VIP was applied to select relevant variables in each block, and a reduced model was fitted based on this selection. Each block was standardised and mean-centred prior to analysis. The multiblock methods were then tested first on the fused full data from all blocks, and then on the variables selected from the single-block analysis. In SO-PLSR, FT-ICR MS was selected as the first block, FTIR as the second and NIR as the third, based on the amount of chemical information, where FT-ICR MS is expected to contain the most, while NIR contains the least. The prediction accuracies for the single-block PLSRs were compared to the R^2 values obtained with MB-PLSR and SO-PLSR to evaluate the gain of using multiblock methods. The results for all methods are shown in Table 1.

Variable selection method	Procedure	Selected variables APPI	Selected variables FTIR	Selected variables NIR	R^2	No. PCs
No variable selection	FT-ICR MS	23800	-	-	0.651	10
	FTIR	-	3201	-	0.874	4
	NIR	-	-	11101	0.486	4
	MB-PLSR	23800	3201	11101	0.835	8
	SO-PLSR	23800	3201	11101	0.856	1,4,3
Variables selected from single-block	FT-ICR MS	1344	-	-	0.874	20
	FTIR	-	1060	-	0.905	20
	NIR	-	-	1060	0.740	7
	MB-PLSR	1344	1060	1060	0.897	4
	SO-PLSR	1344	1060	1060	0.912	3,4,3
Variables selected from multiblock	MB-PLSR	3651	3201	1012	0.896	12
	SO-PLSR	12950	866	5019	0.912	0,4,5

Table 1: Results of single-block PLSR, MB-PLSR and SO-PLSR without variable selection, with variable selection from the single-block methods and with variable selection from MB-PLSR and SO-PLSR. Under 'Procedure', FT-ICR MS, FTIR and NIR indicate single-block analysis of the respective blocks, while MB-PLSR and SO-PLSR indicate the respective multiblock method on fused data. The number of selected variables for each method (when refitting on selected variables) is indicated, and '-' indicates exclusion of that block.

In Figure 4, the variables selected from single-block PLSR are highlighted in pink over the raw spectra for FT-ICR MS and FTIR, and preprocessed NIR spectra. The results from the single-block PLSR model for FT-ICR MS showed an increase in R^2 from 0.651 to 0.874 with a reduction in the number of variables from 23800 to 1344, i.e., 5.6% of the variables, when performing variable selection. Most of the selected variables were gathered in two areas, in m/z ∼350-300 and m/z ∼550-500, and a few variables were found below m/z 200 and around m/z 700-750. For FTIR, 1060 variables were selected, showing a slight increase in R^2, which already was high with all variables included. The selected variables were all positioned between ∼1700-800 cm^{-1}, ∼1850-1750 cm^{-1} and ∼2830-2950 cm^{-1}. Interestingly, the baseline shifts at 3000-4000 cm^{-1} were not selected as important. Lastly, in NIR, 1060 variables were selected and these were spread throughout the spectra, but with a higher abundance between 6500-9500 cm^{-1}. None of the peak maxima were selected, but several of the minima were selected, for instance at ∼8900 cm^{-1}, ∼7700 cm^{-1} and ∼6800 cm^{-1}.

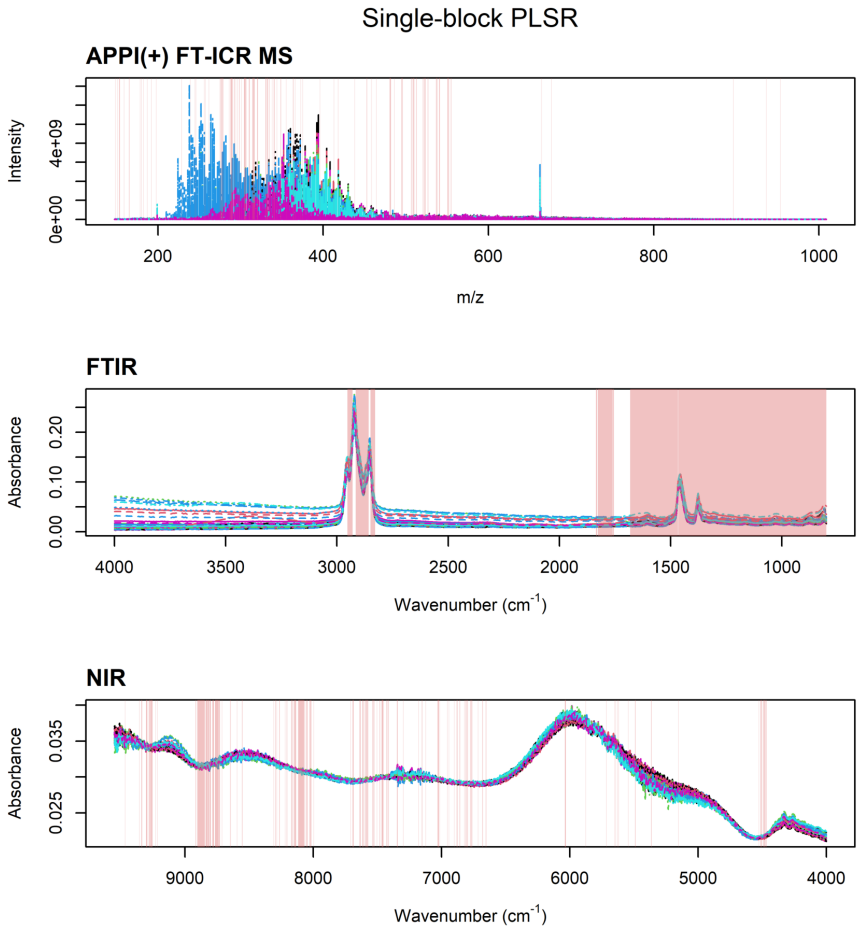


Figure 4: The variables selected from the single-block PLSR models are shown by the pink lines. VIP thresholds for selection were individually optimised for FT-ICR MS, FTIR and NIR.

FTIR gave the lowest reduction in the number of variables after variable selection, and it was also the analysis technique with the fewest variables in the original data set. In addition, FTIR had the highest R^2 with the full data of 0.875, which is quite good in itself, and it therefore makes sense that there is little to gain by removing a large number of variables.

For the multiblock methods, MB-PLSR and SO-PLSR achieved R^2 values of 0.835 and 0.856, respectively, when using the full data set. This was lower than single-block PLSR for FTIR, but higher than for both FT-ICR MS and NIR. When fitting the reduced data using MB-PLSR and SO-PLSR, the R^2 values increased to 0.897 and 0.912, respectively. The R^2 value of 0.912 for SO-PLSR was the highest achieved of all the predictions using the variables selected from the single-block analysis.

3.3 Variable selection from MB-PLSR and SO-PLSR

In the second part of the study, VIP was applied to the multi-block methods to identify which variables MB-PLSR and SO-PLSR deemed important. For MB-PLSR, VIP selected variables from the global model on the concatenated data, and the VIP threshold was determined from the global model and applied to each block. The reduced data set was then fitted to a new MB-PLSR model to evaluate the effect of the variable selection.

For SO-PLSR, VIP was applied to select variables after each PLSR modelling, meaning that for block two (FTIR) and block three (NIR), the VIP scores were determined after orthogonalisation. The VIP threshold for SO-PLSR was then determined from the global model, after all blocks had been fitted to the SO-PLSR model, meaning that one threshold was selected and applied to all blocks. The VIP threshold was then used to determine the selected variables from each block. The reduced data set was fitted to a new SO-PLSR model to evaluate the effect of the variable selection.

For MB-PLSR, Figure 5 shows the variables selected as important for each block highlighted by the pink lines. MB-PLSR selected all variables as important in FTIR, 3651 for FT-ICR MS and 1012 for NIR. This could be due to the scaling that happens during concatenation of the blocks in MB-PLSR, where each block is divided by the root of the number of variables in each block. As the FTIR data has the fewest variables of the three blocks, this scaling will have less impact than in the other two blocks, which can mean that the FTIR variables have higher values and thereby exceed the VIP threshold, so that all of them are selected. However, even though a higher number of variables were selected than for the variable selection in the single-block analysis, the R^2 value was approximately the same.

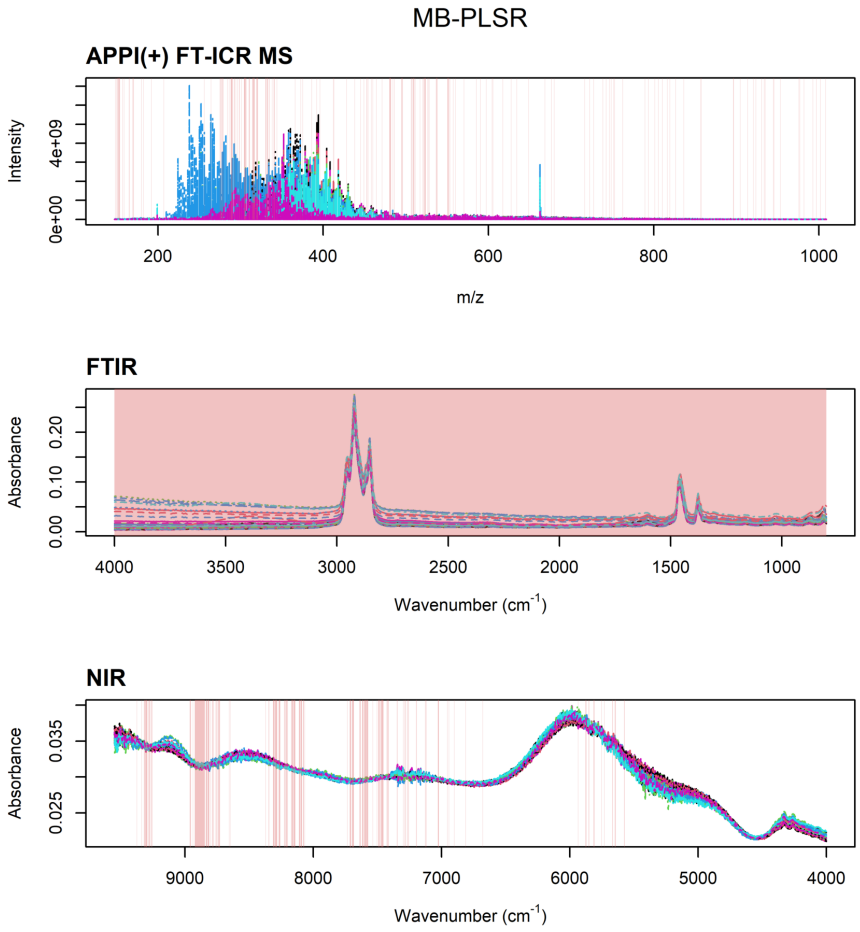


Figure 5: The variables selected from the MB-PLSR model are shown by the pink lines for the three blocks, FT-ICR MS, FTIR and NIR. The VIP threshold was determined from the global model and then applied to the individual blocks.

For SO-PLSR, Figure 6 shows the variables selected as important for each block highlighted by the pink lines. SO-PLSR selected the largest total amount of variables; 12950 for FT-ICR MS, 866 for FTIR and 5019 for NIR. But when the new SO-PLSR model was fitted to the reduced data set, the model selected zero components from the FT-ICR MS block, meaning that it did not find anything of importance from this block. As the variable selection in SO-PLSR was done from the global model, the components used for selection were 1 for FT-ICR MS, 4 for FTIR and 3 for NIR. The R^2 however, was the same when using the variables selected from SO-PLSR and when using the variables selected from the individual blocks.

The results show that SO-PLSR consistently outperforms both single-block PLSR and MB-PLSR. The lowest amount of variables were selected from the single-block

PLSR methods, yet achieving the highest R_2 with SO-PLSR.

