UNIVERSITY OF THE
WITWATERSRAND,
JOHANNESBURG

Towards a Robust, Universal Predictor of Gas Hydrate Equilibria through the means of a Deep Learning Regression

Research Report

Prepared by

M. K. B. Landgrebe (Student Number: 704140)

A research report submitted to the Faculty of Engineering and Built Environment, University of the Witwatersrand, Johannesburg, South Africa, in partial fulfilment of the requirements for the Degree of Master of Science in Engineering.

Supervisor: Dr. DB Nkazi

October, 2019

Declaration

I, M. Landgrebe (Student Number: 704140) student registered for *Master of Science in Engineering* (Chemical) in the year 2018 - 2019. I hereby declare the following:

- I am aware that plagiarism (the use of someone else's work without their permission and/or without acknowledging the original source) is wrong.
- I confirm that the work submitted for assessment for the above course is my own unaided work except where I have explicitly indicated otherwise.
- I have followed the required conventions in referencing the thoughts and ideas of others.
- I understand that the University of the Witwatersrand may take disciplinary action against me if there is a belief that this is not my own unaided work or that I have failed to acknowledge the source of the ideas or words in my writing.

Signature:  Date: October 15, 2019

Acknowledgements

My thanks to my research supervisor, Dr. D. Nkazi for his continuous input, feedback and guidance throughout my research, for the assistance with administrative matters, for encouraging me to pursue such a fascinating topic, and for the time taken to ensure my research was a success

Acknowledgements to the University of the Witwatersrand, the school of Chemical and Metallurgical Engineering, and the staff of the Wartenweiler and Chamber of Mines Engineering libraries for their assistance in administrative matters, and research material requests.

Special thanks to Dr. T.C.W. Landgrebe for the guidance, feedback and valuable insight into the machine learning and data gathering aspects of this research, and for recommending the software which facilitated the development of my model.

Further thanks to the Python and Keras development teams, whose software greatly simplified the procedure of model development, and made my venture into machine learning a satisfying experience.

Abstract

Gas hydrate equilibria of natural gas mixtures has proven to be a highly non-linear, multimodal phenomenon, and extensive investment has been made over decades in order to understand and accurately predict natural gas hydrate equilibrium conditions. While most models applied toward predicting gas hydrate equilibria industrially are computerised thermodynamic models based on intrinsic molecular behaviour, these approaches are often limited in their capability to predict actual phenomena over a wide range of conditions due to the high degrees of non-linearity and complexities resulting from other factors which prove difficult to model explicitly. In this research, an artificial neural network was developed using publicly available experimental gas hydrate equilibrium data. A regression was achieved by means of a deep-learning multi-layer perceptron consisting of three hidden layers with a high neuron count, and an output layer comprised of a single neuron, corresponding with the predicted equilibrium pressure. 9 model features are present in the input layer, consisting of the temperature and the molar fractions of methane, ethane, propane, iso-butane, n-butane, carbon dioxide, nitrogen and a lumped fraction of organic molecules consisting of at least five carbon atoms. Models have been evaluated according to the ability to predict a wide range of data, multicomponent prediction accuracy, and dependency on individual sources of data. 670 multicomponent experimental equilibrium data samples have been obtained from literature. Due to the limited amount of multicomponent equilibrium data published, the incorporation of pure and binary methane mixtures into a second dataset including multicomponent data has proven imperative to achieve the best possible model. The complete dataset consists of 1209 equilibrium data samples. To ensure multicomponent data is accurately modelled, several models have been developed using both datasets to prove that the pure and binary inclusive dataset models do not simply inflate results through inclusion of easily predicted data. Regression scoring was assessed using the coefficient of determination, the R^2 score. Cross-validation and hold-out validation have been employed in conjunction to assess the model's ability to predict unseen data, while facilitating parameter optimization and yielding the bias and variance associated with the model. Cross-validation has been implemented by means of 10-fold validation, with a randomized 70%-30% train-test split performed to determine the test indices for each fold. Hold-out validation has been achieved by means of a 10% stratified-split, whereby the proportion of data from each independent source is held approximately constant across training and hold-out validation sets with the purpose of ensuring a wide range of conditions are tested. A cross validation R^2 score of 0.9860 is achieved with a standard deviation of 0.0035. Hold-out validation yields an R^2 of 0.9926. Results indicate a sufficiently accurate model has been achieved with a low enough variance to consider the model universal over the range of equilibrium data included in this investigation. The dependency on individual experimental data sources is of concern due to the limited amount of multicomponent equilibrium data available, and the age of equilibrium measurement practices for many sources and time frames associated with hydrate equilibrium measurements. However, the inclusion of pure methane and methane binary compounds does assist in reducing the susceptibility of the model to these errors. Dependency on individual data sources has been assessed by means of grouped cross-validation being performed on neural network models. Grouping results do indicate a lack of independently obtained data covering certain ranges of conditions, however binary inclusive models are shown to present a damping effect on the magnitude of experimental or measurement error on the model at large. Due to a lack of independent experimental studies covering a wide range of conditions, hydrogen sulphide could not be included as a feature in model development. As such, the developed model is noted to be applicable to sweet natural gas flow systems, where hydrate structures I or II are exhibited.

Table of Contents

List of Figures

List of Tables

Nomenclature

Complete Dataset: Dataset including all equilibrium points sampled from literature, including pure methane and methane binary mixtures

CCGT: Combined Cycle Gas Turbine

GUMP: Gas Utilization Master Plan

LDHI: Low Dosage Hydrate Inhibitor

LNG: Liquefied Natural Gas

MPa: Pressure Unit - Mega Pascals

Multicomponent Exclusive Dataset: Dataset limited to samples from experimental studies where at least three gas-phase components (excluding water) are present