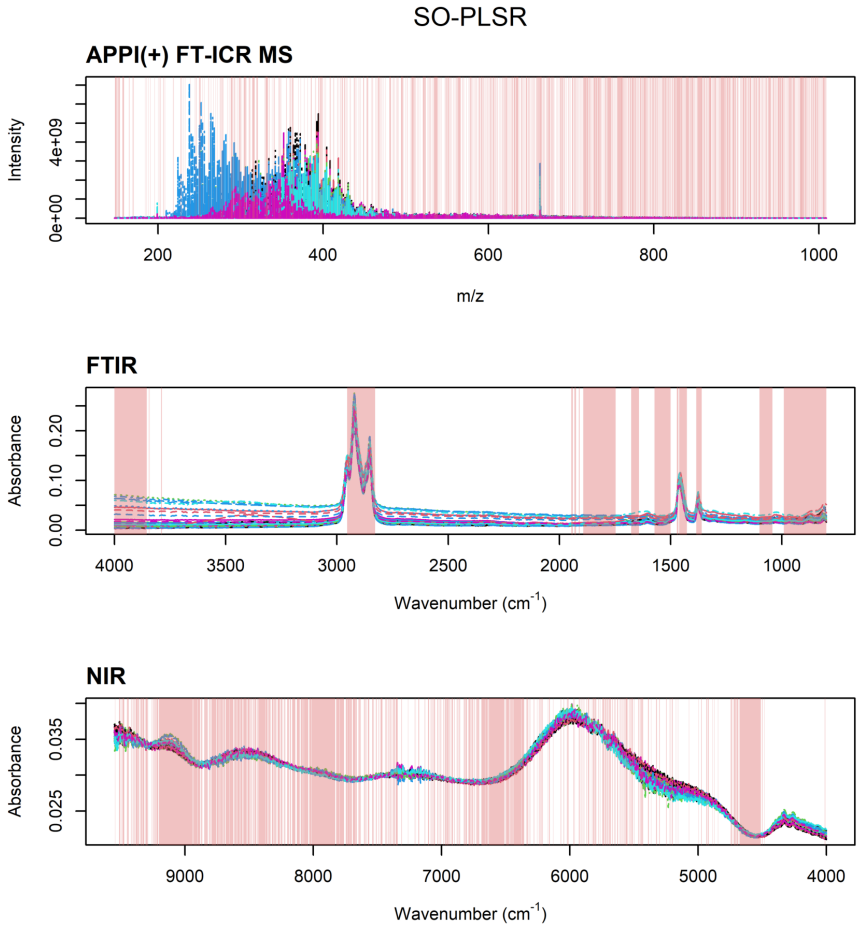


Figure 6: The variables selected from the SO-PLSR model are shown by the pink lines for the three blocks, FT-ICR MS, FTIR and NIR. The VIP threshold was determined from the global model and then applied to the individual blocks.

An interesting difference between the variables selected in SO-PLSR compared to the variables selected in MB-PLSR and in single-block PLSR, can be seen in the FT-ICR MS block. Where MB-PLSR and single-block PLSR selected mainly variables with low m/z-ratios, SO-PLSR showed a high abundance of selected variables between m/z 750-1000. The density of a sample is dependent on the size of the molecules, and large molecules contribute to higher densities. For FTIR, less importance is given to the variables below 1500 cm^{-1} in SO-PLSR than in single-block PLSR. Additionally, in SO-PLSR, variables occurring in the baseline shift area were selected, as opposed to in single-block.

3.3.1 Alternative selection methods

The above results were based on a global variable selection strategy for SO-PLSR, but other strategies for variable selection also exist. As SO-PLSR is a sequential method, the variable selection can also be implemented in a sequential manner, where the variables are selected after each PLSR modelling, and the reduced data is used to orthogonalise the next block. In this strategy, the selection for the first block will be the same as for a single-block PLSR model. This was implemented in this study as well, and resulted in a high R^2 value of 0.955 with 1344 variables selected from FT-ICR MS, 30 from FTIR and 10232 for NIR. The fact that only 30 variables were selected after the orthogonalisation of the FTIR block can be due to that all the chemical information in FTIR was also found in the variables selected in FT-ICR MS. Additionally, almost all variables were selected from the NIR block, meaning that very little information from NIR was captured in FT-ICR MS. These selections make little sense when compared to the variable selections from single-block and MB-PLSR, and the global strategy for variable selection from SO-PLSR was therefore the one selected in this study.

3.4 Interpretation of the peaks

For many chemical applications, the main interest is to identify what makes a variable important, not just to determine that it is important. Visualisation of the selected variables' positions in the spectrum can be a tool to aid interpretation of the variables. In the last part of this study, an effort was made to interpret the selected variables and connect them to the underlying chemistry affecting the density of the oils. Crude oils with high densities are known to contain more poly-aromatic and heteroatom-containing compounds and to have a relatively lower content of alkanes (saturates), while light crudes often contain more alkanes [56]. Additionally, positive correlations between the amounts of nitrogen-containing compounds, N_1 and N_1O_1 and density have been found [8].

3.4.1 FTIR

For FTIR, important peaks for interpreting spectra are for instance the spectral bands between 3500-2700 cm^{-1}, commonly referred to as the hydrogen stretching zone, where the vibrational frequencies of C-H, N-H and O-H manifest. Triple bonds between C≡C and C≡N appear in the region between 2260-2100 cm^{-1}, while double bonds between C=C, C=N and C=O appear around 1800-1650 cm^{-1}. The spectral area below 1500 cm^{-1} is called the fingerprint region, and single bonds, C-H bending and some benzene ring derivative bonds determine the type of functional groups located in this area. The aromatic region is also in the fingerprint region, and appears between 1000-400 cm^{-1}, showing the presence of aromatic rings in the samples [57, 58].

The variables selected from the FTIR block in single-block PLSR and SO-PLSR mainly coincide with the peaks observed at ∼3000-2800 cm^{-1}, the hydrogen stretching zone showing alkanes, the peaks observed at 1470 cm^{-1} and 1450 cm^{-1}, and the

remaining of the fingerprint region. As observed in Figure 5, MB-PLSR selected all variables from FTIR as important. From the single-block variable selection, the entire fingerprint region was determined to be important, and as this is the region determining the characteristics of an oil, it likely has a large effect on establishing the differences between the oils. For the variables selected in SO-PLSR, some parts of the fingerprint region were not selected, but the entire aromatic region was selected. This was expected, since higher amounts of aromatic components are associated with higher molecular weights and densities, especially if the aromatic rings are condensed, like for resins and asphaltenes.

3.4.2 NIR

Even though NIR is more limited when it comes to characterisation of chemical groups, there are some notable peaks for crude oil characterisation [59]. Among these are the spectral bands at 4500-4000 cm^{-1}, which correspond to the combination of C-H stretching and bending of CH_2 and CH_3. The weak band between 4750-4500 cm^{-1} can be assigned to the combination of fundamental vibrations in unsaturated groups. In NIR, the overtones occurring when the molecule transitions from the ground state to the exited state are visible as spectral bands. The first overtone of the fundamental C-H stretching mode is observed between 6050-5500 cm^{-1}, and the weak absorption centred at 7000 cm^{-1} is due to the combination of the C-H fundamental binding and stretching first overtone modes. Lastly, the bond centred at 8000 cm^{-1} can be attributed to the second overtones of C-H stretching. Additionally, baseline offset and a baseline slope are common in NIR, and for crude oils, this often occurs at 9000-6500 cm^{-1}. These effects are characteristic for asphaltene-containing samples.

The largest abundance of selected variables from the variable selection in NIR were present in the areas above 7000 cm^{-1}. These results were consistent between the three methods single-block PLSR, MB-PLSR and SO-PLSR. Inspection of the raw spectra in Figure 3 showed that both baseline offsets and slopes were visible, indicating the presence of varying amounts of asphaltenes and possibly other components with condensed ring structures, except for in four samples not exhibiting a baseline slope. The NIR data set was the block with the largest variations between the selected variables. Single-block PLSR, MB-PLSR and SO-PLSR selected many of the same areas, but slightly different peaks were identified as important for each area. It is, however, interesting that the spectral bands associated with C-H stretching, which are mainly indications of alkanes, showed low importance for all methods. Some spectral bands were selected at ∼4500-4400 cm^{-1}, corresponding to alkane stretching and bending, and in the first overtone centred at 5500 cm^{-1}, in single-block PLSR. SO-PLSR, on the other hand, selected the variables corresponding to the spectral bands for unsaturated groups at 4750-4500 cm^{-1}.

3.4.3 FT-ICR MS

The interpretation of MS spectra is different from that of IR spectra. While IR is sectioned into groups of similar characteristics, the position of a peak in an MS

spectrum is related to the mass of the compound. The m/z for a peak is directly corresponding to the mass when the charge of the molecule is 1, and as APPI ionises molecules to a charge of 1, the peak position reveals the mass of the molecule. To be able to interpret peaks in an MS spectrum accurately, their molecular formulas have to be identified using a suitable spectral processing software. However, some chemical groups have been shown to appear in certain masses. For instance, asphaltenes have an average mass of ∼750 Da, meaning that asphaltenic molecules are seen as peaks around m/z 750. Naphthenic acids, with an average mass between 300-500 Da [60], are usually only found in low concentrations in light oils, but in higher concentrations in heavy crude oils, and then often in the form of high-density naphthenes [61]. Another important factor in interpreting peaks in an MS spectrum is the double bond equivalent (DBE) for the molecule. The DBE is a measure of the degree of unsaturation present in the molecule, and reveals the number of rings or double bonds present. As it is known that light oils mainly contain saturated alkanes, they will have low DBEs, while heavy oils containing many aromatic groups or many double-bonded side chains, will have high DBEs.

For all methods, many of the selected variables were present in the m/z area 300-400, which corresponds to the average mass of naphthenic acids. For single-block PLSR and MB-PLSR, a few variables were also selected between m/z 650-750, i.e., the asphaltene area. For SO-PLSR, Figure 6 shows that a higher abundance of variables was selected above m/z 750, and the variables below this value were more sparsely selected. As mentioned above, higher molecular weights lead to higher densities.

Molecular formulas could be determined for 1905 of the variables selected over all three methods, and these were positioned in the m/z area 165-918. A high mass implies a large number of possible combinations of chemical groups, which makes it more difficult to determine the molecular formula. Many variables are isotope peaks, and in this study, molecular formulas were only calculated for monoisotopic peaks. An isotope peak refers to the peak for isotopes in the molecule, for instance, ^{13}C or ^{2}H. Each molecular peak can usually have up to 4 isotope peaks, meaning that many of the selected variables probably are parts of isotope patterns, and are therefore simply reflections of the molecular peak. For the determined molecular formulas, the number of carbon atoms ranged from 9-61, with 24 being the most frequent. The DBEs for the determined molecular formulas ranged from 0-31, with seven being the most frequent value. The DBEs also revealed that 66.6% of the molecules had DBEs lower than 10, indicating a moderate complexity and density of the ring structures, corresponding to lighter and medium oils. The remaining DBEs between 10-31 could be responsible for the high densities of the heavy oil samples. This fits well with the measured densities, where 71.4% are below 0.90 g/mL.

The variables selected as important in APPI(+) FT-ICR MS which were identified with molecular formulas by Compass DataAnalysis are included in the Supporting Information.

4 Conclusion

In this study, the density of crude oils was predicted from APPI(+) FT-ICR MS, FTIR and NIR spectra. MB-PLSR and SO-PLSR were tested to evaluate the gain of fusing the data with multiblock methods. For each method, VIP was utilised to identify important variables, and the effects of the selected variables on the density, and thereby the oil chemistry, of the samples were interpreted.

For the prediction of density from single-block analysis of the individual blocks, the FTIR data block yielded the highest prediction accuracy, indicating that the strongest relationship to density can be found in the FTIR area. FT-ICR MS was, due to the high mass resolution, expected to contain more chemical information than FTIR and NIR, and therefore expected to be more accurate. The prediction accuracy does increase after variable selection, suggesting that the full spectra contain too many peaks not related to the response which therefore diminish the predictive power of the model. For the NIR data, the prediction accuracy also increased significantly after using variable selection to reduce the number of variables. This indicates that removal of redundant or irrelevant variables is important when analysing data from both FT-ICR MS and NIR.

Out of the three PLSR-based methods, SO-PLSR had the highest prediction accuracy after variable selection, both when using variables selected from the single-block analysis and with variables selected from the SO-PLSR itself. This illustrates the advantage of fusing data from several sources compared to using data from only one analytical technique. During interpretation of the variables, it was discovered that SO-PLSR more frequently selected variables corresponding to aromatic groups, as opposed to MB-PLSR and single-block PLSR. Higher contents of aromatic groups have been related to higher densities, and the fact that SO-PLSR consistently performed better than single-block and MB-PLSR shows that SO-PLSR selected more relevant parts of the spectra. This indicates that the spectral regions corresponding to aromatics are the most efficient to use when predicting density, which also corresponds with the literature.

This study showed that fusing data from multiple spectroscopic sources increases the prediction ability compared to separate analysis of the individual blocks. This illustrates the potential of multiblock analysis for more complicated prediction problems.

Acknowledgement

The authors thank The Research Council of Norway, Equinor ASA, OMV (Norge) AS, Wintershall DEA Norge AS and TotalEnergies for funding. This work is a part of the Knowledge-Building Project for Industry (PETROMAKS 2), Project number: 294636 “New Hydrate Management: New understanding of hydrate phenomena in oil systems to enable safe operation within the hydrate zone”.

We further acknowledge Magnus Fossen Nordborg for conducting the density measurements.

Declarations

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The data analysed during the current study are available in the Zenodo repository DOI: 10.5281/zenodo.7673496

Supporting information

S1 Table. Results from the comparison of the different procedures for preprocessing. Tables showing the R^2 for the different combinations of preprocessings with standardisation and mean-centring.

S2 Table. Molecular formulas for the variables selected from FT-ICR MS. The selected variables for FT-ICR MS over all PLSR based methods which were identified with a molecular formula from Compass DataAnalysis.

References

[1] C. S. Hsu et al. "Petroleomics: advanced molecular probe for petroleum heavy ends". In: *Journal of Mass Spectrometry* 46.4 (Mar. 2011), pp. 337–343. DOI: 10.1002/jms.1893.

[2] A. G. Marshall and R. P. Rodgers. "Petroleomics: The Next Grand Challenge for Chemical Analysis". In: *Accounts of Chemical Research* 37.1 (Jan. 2004), pp. 53–59. DOI: 10.1021/ar020177t.

[3] J. G. Speight. *The Chemistry and Technology of Petroleum*. 4th ed. Chemical Industries 114. Boca Raton, FL: CRC Press, 2006.

[4] F. Brakstad et al. "Prediction of molecular weight and density of distillation fractions from gas chromatographic—mass spectrometric detection and multivariate calibration". In: *Chemometrics and Intelligent Laboratory Systems* 3.4 (June 1988), pp. 321–328. DOI: 10.1016/0169-7439(88)80031-5.

[5] Z. S. Baird and V. Oja. "Predicting fuel properties using chemometrics: a review and an extension to temperature dependent physical properties by using infrared spectroscopy to predict density". In: *Chemometrics and Intelligent Laboratory Systems* 158 (Nov. 2016), pp. 41–47. DOI: 10.1016/j.chemolab.2016.08.004.

[6] E. L. Gjelsvik, M. Fossen, and K. Tøndel. "Current overview and way forward for the use of machine learning in the field of petroleum gas hydrates". In: *Fuel* 334.2 (Feb. 2023), p. 126696. DOI: 10.1016/j.fuel.2022.126696.

[7] C. A. Hughey, R. P. Rodgers, and A. G. Marshall. "Resolution of 11000 Compositionally Distinct Components in a Single Electrospray Ionization Fourier Transform Ion Cyclotron Resonance Mass Spectrum of Crude Oil". In: *Analytical Chemistry* 74.16 (June 2002), pp. 4145–4149. DOI: 10.1021/ac020146b.

[8] M. Hur et al. "Correlation of FT-ICR Mass Spectra with the Chemical and Physical Properties of Associated Crude Oils". en. In: *Energy Fuels* 24 (Aug. 2010), pp. 5524–5532. DOI: 10.1021/ef1007165.

[9] G. C. Klein et al. "Use of Saturates/Aromatics/Resins/Asphaltenes (SARA) Fractionation To Determine Matrix Effects in Crude Oil Analysis by Electrospray Ionization Fourier Transform Ion Cyclotron Resonance Mass Spectrometry". In: *Energy Fuels* 20.2 (Feb. 2006), pp. 668–672. DOI: 10.1021/ef050353p.

[10] T. M. Schaub et al. "Heat-Exchanger Deposits in an Inverted Steam-Assisted Gravity Drainage Operation. Part 2. Organic Acid Analysis by Electrospray Ionization Fourier Transform Ion Cyclotron Resonance Mass Spectrometry". In: *Energy Fuels* 21.1 (Jan. 2007), pp. 185–194. DOI: 10.1021/ef0601115.

[11] D. F. Smith et al. "Characterization of Athabasca Bitumen Heavy Vacuum Gas Oil Distillation Cuts by Negative/Positive Electrospray Ionization and Automated Liquid Injection Field Desorption Ionization Fourier Transform Ion Cyclotron Resonance Mass Spectrometry". In: *Energy Fuels* 22.5 (Sept. 2008), pp. 3118–3125. DOI: 10.1021/ef8000357.

[12] J. V. Headley et al. "Characterization of Naphthenic Acids from Athabasca Oil Sands Using Electrospray Ionization: The Significant Influence of Solvents". In: *Analytical Chemistry* 79.16 (Aug. 2007), pp. 6222–6229. DOI: 10.1021/ac070905w.

[13] M. P. Barrow et al. "Data Visualization for the Characterization of Naphthenic Acids within Petroleum Samples". In: *Energy Fuels* 23.5 (Mar. 2009), pp. 2592–2599. DOI: 10.1021/ef800985z.

[14] F. A. Fernandez-Lima et al. "Petroleum Crude Oil Characterization by IMS-MS and FTICR MS". In: *Analytical Chemistry* 81.24 (Dec. 2009), pp. 9941–9947. DOI: 10.1021/ac901594f.

[15] Y. Cho et al. "Developments in FT-ICR MS instrumentation, ionization techniques, and data interpretation methods for petroleomics". In: *Mass Spectrometry Reviews* 34.2 (Mar. 2014), pp. 248–263. DOI: 10.1002/mas.21438.

[16] J. M. Purcell et al. "Atmospheric Pressure Photoionization Fourier Transform Ion Cyclotron Resonance Mass Spectrometry for Complex Mixture Analysis". In: *Analytical Chemistry* 78.16 (July 2006), pp. 5906–5912. DOI: 10.1021/ac060754h.

[17] A. G. Marshall and R. P. Rodgers. "Petroleomics: Chemistry of the underworld". In: *Proceedings of the National Academy of Sciences of the United States of America* 105.47 (Nov. 2008), pp. 18090–18095. DOI: 10.1073/pnas.0805069105.

[18] M. K. Moro et al. “A review of chemometrics models to predict crude oil properties from nuclear magnetic resonance and infrared spectroscopy”. In: *Fuel* 303 (Nov. 2021), p. 121283. DOI: 10.1016/j.fuel.2021.121283.

[19] E. V. Barros et al. “Characterization of naphthenic acids in crude oil samples – A literature review”. In: *Fuel* 319 (July 2022), p. 123775. DOI: 10.1016/j.fuel.2022.123775.

[20] D. Williams and I. Fleming. *Spectroscopic methods in organic chemistry*. 6th ed. UK: McGraw-Hill Education, 2008.

[21] M. Fossen et al. “Solubility Parameters Based on IR and NIR Spectra: I. Correlation to Polar Solutes and Binary Systems”. In: *Journal of Dispersion Science and Technology* 26.2 (Sept. 2004), pp. 227–241. DOI: 10.1081/DIS-200045605.

[22] N. Aske, H. Kallevik, and J. Sjöblom. “Determination of Saturate, Aromatic, Resin, and Asphaltenic (SARA) Components in Crude Oils by Means of Infrared and Near-Infrared Spectroscopy”. In: *Energy Fuels* 15.5 (Aug. 2001), pp. 1304–1312. DOI: 10.1021/ef010088h.

[23] M. P. Barrow et al. “Athabasca Oil Sands Process Water: Characterization by Atmospheric Pressure Photoionization and Electrospray Ionization Fourier Transform Ion Cyclotron Resonance Mass Spectrometry”. In: *Analytical Chemistry* 82.9 (May 2010), pp. 3727–3735. DOI: 10.1021/ac100103y.

[24] S. Lababidi and W. Schrader. “Online normal-phase high-performance liquid chromatography/Fourier transform ion cyclotron resonance mass spectrometry: Effects of different ionization methods on the characterization of highly complex crude oil mixtures”. In: *Rapid Communications in Mass Spectrometry* 28.12 (June 2014), pp. 1345–1352. DOI: 10.1002/rcm.6907.

[25] S. Chiaberge et al. “Bio-Oil from Waste: A Comprehensive Analytical Study by Soft-Ionization FTICR Mass Spectrometry”. In: *Energy Fuels* 28.3 (Mar. 2014), pp. 2019–2026. DOI: 10.1021/ef402452f.

[26] P. Mishra et al. “Recent trends in multi-block data analysis in chemometrics for multi-source data integration”. In: *TrAC Trends in Analytical Chemistry* 137 (Apr. 2021), p. 116206. DOI: 10.1016/j.trac.2021.116206.

[27] T. I. Dearing et al. “Characterization of Crude Oil Products Using Data Fusion of Process Raman, Infrared, and Nuclear Magnetic Resonance (NMR) Spectra”. In: *Applied Spectroscopy* 65.2 (Feb. 2011), pp. 181–186. DOI: 10.1366/10-05974.

[28] H.-P. Wang et al. “Recent advances of chemometric calibration methods in modern spectroscopy: Algorithms, strategy, and related issues”. In: *TrAC Trends in Analytical Chemistry* 153 (Aug. 2022), p. 116648. DOI: 10.1016/j.trac.2022.116648.

[29] Y.-H. Yun et al. “An overview of variable selection methods in multivariate analysis of near-infrared spectra”. In: *TrAC Trends in Analytical Chemistry* 113 (Apr. 2019), pp. 102–115. DOI: 10.1016/j.trac.2019.01.018.

[30] J. M. Santos et al. “Comparing Crude Oils with Different API Gravities on a Molecular Level Using Mass Spectrometric Analysis. Part 1: Whole Crude Oil”. In: *Energies* 11.10 (Oct. 2018), p. 2766. DOI: 10.3390/en11102766.

[31] *PerkinElmer Frontier FT-IR, NIR and FIR Spectroscopy Brochure*. 2011.

[32] H. Martens and E. Stark. “Extended multiplicative signal correction and spectral interference subtraction: New preprocessing methods for near infrared spectroscopy”. In: *Journal of Pharmaceutical and Biomedical Analysis* 9.8 (1991), pp. 625–635. DOI: 10.1016/0731-7085(91)80188-F.

[33] A. Kohler et al. “Estimating and Correcting Mie Scattering in Synchrotron-Based Microscopic Fourier Transform Infrared Spectra by Extended Multiplicative Signal Correction”. In: *Applied Spectroscopy* 62.3 (2008), pp. 259–266. DOI: 10.1366/000370208783759669.

[34] A. Savitzky and M. J. E. Golay. “Smoothing and Differentiation of Data by Simplified Least Squares Procedures”. In: *Analytical Chemistry* 36.8 (July 1964), pp. 1627–1639. DOI: 10.1021/ac60214a047.

[35] T. De Meyer et al. “NMR-Based Characterization of Metabolic Alterations in Hypertension Using an Adaptive, Intelligent Binning Algorithm”. In: *Analytical Chemistry* 50.10 (May 2008), pp. 3783–3790. DOI: 10.1021/ac7025964.

[36] Bruker Daltonik GmbH. *Bruker Compass ProfileAnalysis*. 2013.

[37] Bruker Daltonik GmbH. *Bruker Compass DataAnalysis*. 2017.

[38] S. Wold, H. Martens, and H. Wold. “The multivariate calibration problem in chemistry solved by the PLS method”. In: *Matrix Pencils*. Lecture Notes in Mathematics 973. Berlin, Heidelberg: Springer, 1983, pp. 286–293.

[39] S. Wold et al. “The Collinearity Problem in Linear Regression. The Partial Least Squares (PLS) Approach to Generalized Inverses”. In: *SIAM Journal on Scientific and Statistical Computing* 5.3 (Sept. 1984), pp. 735–743. DOI: 10.1137/0905052.