ReLU: Rectified Linear Unit

R^2: Coefficient of Determination

σ: Standard Deviation

sI: Structure I Gas Hydrates

sII: Structure II Gas Hydrates

sH: Structure H Gas Hydrates

SGD: Stochastic Gradient Descent

CHAPTER 1: BACKGROUND AND MOTIVATION

1.1 The Role of Natural Gas in South Africa

Some may question the viability of research in the field of natural gas in a world where increasing focus is being placed upon meeting the energy demands of nations through means of green energy. This concern is valid, with the threats of climate change becoming increasingly prevalent, conducting research into hydrocarbon extraction and production seems redundant due to the burning of natural gas leading to more carbon emissions. This argument however, does not account for the immediate energy grid strain affecting developing economies such as South Africa. While the goal is to achieve a significant proportion of energy being generated by renewable means in South Africa by 2030 (IRP, 2010), the time-scale associated with these projects and current strain on the energy grid make large scale non-renewable projects with a high power output worth examining. Investment in renewables such as solar and wind are capital intensive and can thus take many years to replace what a large coal power station could supply. As such, there is a demand for a means to rapidly supply energy while the grid shifts toward a renewable base without producing carbon emissions at a similar magnitude to coal combustion. While natural gas still emits carbon dioxide in addition to other atmospheric pollutants, combustion of natural gas produces 50 to 60% of the carbon dioxide produced by coal, and emits significantly less sulphur and particulate matter (Union of Concerned Scientists, 2019). Combined with heat recovery systems such as those of Combined Cycle Gas Turbines (CCGT), which boast high thermal efficiencies through combining the generation of energy through hydrocarbon combustion with a heat recovery loop which generates steam from waste heat to run a separate turbine, natural gas can be seen that a highly efficient means of generating power compared to coal. Gas turbines additionally provide the great benefit of flexibility, whereby the unit may be activated solely during peak demand periods, then shut off once more (U.S. Energy Information Administration, 2013). This factor proves essential to countries with a struggling power grid, such as South Africa, which instead currently fires carbon unfriendly diesel turbines and energy intensive pumped storage schemes to supplement peak demand. Finally, an additional benefit of investing in gas infrastructure is the potential for a gas market, serving to lower energy consumption while developing a substantial industry. Further details on the strategic planning for a gas infrastructure in South Africa is detailed in the Gas Utilization Master Plan (GUMP) (Strategic Energy Plan 2015-2020). Through domestic use of gas, electricity consumption is reduced. While carbon emissions are still produced, they are significantly lower than that of the coal plant generating

power. These factors illustrate the utility of natural gas serving as a means to transition an energy grid towards renewable energy, developing a subsequent industry, while still meeting energy demands and lowering carbon dioxide emissions. As such, gas hydrate prediction, a significant challenge of natural gas production and distribution will now be discussed, and this research aims to present a solution that is based on experimental data, rather than intrinsic statistical thermodynamics.

1.2 Introduction

Gas hydrates, or clathrate hydrates, are solid, crystalline masses which form as a result of a guest molecule stabilizing a water lattice. These hydrates will form when sufficient water and an appropriate guest molecule are present at the conditions of system temperature being lower than the dew point of the water, and the pressure of the system being high enough for hydrate formation to be initiated (GPSA, 2004). The water lattice may be stabilized by a guest molecule which may be a low carbon number organic molecule or an inorganic gas with a small molecular diameter. While the hydrates may appear crystalline in nature, the structure of gas hydrates varies considerably from ice formed by pure water. The structure by which hydrate lattices arrange themselves may differ depending on the conditions and composition of the gas phase, and several different structures have been identified. While several structures have been identified for certain hydrates, natural gas hydrates have been found to conform to three distinct hydrate structures: sI, sII & sH (Tohidi et al., 2001). The respective components of natural gas additionally do not form other structures in isolation when under normal conditions. These structures are achieved through hydrogen bonds between the encaging water molecules, and are stabilized through interactions between guest and encaging water molecules (Hawtin et al., 2008; Rodger, 1990).

The formation of gas hydrates is of considerable interest in the chemical and petroleum industry, as the formation of gas hydrates often results in flow interruptions leading to additional costs of operation, damage to equipment and possibly leading to hazardous conditions for personnel. As such, a significant amount of research has been directed towards the flow assurance problems resulting from gas hydrates, notably towards predicting the conditions at which these hydrates are likely stable (Shahnazar & Hasan, 2014). Gas hydrates are well documented as forming in transit pipelines of reservoir fluids, including both liquid and gas hydrocarbon flows. The conditions at which hydrates are most likely to be encountered

at high pressure and low temperature conditions (Rajnauth et al., 2012), which are often encountered when dealing with offshore wells and pipelines (Makogon, 1997).

The structure of gas hydrates is responsible for much of the complexities associated with modelling gas hydrate phenomena. Natural gases have been identified as assuming three distinct hydrate structures: the cubic sI, and sII structures, and the hexagonal sH (Sloan, 1998). These different structures exhibit different numbers of cages present, and the sizes of these cages. The size of cages plays a vital role in determining which guest molecules may stabilize the structure, for instance while the largest cage diameter of sI hydrates is too small to accommodate molecules such as iso-butane, iso-butane forms sII hydrates, being present in the large cages of this structure (Sloan & Koh, 2007). The difference between hydrate structures is significant from a thermodynamic standpoint, as different hydrates may form at different conditions. GPSA (2004) attributes the hydrate structure as significantly impacting the hydrate formation temperature and pressure. As such, any thermodynamic models need to account for the hydrate structure, which poses a significant challenge considering natural gas forms three different structures. The systems being examined in this research focus on the formation of gas hydrates from reservoir fluids in production tubing, flow-lines to surface facilities and pipelines. Based on these constraints, it is possible to further narrow down the range of hydrate structures relevant to an equilibrium condition model operating the expected conditions of these flow-lines. Tohidi et al. (2001) investigated the occurrence of structure H hydrates in reservoir fluids including natural gas, and found that structure H hydrates are unlikely to form when operating outside the hydrate stability region, while identifying structure II hydrates as the dominant structure in reservoir fluid hydrate formation. Tohidi et al. (2001) further claims that sI hydrates are more likely to be stable than sH hydrates. GPSA (2004) notes that gas mixtures are likely to form sII hydrates. Natural gas mixtures are more likely to form sII hydrates as molecules above a certain diameter only can fit into the large cages of sII hydrates, and n-butane being too large to even form sII hydrates on its own, requiring a smaller molecule such as methane to assist in fitting into the large cage of sII (Sloan & Koh, 2007). Overall it is thus possible to conclude that the model developed to predict equilibrium conditions for natural gas flow within flow lines at normal operating ranges need only consider gas mixtures and conditions forming sI and sII hydrates. Note that structure H natural gas hydrates may be encountered outside of transit lines and reservoirs, and thus further consideration is required when examining naturally occurring hydrates or systems under abnormally high pressures.

Furthermore, as will be discussed, operating well within the hydrate stability zone may lead to structure H hydrate formation (Tohidi et al., 2001).