[40] H. Martens and T. Næs. *Multivariate Calibration*. 1st ed. Chichester: Wiley, July 1992.

[41] A. K. Smilde, T. Næs, and K. H. Liland. *Multiblock Data Fusion in Statistics and Machine Learning: Applications in the Natural and Life Sciences*. 1st ed. John Wiley & Sons, Ltd, Apr. 2022.

[42] A. K. Smilde et al. “Common and distinct components in data fusion”. In: *Journal of Chemometrics* 31.7 (July 2017), e2900. DOI: 10.1002/cem.2900.

[43] L. E. Wangen and B. R. Kowalski. “A multiblock partial least squares algorithm for investigating complex chemical systems”. In: *Journal of Chemometrics* 3.1 (Jan. 1989), pp. 3–20. DOI: 10.1002/cem.1180030104.

[44] J. A. Westerhuis, T. Kourti, and J. F. MacGregor. “Analysis of multiblock and hierarchical PCA and PLS models”. In: *Journal of Chemometrics* 12.5 (Dec. 1998), pp. 301–321. DOI: 10.1002/(SICI)1099-128X(199809/10)12:5<301::AID-CEM515>3.0.CO;2-S.

[45] J. A. Westerhuis and A. K. Smilde. "Deflation in multiblock PLS". In: *Journal of Chemometrics* 15.5 (June 2001), pp. 485–493. DOI: 10.1002/cem.652.

[46] T. Næs et al. "Path modelling by sequential PLS regression". In: *Journal of Chemometrics* 25.1 (Jan. 2011), pp. 28–40. DOI: 10.1002/cem.1357.

[47] A. Biancolillo and T. Næs. "The sequential and orthogonalized PLS regression for multiblock regression: theory, examples, and extensions". In: *Data Fusion Methodology and Applications*. Vol. 31. Data Handling in Science and Technology. Elsevier, 2019, pp. 157–177.

[48] T. Mehmood et al. "A review of variable selection methods in Partial Least Squares Regression". In: *Chemometrics and Intelligent Laboratory Systems* 118 (Aug. 2012), pp. 62–69. DOI: 10.1016/j.chemolab.2012.07.010.

[49] S. Favilla et al. "Assessing feature relevance in NPLS models by VIP". In: *Chemometrics and Intelligent Laboratory Systems* 129 (Nov. 2013), pp. 76–86. DOI: 10.1016/j.chemolab.2013.05.013.

[50] M. Farrés et al. "Comparison of the variable importance in projection (VIP) and of the selectivity ratio (SR) methods for variable selection and interpretation". In: *Journal of Chemometrics* 29.10 (June 2015), pp. 528–536. DOI: 10.1002/cem.2736.

[51] A. Biancolillo et al. "Variable selection in multi-block regression". In: *Chemometrics and Intelligent Laboratory Systems* 156 (Aug. 2016), pp. 89–101. DOI: 10.1016/j.chemolab.2016.05.016.

[52] T. N. Tran et al. "Interpretation of variable importance in Partial Least Squares with Significance Multivariate Correlation (sMC)". In: *Chemometrics and Intelligent Laboratory Systems* 138 (Nov. 2014), pp. 153–160. DOI: 10.1016/j.chemolab.2014.08.005.

[53] R Core Team. *R: A Language and Environment for Statistical Computing*. Vienna, Austria, 2022.

[54] K. H. Liland, B.-H. Mevik, and R. Wehrens. *pls: Partial Least Squares and Principal Component Regression*. 2021.

[55] K. H. Liland. *multiblock: Multiblock Data Fusion in Statistics and Machine Learning*. 2022.

[56] M. M. Boduszynski. "Composition of heavy petroleums. 1. Molecular weight, hydrogen deficiency, and heteroatom concentration as a function of atmospheric equivalent boiling point up to 1400. °F (760 °C)". In: *Energy Fuels* 1.1 (Jan. 1987), pp. 2–11. DOI: 10.1021/ef00001a001.

[57] R. M. Silverstein and G. C. Bassler. "Spectrometric identification of organic compounds". In: *Journal of Chemical Education* 39.11 (1962), p. 546. DOI: 10.1021/ed039p546.

[58] I. Zojaji, A. Esfandiarian, and J. Taheri-Shakib. "Toward molecular characterization of asphaltene from different origins under different conditions by means of FT-IR spectroscopy". In: *Advances in Colloid and Interface Science* 289 (Mar. 2021), p. 102314. DOI: 10.1016/j.cis.2020.102314.

[59] J. Laxalde et al. "Characterisation of heavy oils using near-infrared spectroscopy: Optimisation of pre-processing methods and variable selection". In: *Analytica Chimica Acta* 705.1-2 (Oct. 2011), pp. 227–234. DOI: 10.1016/j.aca.2011.05.048.

[60] C. Hurtevent et al. "Production Issues of Acidic Petroleum Crude Oils". In: *Emulsions and Emulsion Stability*. 2nd ed. CRC Press, 2005, p. 40.

[61] C. Yang et al. "Characterization of naphthenic acids in crude oils and refined petroleum products". In: *Fuel* 255 (Nov. 2019), p. 115849. DOI: 10.1016/j.fuel.2019.115849.

Paper V

Gjelsvik E.L., Tøndel K., Hierarchical cluster-based deep learning, *Manuscript*

Hierarchical cluster-based deep learning

Elise Lunde Gjelsvik[1] and Kristin Tøndel[1]

[1]*Faculty of Science and Technology, Norwegian University of Life Sciences, Aas, Norway*

Abstract

Prediction models based on data with large inhomogeneity or collinearity often perform poorly because relationships between groups in the data dominate the model. This can be overcome by splitting the data into smaller clusters and creating a local model within each cluster. In this study, the Hierarchical Cluster-based Partial Least Squares Regression (HC-PLSR) procedure was expanded to deep learning. Hierarchical Cluster-based convolutional neural networks (HC-CNNs), Hierarchical Cluster-based recurrent neural networks (HC-RNNs) and Hierarchical Cluster-based Support Vector Regression models (HC-SVRs) were implemented and tested on spectroscopic FT-IR data. The data consisted of FT-IR measurements of raw material dry films from chicken, turkey, mackerel and salmon after enzymatic hydrolysis, for prediction of average molecular weight during hydrolysis. The deep learning models, HC-CNN, HC-RNN and HC-SVR outperformed HC-PLSR, showing the disadvantage of PLSR for non-linear data. An interpretation of the importance of the features for predicting the response based on measures provided by each of the methods was done to evaluate the similarities and differences between the prediction models.

Keywords

Local modelling; Fuzzy C-Means clustering; Spectral clustering; Hierarchical agglomerative clustering; Hierarchical Cluster-based Partial Least Squares Regression (HC-PLSR); Hierarchical deep learning

1 Introduction

For complex data sets with large inhomogeneity or collinearity, prediction models often perform poorly. In these cases, the relationships between groups in the data dominate the model and the prediction approaches the average of the closest group [1]. An important step in achieving a good prediction model, is to identify and model these underlying structures. Poor prediction can also occur when there are several different competing interrelationships overshadowing the prediction problem of interest. Real world data often have unknown structures and hidden relationships which can be difficult to model. To overcome this problem, the data can be split into smaller clusters of more homogeneous data. If the

split is successful, the estimation of the relationships among both variables and observations will achieve an improved prediction. The split can be based on any criterion, i.e. be manually determined or assigned by a clustering algorithm. The clustering can find the hidden structures without having any prior knowledge about the data. This approach yields a set of models, one for each cluster, where each of them are fitted to the data within that cluster, overcoming the shortcomings of a model based on all the data.

Local linear approaches for modelling large and complex data sets have been shown to work well, and Partial Least Squares Regression (PLSR) have been used in several studies where the data has been divided based on prior knowledge [1, 2, 3, 4, 5, 6, 7]. However, the data does not always have known groups to base the separation on, and then other approaches are required. For instance, Tøndel et al. [8, 9] created a hierarchical PLSR model (HC-PLSR), where the observations were assigned into clusters based on Fuzzy C-Means (FCM) clustering which assumes no prior knowledge about the data structure. Fuzzy clustering methods have also been developed within a framework that determines the optimal number of clusters [10, 11].

The advantage of PLSR is its efficient ways of finding latent variables in the data fast and reliably. Nevertheless, a potential pitfall when it comes to PLSR, is that this is a linear model, while the data could have non-linear interrelationships where PLSR struggle to perform well. The input space the data lies in can exist in different planes, be high-dimensional, low-dimensional, linear, non-linear etc. Even though simple non-linearities can be accounted for by e.g. including polynomial terms in the regressor matrix, and more abrupt non-linearities can be handled by local modelling, some types of non-linearities can not be modelled using PLSR. In such cases, there is a need for a method able to handle more complex structures, as even with HC-PLSR, it is a requirement that the data are at least locally linear.

Neural networks are powerful when it comes to modelling non-linear and large data sets. However, the more complex the data becomes, the deeper the network needs to be to achieve adequate prediction. A deep network contains numerous parameters and uses large amounts of computational power to converge. For local modelling, smaller networks can be implemented without loosing predictability, rather increasing predictive power. The interpretability of the model is also higher for simpler models, and the risk of overfitting decreases.

In this study, the framework for HC-PLSR was extended to deep learning based models in an attempt to improve the predictive power further. Local Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs) and Support Vector Regression (SVR) models were created and compared to the local PLSR models.

1.1 Partial Least Squares Regression (PLSR)

PLSR and its algorithm have previously been described rigorously in the literature [12, 13, 14, 15]. In short, PLSR decomposes large data sets into a subspace of latent variables (scores and loadings) representing the main features of covariance between $\boldsymbol{X}$ (regressors) and $\boldsymbol{Y}$ (response), where both $\boldsymbol{X}$ and $\boldsymbol{Y}$ can be multivariate. PLSR uses inter-correlations between the response variables to stabilise the model, and does not require that the variables are linearly independent. The latent variables, the PLS components (PCs), represent the

most relevant subspaces of the regressors. This is beneficial as it makes PLSR able to handle a wide range of data, including chemometric data, where the number of features often exceeds the number of samples.

1.2 Convolutional Neural Networks (CNNs)

CNNs is a type of deep neural networks which uses convolutions to extract information in one or more hidden layers [16, 17]. CNNs are regularized versions of fully connected networks and consist of an input layer, hidden layers (mainly convolutional layers, pooling and fully connected layers) and an output layer. In the convolutional layers, the data is organised in a feature map where the weights are connected to the previous layer. These weights are used to filter for patterns in the data. The pooling layer semantically merges similar features, reducing the dimension of the representation [17]. Commonly used in pattern recognition, CNNs are good feature extractors as they learn the most important features by themselves. In contrast to PLSR, CNNs can handle non-linear data.

1.3 Recurrent Neural Networks (RNNs)

RNNs is a type of neural networks often used for sequential or time-series data, feeding the output from one layer as input to the next layer [18, 19]. Like CNNs, the RNNs learn from the training input, but the RNNs use internal states (memory) to impact inputs and outputs with previous information. Therefore, RNNs have a strong capability of capturing contextual data from a sequence. In a RNN, the input sequence is processed one element at a time with the memory in the hidden units retaining information on all the elements in the sequence [17]. In the case of chemical spectra, the peaks are often related to adjacent peaks or can appear in certain patterns, which is why pattern recognition methods such as CNNs and RNNs often achieve good prediction models for this type of data.

1.4 Support Vector Regression (SVR)

In SVR, a hyperplane or a set of hyperplanes are constructed in a high-dimensional space to separate the observations in the training set [20, 21, 22]. The aim is to find the hyperplane that has the largest distance (margin) to the nearest data point. The margin is defined as the distance between the separating hyperplane, the decision boundary, and the training samples that are closest to this hyperplane. The hyperplane is used to predict the continuous output and the regression solution is the hyperplane that has the maximum number of data points. Decision boundaries with large margins tend to have a lower generalisation error, while decision boundaries with small margins are more prone to overfitting. This makes SVRs proficient at handling non-linearities in data and as the hyperplanes are constructed in a high-dimensional space, SVR can handle data which cannot be separated in the first two or three dimensions.

2 Material and methods

2.1 Fuzzy C-Means clustering (FCM)

Cluster analysis consists of assigning data into groups in a way that the data points in the same group are as similar to each other and as dissimilar to data points in other groups as possible. These clusters are defined based on a similarity measure, for instance, distance, connectivity or intensity. In Fuzzy clustering, each data point can belong to more than one cluster, where cluster membership probabilities define to which degree a sample belongs to the different clusters. The closer to the centroid the sample is, the higher the membership probabilities become. The FCM algorithm [23, 24] chooses a number of clusters where probabilities for being in the clusters are assigned randomly to each data point. This is repeated until the algorithm has converged, i.e. until the changes in the probabilities no longer exceed the set sensitivity threshold. The centroid is then computed for each cluster and finally, the membership probabilities for each sample for being in each of the clusters are computed.

2.2 Alternative clustering techniques

Spectral clustering (SPC) [25] uses the spectrum (eigenvalues) of the similarity matrix of the data to reduce dimensionality, so that the clustering can be done in fewer dimensions. The similarity matrix consists of an assessment of the relative similarity for each pair of points in the data set. SPC is useful when the structure of the clusters is non-convex, when the center and spread of the cluster give a poor description of the properties of the cluster.

In hierarchical agglomerative clustering (HAC) [26], nested clusters are built by successive merging or splitting. This hierarchy of clusters can be presented as a dendrogram, where the branches consist of unique clusters for each sample, merging as pairs of clusters are combined until all samples of the data set are incorporated into one cluster, the root. The distance between two subsets of the data is called the linkage distance and represents the distance between samples in the clusters, and thereby also the cluster regions.

2.3 Local modelling

Implementation of the local modelling was based on the HC-PLSR procedure developed by Tøndel et. al [8]. First, a PLSR model was built using all observations in the training set, and the optimal number of PCs was determined using a Leave-One-Out cross validation (LOOCV) to reduce the risk of overfitting. The selection of PCs was done based on the minimum cross-validated mean squared error (MSE) over the LOOCV models to create the global PLSR model. The X-scores for the samples from the global PLSR model were then clustered by FCM using Euclidean distance. The optimal number of clusters were determined using LOOCV on the training set. For clusters containing less than 10 samples, the samples were reassigned based on their membership probabilities, until they were placed in a cluster with more than 10 samples.

For each of the determined clusters, local models (PLSR, CNN, RNN and SVR) were calibrated individually using LOOCV to find the optimal local model parameters (PCs

for PLSR, training length and epochs for CNN and RNN and hyperparameters for SVR). New (test set) observations, were projected into the global PLSR model and their X-scores were calculated. The resulting X-scores were then classified into the appropriate clusters based on FCM, Linear Discriminant Analysis (LDA) [27], Quadratic Discriminant Analysis (QDA) [28] or Naive Bayes classification (NB) [28]. Prediction of the response for the new observations was done using the local model for the assigned cluster. The resulting predictions were compared to those obtained using the global PLSR, CNN, RNN and SVR models to evaluate the improvements achieved using the local modelling.

In addition to FCM, several other clustering methods were fitted to the data to evaluate the clustering proficiency. HAC and SPC showed some interesting clusters that differed from FCM, and they were implemented in the algorithm. Local modelling using HCA and SPC for clustering instead of FCM were created and the results were compared to the results using FCM. The remaining methods did not yield a meaningful groupings of the data and were therefore rejected. All cluster distributions are presented in the supporting information.

All statistical methods were implemented using Python 3.8 and its machine learning packages. FCM was implemented using the fuzzy-c-means package [29]. All calculations were done on the Orion High Performance Computing Center (OHPCC) at the Norwegian University of Life Sciences (NMBU).

2.4 Visualising important features

To evaluate which features the CNN and RNN estimated to be important during building of the local models, an effort was made to visualise the gradients for each of the local models. This visualisation was based on the variational gradient method (VarGrad) [30, 31] which previously has shown good results compared to similar visualisation methods [32]. Jenul *et al.* showed that this method works well for determining importance of blocks in a multiblock data set [33], and this procedure was therefore adapted to determine the importance of features. In VarGrad, a small random noise (using the Numpy Random Generator) is added to the input layer and then the gradient function for each feature is estimated. The resulting variation in the gradient indicates which of the features the output from the network is most dependent on. These features are deemed important, and their effect on the prediction can be evaluated.

For PLSR, the loadings gives an overview of which features that accounts for the variation explained by the response. SVR on the other hand, does not have a built in method for feature evaluation, and therefore permutation feature importance was used. Permutation feature importance is a model inspection technique that identifies important features based on changes in the prediction accuracy when a feature is randomly shuffled [34]. If the prediction accuracy of the model decreases significantly when a feature is randomly shuffled, this indicates that the feature is important for the models ability to predict the response. Similarly, if the prediction accuracy is unaffected, the feature is not important for the prediction.

2.5 Data

The data set used to compare the methods consists of Fourier-transform infrared spectroscopy (FT-IR) measurements of raw material dry films from chicken, turkey, salmon and mackerel hydrolysed by six different enzymes, and was retrieved from Kristoffersen et al. [3]. The pre-processing was performed using Savitzky-Golay 2^{nd} derivative smoothing (with window width 11 pt and 3^{rd} order polynomial smoothing) followed by extended multiplicative signal correction (EMSC) with 2^{nd} order polynomial correction, with the mean spectrum as reference. Lastly, the spectra were cut to contain the region between 1800 cm^{-1} and 700 cm^{-1} based on prior knowledge about the relevance of different parts of the spectra.

All samples were measured in replicates, and the average spectrum was calculated over the replica for each sample, resulting in 332 unique samples. The data set included information about 28 different subgroups consisting of the six enzymes used for hydrolysis combined with the 4 different raw materials. To gain an understanding of the data, the mean of the samples for each animal and mean of the samples for each enzyme was plotted. This was done for both the raw data and for the pre-processed data, and the spectra are shown in Figure 1. The enzymes used were Alcalase, Papain, Protamex, Flavorzyme, Corolase 2TS, additionally some of the mackerel samples were self-hydrolysed (labelled NaN in Figure 1). The response to be predicted was the average molecular weight (M_w) during the enzymatic hydrolysis of the raw materials. The data was mean centred (mean of 0) before PLSR and mean centred and standardised (mean of 0, std of 1) before CNN, RNN and SVM. The data was divided into training and test sets with half of the data set (50 %) in each. This split was chosen to get a good representation for each of the 28 groups in both the training and test data. Determination of the parameters for the global PLSR model and all local models was done on the training data.

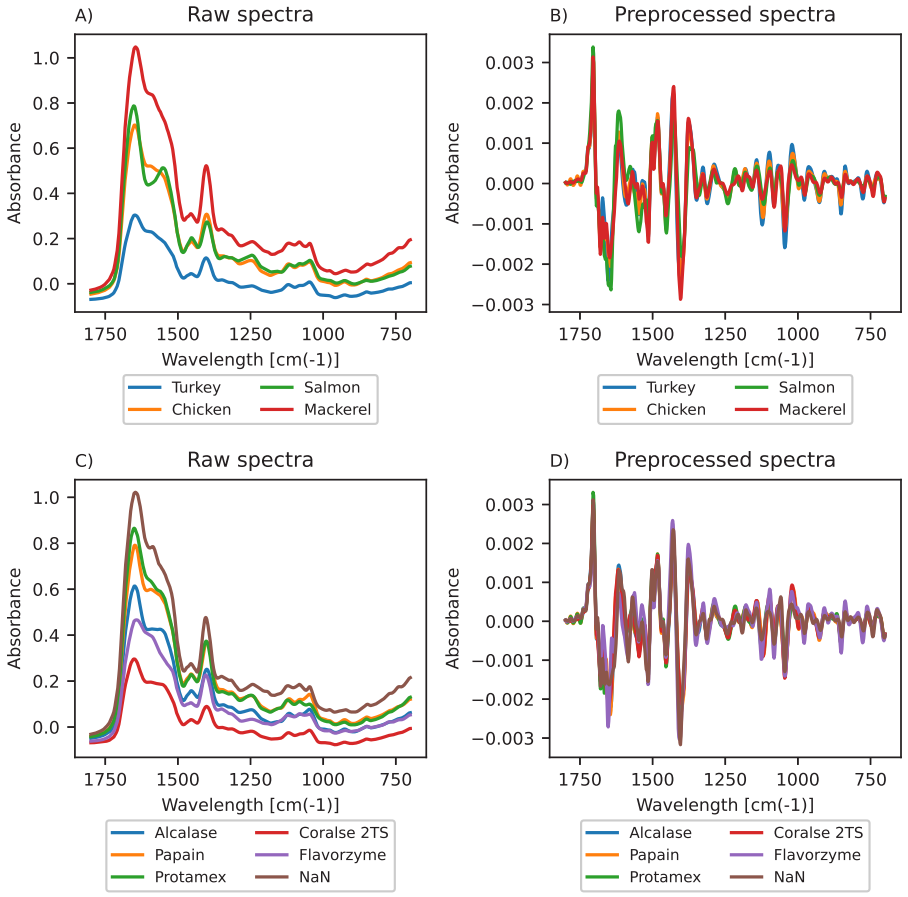


Figure 1: FT-IR spectra of the mean of the samples for each animal and mean of the samples for each enzyme. The mean raw spectra for each animal (A), the mean pre-processed spectra for each animal (B), the mean raw spectra for each enzyme (C) and the mean pre-processed spectra for each enzyme (D).

2.6 Model parameters

The CNN consisted of one convolutional layer with 5 filters and a kernel size of 11, the Exponential Linear Unit (Elu) as the activation function, and one max pooling layer. The RNN consisted of two recurrent layers, the first with return sequences activated on 32 nodes, the second on 16 nodes and activation Elu in the recurrent layers and linear activation in the last dense layer. For CNN and RNN, the networks were trained on 1000 epochs, and the number of epochs used were determined by LOOCV on the training set, for both the global model and the local models. The SVM used a grid search to find the local model parameters from linear, rbf or sigmoid kernel with the regularisation parameter between 0.0001 and 1000000.

3 Results

A global PLSR model was built on the training data from the FT-IR data set, and using LOOCV, the optimal number of components was determined to be 28. The screeplot in Figure 2 shows the first 50 PCs, and illustrates that after 5 PCs, the increase in total explained variance is small. Therefore, an additional restraint that each included component should account for more than 1 % of the explained variance was added. With this constraint, the optimal number of components was determined to be 11, explaining 93.34 % of the variance. A PLSR scoreplot of the first three PC's is shown in Figure 3 where **A** and **C** are coloured by the enzyme used for hydrolysis while **B** and **D** are coloured by the animal the raw material originates from. The scoreplot shows the distributions of samples and how the data is clustered based on prior knowledge. The first three components explain 67.36 % of the variance, however there are no easily separable clusters, either for animal or enzymes.