Hydrate formation is often considered an unwanted process, however due to the volatile and highly flammable nature of natural gas, hydrates have been viewed as a safer means of storing and transporting natural gas, or as an energy source when considering naturally occurring hydrates offshore or in reservoirs. Veluswamy et al. (2018) notes artificially solidified natural gas hydrates as the safest means of storing and transporting natural gas, due to lowered risk of ignition or explosion, and thus additionally serves to increase the attractiveness of clathrate hydrates for long term storage of natural gas. Converting natural gas into hydrates would prove particularly useful for offshore gas production which occurs far from the shore or from a pipeline, thus allowing safer storage of natural gas on platform, and safer transportation of natural gas to its destination. This approach would contrast current Liquefied Natural Gas (LNG) approaches which seek to transport natural gas as a liquid using specifically designed tankers and storage facilities, by instead converting natural gas into solid hydrate crystals. While technological limitations have largely prevented significant application of converting natural gases into hydrates for transportation and storage, increasing interest has arisen over the potential use of naturally occurring gas hydrates and artificially produced hydrates for storing and transporting natural gas, partially driven due to several countries attempting to reduce carbon emissions or increase energy production by adopting gas as a cleaner fossil fuel and increasing the share of gas in the generation of electricity (Chong et al., 2016; Veluswamy et al., 2018). While a considerable number of publications have been conducted on the economic viability of implementing large-scale gas hydrate conversion, transportation and storage, Veluswamy et al. (2018) in a review reports that a significant degree of conflicting studies are present, and that further investigation into the economic viability of storing and transporting hydrates as sII rather than sI hydrates is required, as most studies are based on pure methane sI hydrates rather than natural gas sII hydrates. Naturally occurring hydrates found on the sea-floor or within reservoirs have been identified as an extremely large potential source of carbon-based energy (Chong et al., 2016). Makogon et al. (2007) details that naturally occurring gas hydrates can be commercially exploited to yield significantly more energy than consumed during extraction, pressurization and transportation depending on the composition of the hydrate and hydrate concentration within the bed being developed, thus highlighting the potential of naturally occurring gas hydrates as an energy source. Much research has been conducted on the properties, geographic distribution and potential for use as an energy source

of naturally occurring hydrates. Chong et al. (2016) has conducted a recent review of the state of naturally occurring clathrate hydrate developments.

1.3 Effect of Hydrates and Safety Concerns

The presence of hydrate plugs in a pipeline poses serious risks to operational and potentially personnel safety. These plugs can completely block flow through a pipeline, with the subsequent blockage halting production and requiring a pipeline shutdown to remove the plug, thus resulting in financial losses (Perrin et al., 2013). In addition to production losses, halting the flow of a pipeline can further compound hydrate problems and requires immediate removal of the plug, which can in itself pose serious hazards. Common means of plug removal include depressurizing both ends of the pipeline so as to dissociate hydrates (Austvik et al., 2000). During depressurization, as hydrate plugs begin to dissociate from the outer radius, they may separate from the pipe wall and travel downstream as a solid mass of significant density, thus potentially serving as a projectile which can rupture the pipeline when striking obstructions or bends in piping, or cause a rupture downstream due to dense gas pockets resulting from the plug's momentum (Sloan & Koh, 2007). Other techniques for removal of hydrates have been developed such as chemical injection and external heating to dissolve hydrates (Austvik et al., 2000; Turner et al., 2014), however these are also subject to various challenges. Thus, caution must be exercised when removing hydrate plugs, which can take extended periods of time and can incur significant costs. Due to the difficulties associated with removing hydrate plugs, avoiding the presence of hydrates entirely or significantly slowing the rate at which hydrate plugs form has become a highly popular means of managing gas hydrates, particularly for deep offshore wells where operating conditions are well within the hydrate stability zone.

1.4 Hydrate Formation Mechanism

The formation of hydrates is prevalent throughout the production of offshore reservoirs, affecting the production well, surface equipment, transit lines from well to surface and pipelines. Within reservoirs undergoing production or drilling, hydrates are known to form in oil wells as well as gas wells through water, which enters the production zone during drilling (Makogon, 1997). As most shallow reservoirs with favourable depth and temperatures have been produced, offshore production is trending towards deeper wells and arctic reservoirs where adverse conditions leading to hydrate formation are likely to be encountered (Makogon, 1997; Sloan & Koh, 2007). As such, there is an increased likelihood that the presence of gas hydrates will be a significant factor when considering the profitability of a potential reservoir.

This increased prevalence of hydrates has led to the need to establish effective means of controlling the rate at which hydrates form, or preventing the formation entirely.

While the exact mechanism for how hydrate nucleation and growth occur is still debated, with competing nucleation hypotheses discussed in Perrin et al. (2013), it is well established that in order for hydrate masses to form, hydrate crystal nucleation and growth must occur (Perrin et al., 2013). Following nucleation and growth, hydrate plugs may form through the agglomeration of smaller hydrate particles (Hawtin et al., 2008, Zerpa, 2013). As the focus of this research is the formation of hydrates in flow lines from natural gas wells, the case of hydrate formation in gas dominated pipelines is examined. While a detailed description for hydrate formation for both gas-water and multiphase systems may be found in Zerpa (2013), a brief description of hydrate formation in gas flow pipelines is provided to facilitate a discussion on the types of hydrate inhibition commonly applied in industry. Hydrates may initially nucleate and grow either at the pipe wall if the temperature at the wall is lower than that of the bulk fluid, and where defects in the surface of the wall lead to sites promoting hydrate formation (Bassani et al., 2017). The likelihood of the wall temperature is lower than that of the bulk flow is especially plausible for deep water pipelines, where line burial and insulation may prove insufficient due to low ambient temperatures, thus leading to large differentials in temperature across the inner radius of a pipe due to radial heat transfer. Hydrates may alternatively form at the gas-water interface as a dispersion where sufficient contact between phases is present (Bassani et al., 2017). Following hydrate particle nucleation and growth, formed hydrates may agglomerate to form a plug, which may be identified through a highly fluctuating pressure drop (Zerpa, 2013). Having nucleated and grown, hydrates may break from the wall and enter the fluid phase, and thus may agglomerate further downstream and result in hydrate plugs forming downstream from the site of hydrate nucleation, possibly even locations outside the expected regions of hydrate build up (Sloan & Koh, 2007). Having discussed how hydrate plugs form, it is now possible to discuss means of preventing or controlling the unwanted formation of hydrates.

1.5 Hydrate Control

Several means of preventing formation or controlling the rate of formation have been established. These approaches range from mechanical means such as insulation and routine pipe cleaning, to the addition of chemicals into the produced reservoir fluids which prevent or alter the mechanism of hydrate formation. Before examining the chemical means of hydrate

control, it is worth discussing factors which promote hydrate formation, and how these can be managed. Naturally, the presence of water is required to facilitate the formation of hydrates. Water enters pipelines through construction of the line and condensation of water dissolved in the reservoir fluid during pipeline stoppages, whereby the temperature profile in the pipeline begins to tend toward ambient conditions leading to condensation, furthermore, stoppages allow for adhesive bonds to form between the pipe and hydrate which would have been carried by the flowing phase (Makogon, 1997). Makogon (1997) attributes water in the pipeline tending to accumulate to form stagnant zones in the lower sections of an elevation in the pipeline as the main cause for the agglomeration of hydrate plugs, however these stagnant zones may be removed from the pipeline by means of pigging, whereby a pigging piston passed through removes water from the pipeline, provided this is physically possible due to pipeline structure. However, it is often not possible to avoid the presence of water in the pipeline, and eliminating water from the line may result in costly stoppages which may lead to further hydrate formation. Aside from managing the water content inside the pipeline where possible, the risk of hydrate formation may be reduced by burying and insulating the pipeline. Through burial and insulation, the heat transferred from the fluid phase into the ambient surroundings is significantly lowered. As such, this approach can be taken in an attempt to prevent the fluid phase of the pipeline from dropping below the hydrate equilibrium temperature into the zone in which hydrates are stable. In addition to this, electrical heating can be applied to sections of the pipeline, thus further increasing the fluid phase temperature and reducing the required concentration of inhibitors (Turner et al., 2014). Further consideration additionally needs to be given to the ambient temperature of the region through which the pipeline is laid, which greatly affects the hydrate equilibrium temperature. As such, pipeline insulation is often added to slow the rate of heat transfer between pipe contents and the sea floor if the pipe is buried, or the surrounding sea water if left exposed. The knowledge of this heat transfer is an essential quality used in determining the driving force of hydrate formation in addition to predicting the equilibrium conditions, however complexity arises as a result of seasonal variations in ambient conditions (Makogon, 1997), and must be considered when establishing a hydrate control strategy. As drilling operations move into deeper and colder waters however, these approaches may not be enough to prevent or reduce hydrate formation to an acceptable extent from an economic standpoint. As ambient conditions grow colder and pressures increase, the costs of adequate insulation may prove unacceptable. Furthermore, line burial may not be possible for sections of a pipeline such as vertical lines to the surface or pipeline crossings on the sea-floor. As such, chemical means of managing and preventing hydrates are popular, with the use of

thermodynamic inhibitors such as methanol are widespread throughout the industry, and are used for preventing hydrate formation inside wells, production lines, surface facilities and pipelines. These approaches are not mutually exclusive, chemical inhibitors are usually applied in conjunction with pipeline insulation and burial.

Three types of chemical inhibitors are currently used in industry: thermodynamic inhibitors, kinetic inhibitors and anti-agglomerants. Thermodynamic inhibitors such as methanol and ethylene glycols serve to reduce the hydrate equilibrium temperature, thus allowing the fluid phase of the flow line to operate at low temperatures without significant risk of hydrate formation occurring (GPSA, 2004). Thermodynamic inhibitors operate independently of formation kinetics by preventing the formation of hydrates (Perrin et al., 2013). These inhibitors function by lowering the temperature required for hydrate formation for a fluid phase of a specific pressure and composition. When using thermodynamic inhibitors, hydrate formation is completely unwanted, and the flow line is to be operated at temperature and pressure conditions outside of the hydrate stability zone. The use of thermodynamic inhibitors is not without disadvantages; these inhibitors do require addition into the fluid phase in large concentrations, often as high as 50%, and requires separation from the product, thus requiring separation equipment to handle this high volume of inhibitors which is often added continuously and thus proving costly (GPSA, 2004). Methods of thermodynamic inhibition when applied to deep or cold-water drilling so as to prevent hydrate formation may prove highly costly, so much so that the possibility of controlling the rate of hydrate formation rather than prevention becomes economically viable (Sloan & Koh, 2007). A separate class of inhibitors termed low dosage hydrate inhibitors (LDHIs) has been developed for this purpose, to control hydrate formation rather than prevent it. These LDHIs are branched into two categories: kinetic inhibitors and anti-agglomerants. These inhibitors prove more suited to production well within the hydrate stability zone, as the quantity of inhibitors added into the pipeline proves far lower than that of thermodynamic inhibition, thus facilitating lower costs of hydrate control in addition to capital savings through reduced inhibitor storage requirements (Perrin et al., 2013; Hawtin et al., 2008; GPSA, 2004). As opposed to thermodynamic inhibitors which prevent hydrate formation, kinetic inhibitors act in a dynamic manner, allowing hydrate nucleation while slowing the growth of nucleated hydrates, thus delaying plug formation. (GPSA, 2004). While using LDHIs, further consideration is required to account for expected stoppages in production, whereby the contents of the pipeline cease to flow and the nucleated hydrates are given greater time to grow and pipeline contents experience cooling towards

ambient conditions, which may result in the maximum subcooling tolerated by the inhibitor being exceeded (Sloan & Koh, 2007). Unlike kinetic inhibitors which slow the rate of hydrate formation, anti-agglomerants allow hydrates to form, while dispersing and transporting these hydrates downstream with the pipeline flow, thus limiting hydrate particle size without slowing formation rate (Frostman, 2000). Through preventing hydrates from aggregating downstream of the site of hydrate growth, plug formation is avoided. Instead, hydrate particles travel downstream and are later removed. Unlike thermodynamic or kinetic inhibitors which may operate with or without liquid hydrocarbons present in the pipeline, anti-agglomerants require the presence of liquid phase hydrocarbons (Perrin et al., 2013). As gas-phase wells and flow-lines are the focus of this research, the presence of a significant liquid flow in the pipeline cannot be ensured, as such this is not recommended as an inhibition strategy for use with the model developed in this research.

Hydrate control is not exclusive to flow-lines and surface equipment. Control of hydrates which form in reservoirs being drilled or produced is performed by means of adding inhibitors to drilling fluids to prevent formation in the production lines; additionally, external heating could be applied by means of heated fluids such that the temperature of the line exceeds hydrate equilibria (Makogon, 1997). Through use of thermodynamic inhibitors, the well can be operated outside the hydrate stability zone (Makogon, 1997).