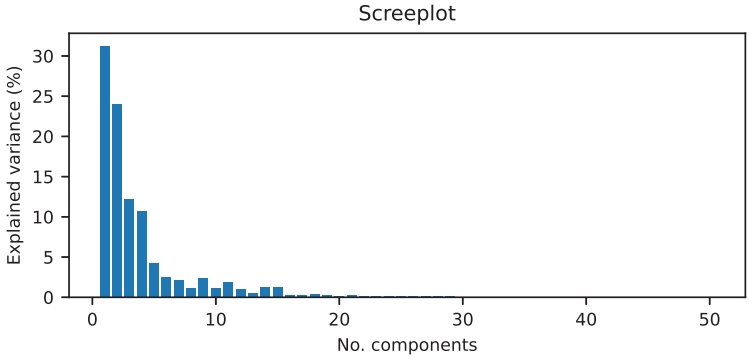


Figure 2: PLSR screeplot of the explained variance for the first 50 PCs

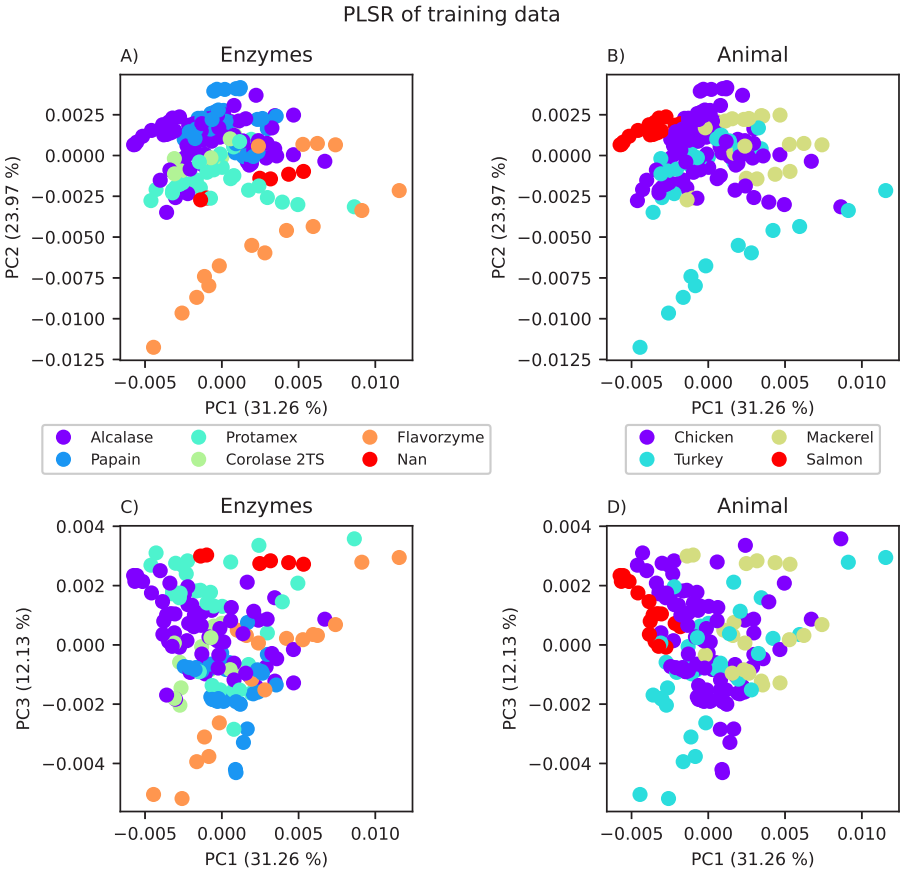


Figure 3: PLSR scoreplot where in the plots to the left (A and C) samples are coloured by hydrolysis enzyme, while in the plots to the right (B and D) samples are coloured by animal of origin. The top plots (A and B) show PC1 against PC2 while the lower plots (C and D) show PC1 against PC3.

3.1 Optimal number of clusters

To identify the optimal number of clusters to use, models with 2-10 clusters were calibrated using the training set. The training data was divided into a calibration set (50 %) and a validation set (50 %) where the calibration set was used to calibrate the local models using LOOCV. The samples in the validation set were then classified and predicted by the respective local models. This was done to validate the classification done by FCM, LDA, QDA and NB. The validation was done by test set validation instead of LOOCV, because of the high computational time and demand of running 166 models 166 times. The results are shown in Figure 4. The optimal number of clusters for each method was determined using the maximum of the mean prediction accuracy (R^2) over the four classification methods and confirmed by visual inspection of Figure 4. The optimal number of clusters was

determined to be 3 for PLSR and 2 for CNN, RNN and SVR as all the four classification methods achieved high R^2 values. For higher numbers of clusters, the classification models are fluctuating and for HC-PLSR and HC-CNN they even show a decreasing trend in classification accuracy. Additionally, with the limited amount of samples available, a high number of clusters will mean few samples in each cluster, something that yields poorer predictions/generalisability. The risk of overfitting clearly increases with an increasing number of clusters.

The distributions of samples in cluster 2-10 for all the clustering methods tested are shown in the supporting information along with the figures determining the optimal number of clusters for HAC and SPC.

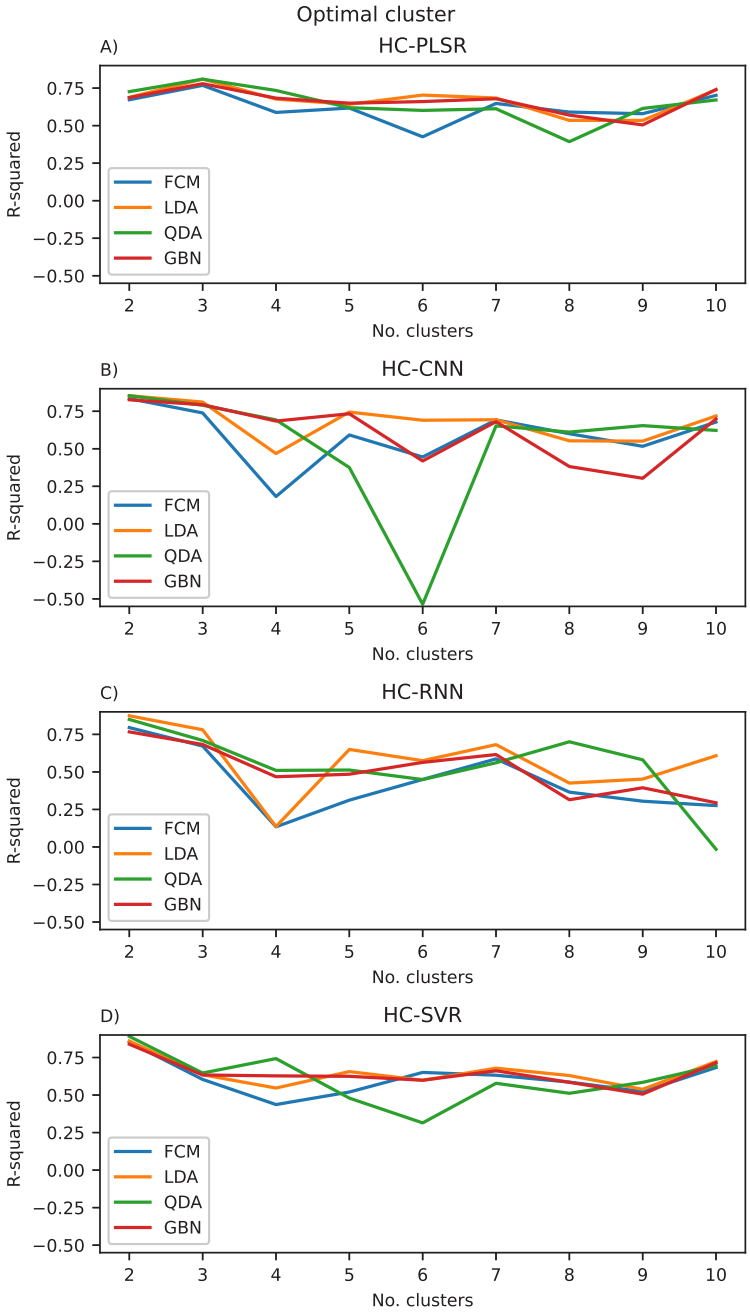


Figure 4: Optimal number of clusters for HC-PLSR (A), HC-CNN (B), HC-RNN (C) and HC-SVR (D) determined by the training set.

To evaluate the properties of the clusters, scoreplots from PLSR with results from FCM using 2 and 3 clusters are shown in Figure 5. For the 2-cluster model, there were 77 samples in cluster 1 and 89 samples in cluster 2. For the 3-cluster model, there were 65 samples in cluster 1, 57 in cluster 2, 44 in cluster 3. However, the clusters are not similar to those given by animal type or enzyme from Figure 3.

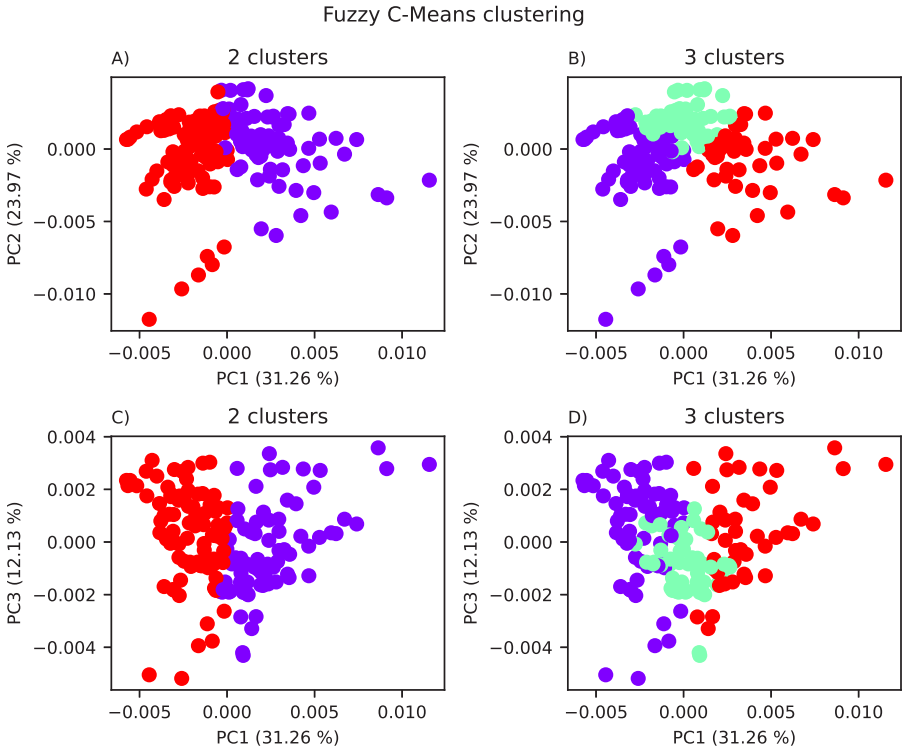


Figure 5: PLSR scoreplot showing the FCM results using the optimal number of clusters; 2 for SVR, CNN and RNN (A and C) and 3 for PLSR (B and D). The top plots show PC1 against PC2 (A and B) while the bottom plots show PC1 against PC3 (C and D).

3.2 Prediction

With the optimal number of clusters determined for HC-PLSR, HC-CNN, HC-RNN and HC-SVR, the local models were trained and applied on the test set. The performance of the models over the four classification methods is shown in Table 1. The results from the FCM clustering were compared to those from HAC and SPC to evaluate whether a different clustering technique could find more informative clusters. The global models were also applied to the test data to evaluate the effect of the local modelling. FCM was unable to classify samples when using HAC and SPC, as the FCM did not have the possibility to train on labels, and was therefore not able to assimilate the clusters from the other two methods.

Model	FCM	LDA	QDA	GNB	# clusters	Global model	Clustering
HC-PLSR	0.696	0.654	0.691	0.804	3	0.797	FCM
HC-CNN	0.748	0.737	0.759	0.732	2	0.795	FCM
HC-RNN	0.810	0.798	0.814	0.818	2	0.836	FCM
HC-SVR	0.850	0.831	0.871	0.812	2	0.871	FCM
HC-PLSR	-	0.593	0.641	0.710	2	0.797	HAC
HC-CNN	-	0.787	0.795	0.786	2	0.795	HAC
HC-RNN	-	0.781	0.779	0.789	2	0.836	HAC
HC-SVR	-	0.858	0.879	0.843	2	0.871	HAC
HC-PLSR	-	0.824	0.844	0.736	4	0.797	SPC
HC-CNN	-	0.752	0.776	0.771	2	0.795	SPC
HC-RNN	-	0.799	0.808	0.823	2	0.836	SPC
HC-SVR	-	0.860	0.871	0.848	2	0.871	SPC

Table 1: R^2-scores for HC-PLSR, HC-CNN, HC-RNN and HC-SVR for their optimal number of clusters using FCM, HAC and SPC.

3.3 Important features

The important features for each of the local modelling methods were visualised as a heatmap and the mean pre-processed spectra for the corresponding cluster were overlayed to simplify the interpretation. For HC-PLSR, the important features were visualised using the loadings for each of the local models, and the results are shown in Figure 6. For each of the local models in HC-CNN and HC-RNN, VarGrad was applied to identify the important features. The results for CNN are shown in Figure 7, where the important features are yellow and the unimportant are blue. For RNN, the results are shown in Figure 8. Lastly, for HC-SVR, the important features were obtained using permutation feature importance for each local model, and the results are shown in Figure 9.