Overall this short review of hydrate control methods is provided to illustrate that the procedure for preventing or controlling hydrate formation is a highly varied process with several options available to achieve the desired hydrate control strategy objectives. Furthermore, various control strategies may incur large capital and operational costs (GPSA, 2004; Perrin et al., 2013). As such, due to the multiple competing strategies of hydrate control it is essential to have available accurate information regarding the hydrate equilibrium conditions, which are central to determining whether hydrates will form or the driving force promoting hydrate growth. Accurate predictions could allow for lower margins of heating above the predicted equilibrium temperature due to uncertainty in predicted results being reduced, and similarly would allow higher operating pressure for flow-lines. Furthermore, this would also prove useful for kinetic models requiring equilibrium conditions by once more reducing uncertainty for subsequent calculations and thus the need to significantly adjust model results to account for this. As will be discussed in detail in the following sections, a significant number of models have been developed which determine the conditions of hydrate formation when provided with the composition of the fluid and either the temperature or pressure of said fluid. As the field of

gas hydrate research has an extensive history, many different classes of model have been created, with the use of models based on statistical thermodynamics widespread throughout the industry. Over time, the accuracy of these models has greatly increased. As an increasingly wide range of conditions and gas compositions are likely to be encountered with wells being developed in deeper and colder waters, it is required for general hydrate prediction models to cover a universal range of conditions. This becomes challenging due to the highly complex nature of hydrate formation, particularly considering multicomponent systems such as natural gas, where a great many factors influence hydrate equilibria. Recently, new methodologies for developing hydrate equilibrium prediction models have been established with the advent of machine learning becoming increasingly viable. Of particular note has been the popularization of Artificial Neural Networks, which can be used to provide a continuous real number output through means of a regression.

Thus, it can be seen that there is a need for accurate models capable of predicting the conditions of hydrate equilibria. A review of several current models is provided later in this report. This research aims to develop a neural network capable of predicting the hydrate equilibrium pressure, trained and tested on datasets comprised of equilibrium points found from published experimental studies on gas hydrates. A data sampling campaign has been undertaken with the obtaining of natural gas or synthetic natural gas data prioritised. An emphasis is placed on ensuring the model performs accurately when specifically considering multicomponent gases, so as to increase confidence that the model is applicable to natural gas.

1.6 Gas Hydrate Equilibria and the Metastability Effect

Before discussing the model development, it is essential to clarify exactly what prediction the model is making, and why the predicted quality proves useful in predicting, preventing and possibly treating gas hydrate formation. The final goal of all models developed in this research, is the prediction of the pressure under which gas hydrate equilibria will occur for gas of a certain molar composition, at a certain temperature. The nature of this equilibrium point is complex, and elaboration on this occurrence is worthwhile as equilibrium data in this investigation is taken from a wide range of experimental studies, thus resulting in the definition of hydrate equilibrium having significant importance to this study.

While it is clear that gas hydrates will not form before certain temperature and pressure conditions are met, hydrate formation for mixtures of components has proven to be a complex issue. A significant amount of research has been performed on the nature of hydrate formation

thermodynamics and kinetics. While predicting the conditions of hydrate equilibria yields a definitive result, the significance of this point requires elaboration. Unlike a saturation curve, where either side of the curve represents 100% of one phase, the equilibrium curve of gas hydrates for mixed components does not exactly exhibit this behaviour due to the metastability effect of the system, referring to the persistence of hydrates outside of the conditions where hydrates are thermodynamically stable (Ward, 2015). This metastability effect is prevalent for systems containing propane, an sII hydrate former which is prevalent in natural gas systems (Ward, 2015). Many publications in the field of predicting gas hydrate equilibrium conditions use the terms of formation and equilibrium interchangeably, which can result in a degree of confusion as to what state the model is actually predicting. As the purpose of this research is to predict the equilibrium conditions for a given gas mixture, it is worth further elaborating on the nature of hydrate equilibria so as to dispel uncertainty over the definition of the equilibrium point, and to provide certainty that a useful quality is being predicted by the model. The discussion which follows is based on existing methodologies of obtaining isochoric gas hydrate equilibrium measurements, with experimental methodologies presented in Stringari et al. (2014) and Ward (2015).

In order to illustrate the significance of the equilibrium conditions for a mixed gas hydrate, consider Figure 1.1, which has been drafted to detail the overall concept of hydrate stability through investigating an isochoric system which forms then dissociates hydrates. Figure 1.1 represents a simplification of isochoric experimental observations for hydrate formation in isochoric, closed conditions, further details of the experimental approach itself can be found in Stringari et al. (2014) and Ward (2015). Note that other experimental approaches aside from isochoric methods have been applied, such as the isothermal pressure search method (Bishnoi & Dholabhai, 1999). Assuming natural gas is subject to temperature and pressure measurements within a closed system of constant volume with the purpose of determining hydrate equilibrium conditions, point (i) in Figure 1.1 represents the initial temperature and pressure conditions of the gas in the experimental apparatus. In order to form hydrates, the gas is cooled from temperature (i) to (ii). At conditions (ii), the metastability limit of the gas is reached, and a large quantity of hydrates rapidly form, with the pressure of the system dropping to point (iii) due to the gas in the isochoric system becoming encaged (Sloan & Koh, 2007). Having completely formed all hydrates the system can accommodate, the system is heated to (iv), after which the system is gradually heated in a stepwise manner (Stringari et al., 2014), and hydrate dissociation begins to occur. The dissociation is marked by an increase in pressure,

indicating that some of the encaged molecules have been released as gases. Further heating is accompanied by a sharp gradient change in the pressure-temperature curve, with complete dissociation of hydrates is marked by the point at which the last hydrate crystal dissociates (Stringari et al., 2014), corresponding with (v) in Figure 1.1. The pressure gradient at total dissociation follows approximately the same gradient as the gas prior to hydrate formation. (v) is defined as the equilibrium point of the gas mixture (Ward, 2015). Thus, it can be seen that while hydrate equilibrium occurs at point (v), for a system at initial conditions (i) experiencing cooling to (v), hydrate formation is not likely to occur. Sloan & Koh (2007) attributes this phenomenon to the metastability of the system, and comments that during the cooling from point (i) to (ii), no hydrates are observed after extended time periods before the metastability limit (ii) is reached.

In order to ensure more replicable and meaningful measurements, Sloan & Koh (2007) recommends defining the hydrate equilibrium point as the conditions required for complete dissociation of hydrates to occur. Measurements of the equilibrium point, upon which the models in this investigation are based, are far more reproducible than those of formation, be they at nucleation or the metastability limit. While measurements of the equilibrium point may also be subject to metastability should the heating between stages (iv) and (v) in Figure 1.1 occur too rapidly, a low rate of heating during dissociation will overcome metastability and ensure a minimized error of the dissociation point measurement, at the cost of increasing the time taken to perform experimental measurements (Tohidi et al., 2000). Failure to account for the metastability effect during heating may lead to significant error (Ward, 2015). Furthermore, hydrate formation may occur in pipelines in a different manner compared with experimental data obtained from pressurized reactor due to the velocity of the contents within the pipeline, or the presence of hydrate seed crystals (Ahmed & McKinney, 2011), which Ikoku (1980) lists as a hydrate promoting factor. As such, to prevent hydrates forming prior to the metastability limit being reached, adequately accounting for metastability and ensuring no seed crystals are present is essential. This factor further illustrates the utility of a dissociation point based definition of hydrate equilibria. Additionally, the memory effect present in systems which previously contained hydrates could result in a narrowed metastability range (Sloan & Koh, 2007). Wu & Zhang (2010) determined that residual hydrate structures following dissociation, could act as seeds for subsequent hydrate formation, thus reducing the time for hydrate nucleation to occur. This memory effect is of particular importance to the experimental studies from which data was sampled to develop the models which are the focus of this research. There

is no way of knowing as to whether after successive experiments using the same apparatus and initial sample, sufficient time or heating was given between experiments to account for the persistence of a memory effect. Furthermore, use of formation metrics would involve the need to specify whether the measured point took place at nucleation or the metastability limit, both of which are subject to the host of influencing factors discussed. It can thus be seen that ensuring replicable results when using hydrate formation metrics proves highly challenging, and poses a significant potential source of error to data sampling campaigns requiring a wide range of data. Defining the equilibrium point as the conditions at which formed hydrates completely dissociate involves far fewer degrees of variability and influencing factors, and is well suited to the development of models designed to operate over a universal range of conditions.

Overall, the equilibrium curve for a gas of a given composition yields the boundary between the region where hydrates will not occur, and the region wherein formed hydrates are stable. This proves a useful quality, as thermodynamic approaches to avoiding hydrate formation in natural gas systems involve operating outside the zone wherein hydrates are stable (Ahmed & McKinney, 2011). Through equilibrium measurements being replicable and possessing a set definition, it is possible to sample a large quantity of experimental data from literature for the purpose of model development.

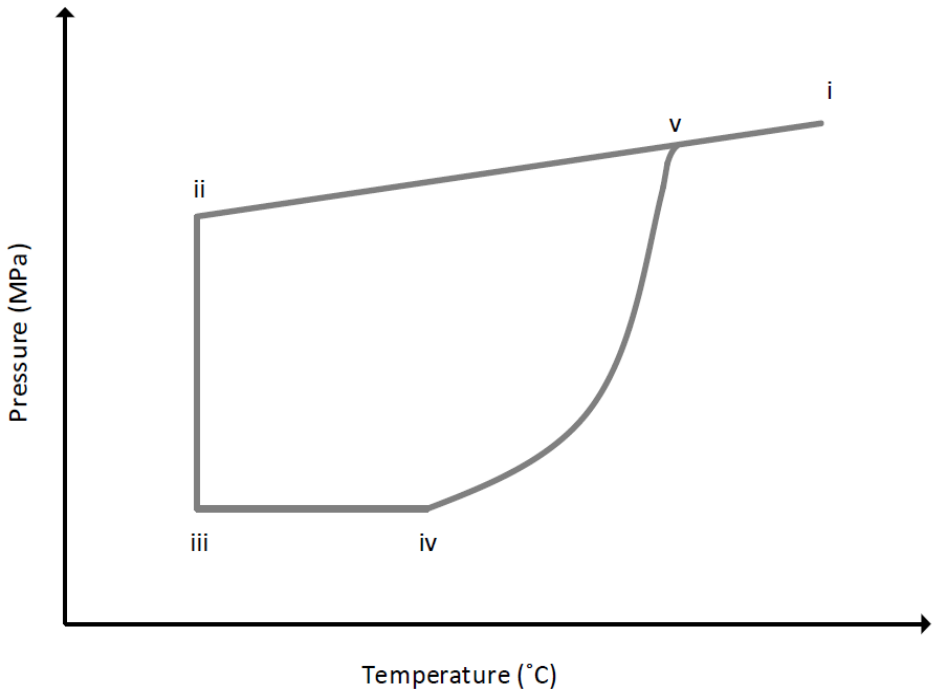

Figure 1.1: Isochoric Hydrate Equilibrium Measurement. Discussion and diagram influenced by the works of Stringari et al. (2014), Sloan & Koh (2007) and Ward (2015)

CHAPTER 2: CURRENT AND HISTORICAL MEANS OF PREDICTING GAS HYDRATE EQUILIBRIA

Due to the costs incurred by hydrate formation, significant incentives into the research of gas hydrates are present through practical application being achievable with significant results. A significant number of models designed to predict gas hydrate equilibrium conditions have been developed, many of which are models used by corporations with specific criteria to suit the range of conditions dealt with. Predictive models to date have been developed using a wide range of approaches. Amongst the early models to prove sufficiently accurate to be used was the Katz (1945) gas gravity plot, which involves graphically determining the expected formation temperature or pressure by providing the gas gravity of the sample gas being investigated. In this case, determining the gas gravity requires the molecular weight of the sample being investigated, and hence the composition of the gas. These gravity charts have been developed based on experimental data and calculations (Shahnazar & Hasan, 2014). Sloan & Koh (2007) details that since accurate hydrate prediction models formulated and tested on a wide range of data have been developed, gas gravity methods are often used as a preliminary estimate of the equilibrium conditions due to the ease of obtaining a prediction. Shahnazar & Hasan (2014) details the shortcomings of specific K-factor models, which are also graph-based hydrate predictors. GPSA (2004) recommends not using the Katz gravity charts for gases with significant quantities of carbon dioxide or hydrogen sulphide, but states that the method yields potentially acceptable results for systems excluding hydrogen sulphide. Furthermore, GPSA (2004) does not recommend using this method for calculating hydrate formation conditions above approximately 10 MPa.

Amongst the models aiming to yield more reliable results than the gas gravity model, are thermodynamic models. One of the first successful thermodynamic models was developed in Van der Waals & Platteeuw (1959), which has served as a basis for many subsequent models (Giavarini & Hester, 2011; Shahnazar & Hasan, 2014; Antunes et al., 2018). Such models entail the prediction of hydrate equilibrium conditions through use of a statistical thermodynamic approach based on intrinsic factors of the gas investigated, and depending on the model, the minimization of Gibbs free energy (Giavarini & Hester, 2011). The advantages of a thermodynamic model are discussed in Sloan & Koh (2007). A particularly advantageous factor is that thermodynamic models can be developed without requiring a massive compilation of experimental data. The lack of experimental data involved during model development avoids