In dry-film FTIR spectra, prominent hydrolysis markers are $\sim$1400 cm^{-1} corresponding to carboxylate (COO^-), 1516 cm^{-1} to ammonia (NH^+_3), $\sim$1550 cm^{-1} to amide II and $\sim$1650 cm^{-1} to amide I [3, 35, 36].

From Figures 7 and 8, it is easy to spot the peak corresponding to the carboxylate-group at $\sim$1400 cm^{-1}, and in the HC-CNNs, this peak has a light yellow colour indicating that it is given weight by the networks in both clusters, but slightly more in cluster 1. For the HC-SVRs, this peak also has a lighter yellow colour for both clusters, and it is lighter and therefore of higher importance in cluster 2. For the HC-PLSR models, this peak is lightest in cluster 3.

The spectra show that the samples with a large effect from the ammonia peak at 1516 cm^{-1} are placed in cluster 1, and in HC-CNN and HC-RNN, this peak is deemed important. In HC-SVR, this peak is of higher importance in cluster 2. In HC-PLSR these samples are placed in cluster 3, which is also where this peak is most important.

The amide II groups at $\sim$1550 cm^{-1}, show importance in cluster 1 for all methods and in cluster 2 for HC-SVR and HC-PLSR. Lastly, the samples with a larger peak at $\sim$1650 cm^{-1}, are placed in cluster 1, and this peak seems to have a higher importance in cluster 1 compared to cluster 2 for HC-CNN, HC-RNN and HC-SVM, while for HC-PLSR this

peak is important in all three clusters.

The models seem to find a peak of importance at $\sim$1350 cm^{-1} and at $\sim$1100 cm^{-1}, and cluster 2 in HC-CNN shows a strong importance of a peak at $\sim$900 $^{-1}$. $\sim$1350 cm^{-1} could possibly correspond to nitro groups, $\sim$900 cm^{-1} could be aromatic rings and $\sim$1100 cm^{-1} could be ether absorption [37]. All could be heteroatom subtitutions (for example oxygen, sulphur or nitrogen groups) as well. These three areas are in the fingerprint region, between 1400-700 cm^{-1}, which usually contains a large number of peaks, which is also noticeable by how the local models all highlight different parts of this area. Additionally, peaks in the fingerprint region often correspond to asymmetric stretching vibrations which also could explain the peaks at $\sim$1350, $\sim$1100 and $\sim$900 $^{-1}$.

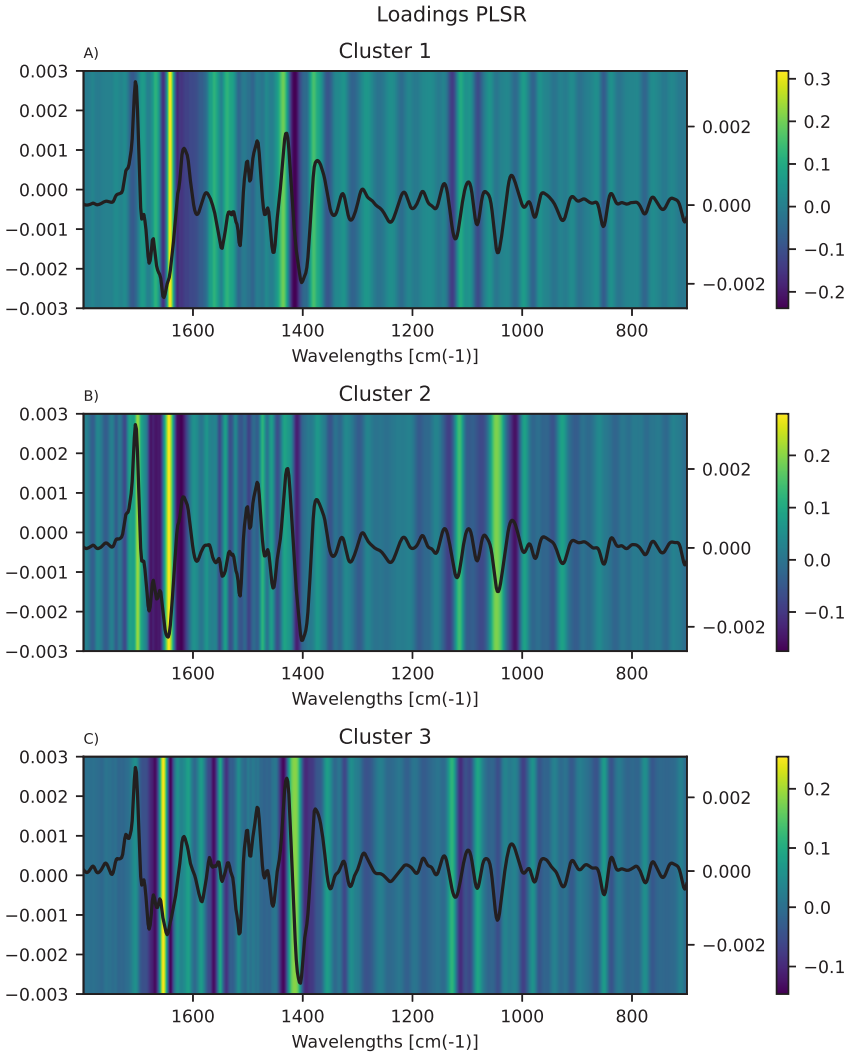


Figure 6: Visualisation of feature importance for the HC-PLSR with three clusters, with the mean spectra of the samples in the corresponding cluster overlayed. Important features are coloured yellow and unimportant are blue.

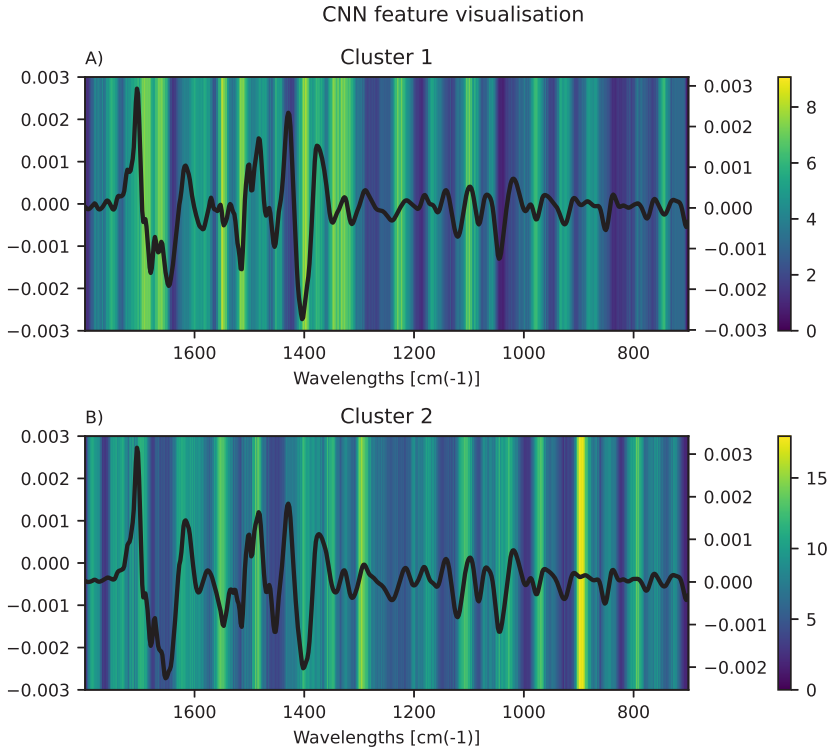


Figure 7: Visualisation of feature importance for the HC-CNN with two clusters, cluster 1 in A and cluster 2 in B, with the mean spectra of the samples in the corresponding cluster overlayed. Important features are coloured yellow and unimportant are blue.

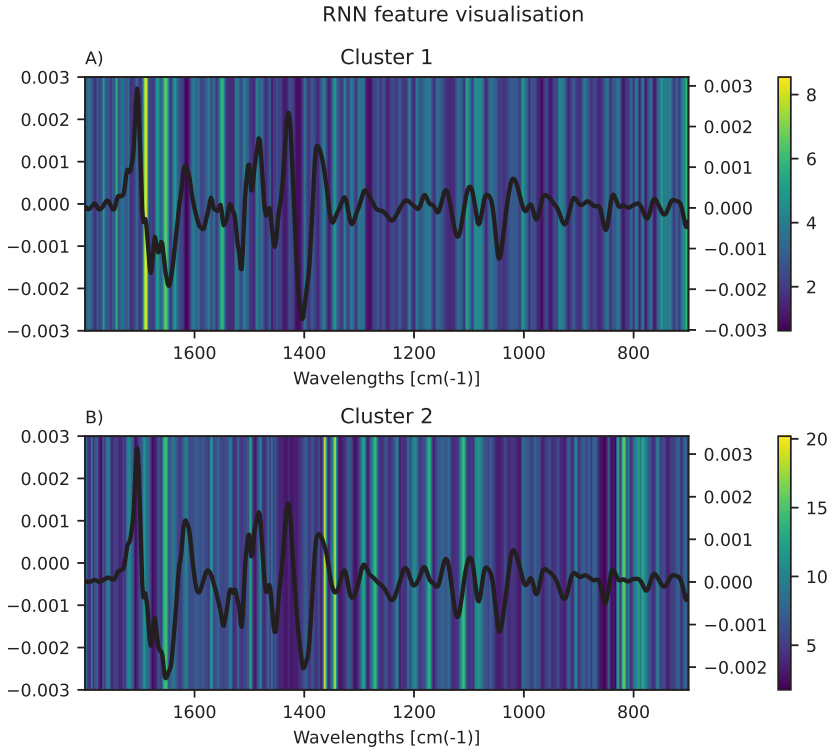


Figure 8: Visualisation of reature importance for the HC-RNN with two clusters, cluster 1 in A and cluster 2 in B, with the mean spectra of the samples in the corresponding cluster overlayed. Important features are coloured yellow and unimportant are blue.

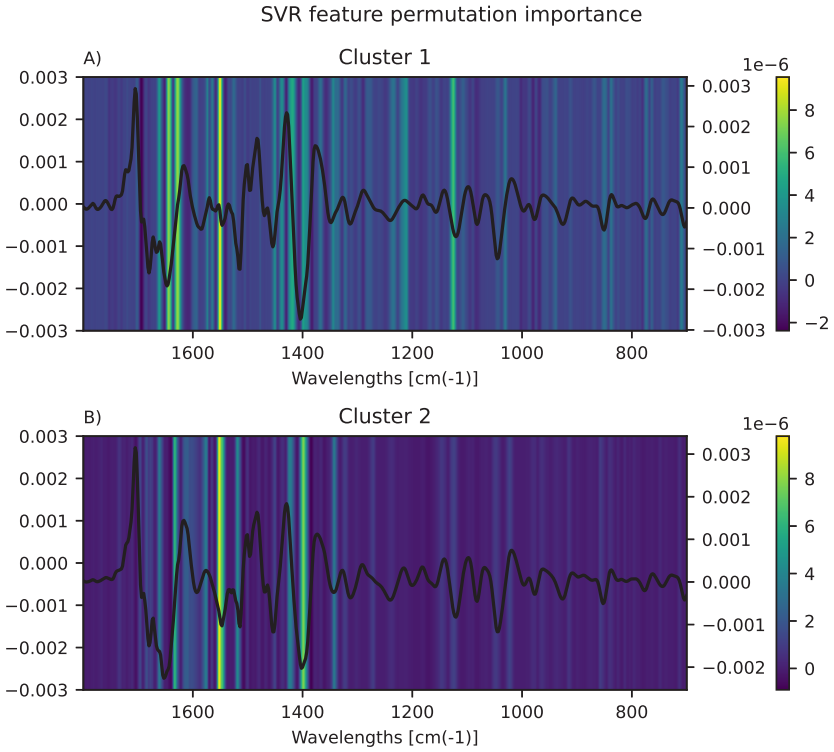


Figure 9: Visualisation of feature importance for the HC-SVR with two clusters, cluster 1 in A and cluster 2 in B, with the mean spectra of the samples in the corresponding cluster overlayed. Important features are coloured yellow and unimportant are blue.