potential experimental errors leaking into the model. GPSA (2004) asserts that presently equation of state based computerized models offer the most consistent means of prediction gas hydrate equilibria, with a low variance in results on average. These thermodynamic models are capable of determining hydrate structure, even multiple hydrate structures in the same environment (Giavarini & Hester, 2011). Ballard & Sloan (2002) investigates the development of a multiphase model minimizing the Gibbs free energy. Many of these popular models use as a base the Van der Waals & Platteuw (1959) model or later variations to model the hydrate phase, while an equation of state such as Peng-Robinson is applied to modelling the fluid phase (Antunes et al., 2018). Ballard & Sloan (2004) and Giavarini & Hester (2011) list computerized thermodynamic models commonly applied in industry, and Ballard & Sloan (2004) tests the accuracy of several computerised programs developed to predict hydrate equilibria based on thermodynamic models which are commonly used in industry, and results indicate that the models do accurately predict unseen test data. The results of the Ballard & Sloan (2004) investigations include ternary and natural gas component tests, which is the focus of this research, and indicates accurate predictions for a moderate sample size. It can thus be seen that successful gas hydrate prediction models have been achieved through a statistical thermodynamics base, and any subsequent models attempting to predict a universal range of conditions need to be able to compete with these models which have achieved a degree of industrial acceptance. Chapoy et al. (2007) explains several shortcomings of the thermodynamic model, notably the difficulty of adjusting model parameters, and uncertainty as to whether or not the model yields the optimal result as opposed to a result corresponding to a local minimum.

A modern approach to modelling gas hydrate equilibrium has been achieved through the application of machine learning. Unlike intrinsic thermodynamic models, machine learning allows for easily measured variables, such as those listed in Table 3.1, to be used to generalize behaviour when trained over a wide range of data. The purpose of this is to attempt the development of a model more capable of accurately predicting highly non-linear, multimodal phenomena such as the equilibrium conditions of gas hydrates than models based on statistical thermodynamics. Prior to popularization of the back-propagation algorithm and the commercialization of increasingly powerful single-core processors, allowing for neural networks to gain traction in the field, the majority of machine learning application to the field of gas hydrate equilibrium prediction has been achieved through Support Vector Machine programs. With increasingly powerful processors becoming commercially viable, artificial

neural network approaches have become prevalent in the field of gas hydrate equilibrium research.

Note that the following models listed are discussed under the lens of natural gas hydrate prediction, with specific focus on multicomponent hydrocarbon gas systems. As such, models designed exclusively to predict the hydrate equilibrium conditions for pure and binary components are not discussed.

Within the field of predicting the equilibrium conditions of gas hydrates, a wide range of studies into machine learning models have been published to date. The input features of the models which have been developed vary considerably. While the output of each regression model is either the temperature or pressure at which hydrate equilibria forms for either the temperature or pressure specified, different models may account for the compositions of the gas phase differently, and may include the presence of inhibitors or electrolytes as a feature. Heydari et al. (2006) and Hesami et al. (2016) developed neural networks to predict the temperature of hydrate formation from an input of two features, gas phase pressure and the specific gravity of the gas. The specific gravity of a gas acts as a form of lumped factor which is used as opposed to inputting the exact composition of each component of the gas phase. The Heydari et al. (2006) and Hesami et al. (2016) neural network models were developed using data sets consisting of under 400 data points. In both cases, the dataset the models were developed from was separated into a training and test data by a pre-defined split, such that the model performance could be tested on data unseen by the model. Other machine learning models have been developed to account for the fraction individual components within the gas phase as inputs to the network. Zenali et al. (2012) developed machine learning models capable of predicting gas hydrate equilibrium conditions through means of both Neural Network and Adaptive Neural-Fuzzy Interface (ANFIS) models for non-inhibited systems based on a dataset consisting of over 700 points. The dataset is split randomly such that two-thirds of the data which is randomly selected serves as the training data, while the rest is used for model testing. Inputs consisted of temperature, and the compositions of methane, ethane, propane, iso-butane, normal-butane, carbon dioxide and hydrogen sulphide.

Ghavipour et al. (2013) reported difficulty in developing a convergent model for the input of specific gravity, and subsequently developed a successful model predicting hydrate equilibria for the inputs of pressure and the gas phase composition of methane, ethane, propane, and normal-butane. The dataset for this model consisted of 130 data points. An interesting feature

of this model is how validation is performed. Unlike many other models, the split of training-test data is governed by the process of leave-one-out validation. This is a form of cross validation where the data set is split into partitions, such that for a dataset with n samples, the training data corresponds to size (n-1) for partitions of size 1, thus each partition is tested once (Ghavipour et al., 2013). This validation practice does provide a means of judging how well the model performs across the entire dataset, rather than a narrow testing range.

Several neural networks have been developed so as to predict hydrate equilibrium conditions in systems where thermodynamic inhibitors, and electrolytes are present. Thermodynamic inhibitors serve to alter the conditions at which hydrate equilibria occurs, typically to cause hydrates to achieve stability at a lower temperature and higher pressure. The presence of electrolytes additionally alters the mechanism of hydrate formation, and affects the conditions at which hydrate equilibria occurs. Elgibaly & Elkamel (1998) developed several models with inputs of pressure and gas gravity, in addition to separate models accounting for the composition of methane, ethane, propane, iso-butane, normal-butane, pentanes, carbon dioxide, nitrogen, hydrogen sulphide and the presence of hydrocarbons with a carbon number greater than five. Of particular interest is the neural network model of Elgibaly & Elkamel (1998), developed to account for a total of 16 inputs, including the presence of several thermodynamic inhibitors and electrolytes. Validation was performed by separating the dataset into two separate sets, one for training and another as held-out data unseen by the neural network used to test the predictive capabilities of the model. This model was later expanded on in Elgibaly & Elkamel (1999) where an optimization was performed such that the cost of inhibitors was accounted for and the optimal thermodynamic inhibition strategy is provided. This was achieved through the development of a neural network which is trained on the inputs of pressure, gas composition and the presence of thermodynamic inhibitors and electrolytes to predict the temperature at which equilibrium is attained. Elgibaly & Elkamel (1999) then developed a subsequent neural network taking into account the gas composition or gas gravity, in addition to both the temperature and pressure of the system and generated an output reporting the concentration of each specific inhibitor that would be required to shift equilibrium conditions to the desired temperature and pressure. Upon providing cost data concerning inhibitors, it is possible as is shown in Elgibaly & Elkamel (1999) that such a model can be used to optimize the thermodynamic inhibition strategy through achieving the desired equilibrium conditions at the lowest cost. Chapoy et al. (2007) developed a neural network from 19 inputs, including gas compositions, temperature, inhibitor and electrolyte

concentration in solution in addition to the hydrate structure. The model separates the dataset into training data, used to develop the model, testing data which is used to tune the model, and validating data which is held-out and unseen by the neural network and used to assess the overall accuracy of the model. The dataset for the Chapoy et al. (2007) model proved extensive, consisting of over 3000 data points, including data for methane gas phase compositions less than 50%.