4 Discussion

In this study the previously established HC-PLSR model was expanded into deep learning and SVRs for non-linear predictions of heterogeneous data. The methods were tested on an FT-IR data set of raw materials from poultry and fish. The results showed that HC-CNN, HC-RNN and HC-SVR yielded better predictions than HC-PLSR. This indicates that the linearity in PLSR represents a disadvantage as the data set used here probably is not locally linear in structure, and hence HC-PLSR is not performing as well as the deep learning-based hierarchical methods. HC-PLSR also needs a higher number of clusters than the other methods to handle non-linearities in the data, since this lies intrinsic in the other methods.

The different classification methods used to assign labels to the test set samples yielded varying prediction results. Over all the methods, QDA has the most unstable predictions and particularly when using 6 clusters. The difference between LDA and QDA is that in LDA, it is assumed that each class share the same covariance matrix, while QDA has

no such assumption. This makes LDA simpler with fewer parameters to determine, and therefore when there are few samples in the data set, which is the case here, LDA tends to perform better than QDA. For all classification methods, HC-RNN achieved a slight increase in R^2 compared to the global models.

The visualisation of the feature importance shows that there are differences when it comes to which samples that are assigned to the various clusters. Additionally, the local models evaluate different features as important between the clusters. However, the clusters determined by the clustering methods used here show no similarities to those based on prior knowledge.

Local modelling also allows for simpler models, models with a low number of PCs, few convolutional layers etc. to predict with high accuracy, while simultaneously easing interpretation. The best prediction results for both the global CNN and RNN models is achieved with 3 layers, which is higher than what was found optimal for the local models. However, in order to avoid overfitting, the number of clusters used should be kept relatively low. Additionaly, as the number of clusters increases, the number of empty clusters also increases since clusters with less than 10 samples are removed. This can result in more unstable predictions. The data set used in this study has a limited number of samples and it is therefore an advantage to keep the number of clusters used low. However, the prediction methods can handle unlimited numbers of clusters, the limitations only lie in the number of samples. In HC-PLSR, using a low number of PCs is beneficial for the clustering, creating simpler clusters and subsequent models. With either method, one should strive to keep the model complexity as low as possible, but without sacrificing prediction accuracy, both to limit the risk of overfitting and to ease interpretation.

For the deep learning models, a disadvantage is that the data set has to be large enough for training and validation. Small data sets can result in unstable networks, especially if the number of features is substantially larger than the number of samples. As the parameters for the global model and for the local models are determined using LOOCV, the training of the models is time consuming. The training data used here contains 166 samples which therefore also is the number of models needing to be built. However, once the models are trained, prediction of the response for new samples is fast.

Expansion of the HC workflow to deep learning has the advantage of being able to handle any kind of data. In data which consists of chemical spectra, different parts of the spectra are often connected by for instance homologeous series, isotope peaks, adjacent groups or chemical bonds, which are essential for the identification and analysis of chemical structures. Creating models which can easily detect these patterns even in data sets with large differences between the samples, is of great value. The comparison of the local models to the global models can illuminate structures in the data which the global model is not able to capture. The local models can also be used to identify problem areas in the data if any of the clusters yield significantly poorer prediction abilities than the remaining clusters.

The local modelling procedure described in this paper is fully automatic, i.e. no prior information about groups in the samples is required. This is an advantage when the structures in the data are unknown. Additionally, our results show that the models achieve a high prediction accuracy even without using prior knowledge.

Acknowledgement

The authors thank The Research Council of Norway, Equinor ASA, OMV (Norge) AS, Wintershall DEA Norge AS and TotalEnergies for funding. This work is a part of the Knowledge-Building Project for Industry (PETROMAKS 2), Project number: 294636 "New Hydrate Management: New understanding of hydrate phenomena in oil systems to enable safe operation within the hydrate zone".

References

[1] L. Eriksson, J. Trygg, and S. Wold, "PLS-trees®, a top-down clustering approach," *Journal of Chemometrics*, vol. 23, pp. 569–580, Nov. 2009.

[2] L. Eriksson, M. Toft, E. Johansson, S. Wold, and J. Trygg, "Separating Y-predictive and Y-orthogonal variation in multi-block spectral data," *Journal of Chemometrics*, vol. 20, pp. 352–361, Oct. 2006.

[3] K. A. Kristoffersen, K. H. Liland, U. Böcker, S. G. Wubshet, D. Lindberg, S. J. Horn, and N. K. Afseth, "FTIR-based hierarchical modeling for prediction of average molecular weights of protein hydrolysates," *Talanta*, vol. 205, p. 120084, Dec. 2019.

[4] A. Lindström, F. Pettersson, F. Almqvist, A. Berglund, J. Kihlberg, and A. Linusson, "Hierarchical PLS Modeling for Predicting the Binding of a Comprehensive Set of Structurally Diverse Protein-Ligand Complexes," *Journal of Chemical Information and Modeling*, vol. 46, pp. 1154–1167, Apr. 2006.

[5] S. Wold, N. Kettaneh, and K. Tjessem, "Hierarchical multiblock PLS and PC models for easier model interpretation and as an alternative to variable selection," *Journal of Chemometrics*, vol. 10, pp. 463–482, Dec. 1996.

[6] S. Wold, A. Berglund, and N. Kettaneh, "New and old trends in chemometrics. How to deal with the increasing data volumes in R&D&P (research, development and production)—with examples from pharmaceutical research and process modeling," *Journal of Chemometrics: Special Issue: Proceedings of the 7th Scandinavian Symposium on Chemometrics*, vol. 16, pp. 377–386, Oct. 2002.

[7] M. Bevilacqua and F. Marini, "Local classification: Locally weighted–partial least squares-discriminant analysis (LW–PLS-DA)," *Analytica Chimica Acta*, vol. 838, pp. 20–30, Aug. 2014.

[8] K. Tøndel, U. G. Indahl, A. B. Gjuvsland, J. O. Vik, P. Hunter, S. W. Omholt, and H. Martens, "Hierarchical Cluster-based Partial Least Squares Regression (HC-PLSR) is an efficient tool for metamodelling of nonlinear dynamic models," *BMC Systems Biology*, vol. 5, p. Article 90, June 2011.

[9] K. Tøndel, U. G. Indahl, A. B. Gjuvsland, S. W. Omholt, and H. Martens, "Multi-way metamodelling facilitates insight into the complex input-output maps of nonlinear dynamic models," *BMC Systems Biology*, vol. 6, p. 88, July 2012.

[10] I. Gath and A.B. Geva, "Unsupervised optimal fuzzy clustering," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 11, pp. 773–780, July 1989.

[11] H. Frigui and R. Krishnapuram, "A robust competitive clustering algorithm with applications in computer vision," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 21, pp. 450–465, May 1999.

[12] S. Wold, H. Martens, and H. Wold, "The multivariate calibration problem in chemistry solved by the PLS method," in *Matrix Pencils*, no. 973 in Lecture Notes in Mathematics, pp. 286–293, Berlin, Heidelberg: Springer, 1983.

[13] S. Wold, A. Ruhe, H. Wold, and W. J. Dunn III, "The Collinearity Problem in Linear Regression. The Partial Least Squares (PLS) Approach to Generalized Inverses," *SIAM Journal on Scientific and Statistical Computing*, vol. 5, pp. 735–743, Sept. 1984.

[14] H. Martens and T. Næs, *Multivariate Calibration*. Chichester: Wiley, 1 ed., July 1992.

[15] S. Wold, M. Sjöström, and L. Eriksson, "PLS-regression: a basic tool of chemometrics," *Chemometrics and Intelligent Laboratory Systems*, vol. 58, pp. 109–130, Oct. 2001.

[16] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, pp. 2278–2324, Nov. 1998.

[17] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, pp. 436–444, May 2015.

[18] D. P. Mandic and J. A. Chambers, *Recurrent Neural Networks for Prediction: Learning Algorithms, Architectures and Stability*. Wiley, Aug. 2001.

[19] I. H. Sarker, "Deep Learning: A Comprehensive Overview on Techniques, Taxonomy, Applications and Research Directions," *SN Computer Science*, vol. 2, p. 420, Aug. 2021.

[20] H. Drucker, C. J. C. Burges, L. Kaufman, A. Smola, and V. Vapnik, "Support vector regression machines," in *Advances in Neural Information Processing Systems 9*, Cambridge, MA: MIT Press, 1997.

[21] V. Vapnik, G. S., and S. A., "Support vector method for function approximation, regression estimation, and signal processing," in *Advances in Neural Information Process Systems 9*, pp. 281–287, Cambridge, MA: MIT Press, 1 ed., 1997.

[22] C. J. Burges, "A Tutorial on Support Vector Machines for Pattern recognition," *Data Mining and Knowledge Discovery*, vol. 2, pp. 121–167, June 1998.

[23] J. C. Dunn, "A Fuzzy Relative of the ISODATA Process and Its Use in Detecting Compact Well-Separated Clusters," *Journal of Cybernetics*, vol. 3, pp. 32–57, Sept. 1973.

[24] J. C. Bezdek, *Pattern Recognition with Fuzzy Objective Function Algorithms*. Advanced Applications in Pattern Recognition, Springer US, 1 ed., 1981.

[25] U. von Luxburg, "A tutorial on spectral clustering," *Statistics and Computing*, vol. 17, pp. 395–416, Aug. 2007.

[26] F. Nielsen, "Hierarchical Clustering," in *Introduction to HPC with MPI for Data Science*, Undergraduate Topics in Computer Science, pp. 195–211, Springer, 1 ed., Feb. 2016.

[27] G. J. McLachlan, *Discriminant Analysis and Statistical Pattern Recognition*. Wiley Series in Probability and Statistics, Wiley-Interscience, 1 ed., Mar. 1992.

[28] T. Hastie, J. Friedman, and R. Tibshirani, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer Series in Statistics, New York, NY: Springer, 1 ed., 2001.

[29] M. L. D. Dias, "fuzzy-c-means: An implementation of Fuzzy C-means clustering algorithm.," May 2019.

[30] J. Adebayo, J. Gilmer, I. Goodfellow, and B. Kim, "Local Explanation Methods for Deep Neural Networks Lack Sensitivity to Parameter Values," *arXiv*, p. 1810.03307, Oct. 2018.

[31] J. Adebayo, J. Gilmer, M. Muelly, I. Goodfellow, M. Hardt, and B. Kim, "Sanity Checks for Saliency Maps," in *Advances in Neural Information Processing Systems 31*, (Montréal, Canada), pp. 9505–9515, 2018.

[32] S. Hooker, D. Erhan, P.-J. Kindermans, and B. Kim, "A Benchmark for Interpretability Methods in Deep Neural Networks," in *Advances in Neural Information Processing Systems*, vol. 32, (Vancouver, Canada), Curran Associates, Inc., 2019.

[33] A. Jenul, S. Schrunner, B. N. Huynh, R. Helin, C. M. Futsæther, K. H. Liland, and O. Tomic, "Ranking Feature-Block Importance in Artificial Multiblock Neural Networks," in *Artificial Neural Networks and Machine Learning*, vol. 13532 of *Lecture Notes in Computer Science*, (Bristol, UK), Springer, Cham, Sept. 2022.

[34] L. Breiman, "Random Forests," *Machine Learning*, vol. 45, pp. 5–32, Oct. 2001.

[35] U. Böcker, S. G. Wubshet, D. Lindberg, and N. K. Afseth, "Fourier-transform infrared spectroscopy for characterization of protein chain reductions in enzymatic reactions," *Analyst*, vol. 142, no. 15, pp. 2812–2818, 2017.

[36] S. G. Wubshet, I. Måge, U. Böcker, D. Lindberg, S. H. Knutsen, A. Rieder, D. A. Rodriguez, and N. K. Afseth, "Fourier-transform infrared spectroscopy for characterization of protein chain reductions in enzymatic reactions," *Analytical Methods*, vol. 9, pp. 4247–4254, June 2017.

[37] D. Williams and I. Fleming, *Spectroscopic methods in organic chemistry*. UK: McGraw-Hill Education, 6 ed., 2008.

ISBN: 978-82-575-2063-2
ISSN: 1894-6402

Norwegian University
of Life Sciences

Postboks 5003
NO-1432 Ås, Norway
+47 67 23 00 00
www.nmbu.no

116531 / ANDVORDGRAFISK.NO