While neural networks have largely been of the focus of recent model development, support vector machine models have continued to be developed. Ghiasi et al. (2016) developed a support vector model with inputs considering the composition of the gas, in addition to possible thermodynamic inhibitor and electrolyte presence, from a dataset of nearly 4000 points to predict the dissociation temperature of formed hydrates. This model employed the practice of separating the compiled dataset into training, test and validating data. This practice is performed such that model tuning can occur while lowering the risk of overfitting.

While many studies cover an extensive range of conditions, many include a copious amount of data used to both train and test the neural network which consists of pure and binary components, which is largely irrelevant to the application of natural gas itself. While the inclusion of pure and binary component data, particularly methane-propane interactions, can assist a model in distinguishing between sI and sII states of hydrates, and possibly mixtures of the two, testing the model on this data could possibly serve to inflate the final regression score while potentially yielding little information regarding the model's ability to predict complex systems consisting of multiple components, which is of significant importance when dealing with natural gas. This issue can result in a loss of confidence in expecting to achieve results within the reported accuracy of the model when predicting natural gas hydrate equilibrium conditions. Several studies have attempted to develop models exclusively using multicomponent databases, however due to a lack of reliable multicomponent data publically available, this has proven challenging. Soroush et al. (2015) developed a neural network consisting of two Hidden Layers, trained on a dataset consisting of just under 300 equilibrium points obtained from literature. The model was developed by randomly selecting 42 data points to be held out for validation purposes, while the model itself was developed and tuned on a split of 201 training points with the remainder used for testing purposes. Following model training and tuning, the model of Soroush et al. (2015) was tested through using all training data to train the neural network, and performing validation using the held-out data, resulting in a R^2 coefficient of 0.998 being obtained. While this model is limited through use of a small

dataset, what is remarkable about this study is the restriction of the methane + ethane molar gas phase composition limited to a minimum value well above 50%, and still yields an accurate regression score (Soroush et al., 2015). This restriction raises confidence in the models capability of predicting multicomponent systems through outright rejecting ranges of data irrelevant to natural gas systems, and thus resulting in the final result being a good indicator of performance. This study revealed the possibility of developing a convergent neural network on a limited dataset. This contrasts traditional neural network models in the field which largely consist of a large amount of pure and binary component equilibrium points in the datasets used to train and test models.

As it can be seen, a significant degree of study has been directed toward the development of models capable of predicting gas hydrate equilibrium conditions. Shahnazar & Hasan (2015) indicates that at the time of their publication, the majority of publications in the field had still been directed toward the development of thermodynamic models. While many neural network and other machine learning models have been able to successfully predict gas hydrate equilibrium conditions based on a wide range of conditions when tested with experimental data, there is a noticeable lack of an indication as to the error margins of the results from many models. As such, there is significant concern as to the statistical significance of the results of many machine learning studies in the field. Furthermore, as datasets used to train these machine learning models are likely to be non-gaussian, it is imperative that model testing occurs over a significantly wide range of data, such that surety is obtained that model results are not merely representative of a favourable randomized selection of training data. Non-Gaussian datasets imply that randomly sampling indices for training and testing is unlikely to be representative of the entire dataset. This research aims to address these issues which are current weaknesses in many existing machine learning models in the field, and in doing so aims to yield statistically sound model results and to prove the viability of the neural network methodology as a whole for the prediction of gas hydrate equilibrium conditions. Through application of the practice of cross-validation, the model may be trained and tested for several different combinations of data prior to validation with a held-out dataset, while providing the bias and variance, which are essential quantities used to assess machine learning models based on experimental data. This will assist in identifying whether or not unexpectedly weak predictions are made by the model, be it as a result of overfitting or a lack of data for certain conditions. Overall, while hold-out validation will yield a single reported accuracy of the model, cross validation will allow for multiple performance metrics to be obtained for different combinations of train-test datasets,

thus yielding the bias and the model variance across a number of combinations. Furthermore, as the overwhelming majority of data used to develop most models has been sampled from literature, with work on hydrate equilibrium experiments dating as far back as the publication by Hammerschmidt (1934), confidence in some sources of data may be raised due to now outdated methodologies such as optical measurements of hydrate formation. The practice of grouping data per unique source and prohibiting the training and testing of data from the same group will provide an indication as to the dependency of a model on individual sources for data covering certain ranges. Establishing the difference between results trained on grouped and ungrouped data will provide an indication as to how susceptible the model is to experimental, methodology or measurement errors, and will assist in providing a lower estimate of the predictive capabilities of the model when testing on truly unseen data. Finally, through use of datasets that exclusively contain multicomponent data, further confidence in the methodology could be achieved when high accuracy predictions are made by the model when tested by multicomponent data in the absence of pure or binary data which could serve to provide a poor estimation of model performance concerning natural gas environments. The research performed in this publication will seek to develop several models trained on two different datasets, a multicomponent exclusive dataset, and a dataset similar to many used in literature, consisting of a roughly equal mixture of multicomponent data and pure and binary component equilibrium points. Furthermore, the models developed will undergo extensive cross-validation practices to ensure that the model does not perform well only for select combinations of training and test data, while the effect of grouping data by source will be investigated. Cross-validation will provide the bias and variance of results, while Hold-out validation is performed to provide the overall indication of model accuracy while ensuring that overfitting has not occurred through testing the cross-validated model on data completely unseen during the training process.

CHAPTER 3 METHODOLOGY

3.1 Overview

The methodology selected to model the conditions of gas hydrate equilibria is that of a neural network. While significant discussion is presented regarding the topology, training and validation of the model, the neural networks developed in this investigation may be summarized as feed-forward, multilayer perceptron, artificial neural networks. The various neural networks developed in this investigation are trained in a supervised manner by means of one of two datasets comprising of 670 and 1209 experimental equilibrium measurements sampled from various publications which are listed in appendix A. Results obtained include various metrics which determine the ability of the model to predict unseen data, and a final estimate along with possible variance are provided. Deep learning has been incorporated into model development, and all neural networks developed in this investigation consist of at least two hidden layers. The trained models have been evaluated according to 10-fold cross-validation, and additionally in the case of several models, hold-out validation. By incorporating both cross-validation and hold-out validation, models have been tuned by a Grid-search iterative procedure to yield final models which are likely close to the optimal configuration. Due to a slight variance in results over repeating the same model configuration and parameters, true optimization cannot be performed in this case. The main scoring metrics used in this investigation include the R^2 regression coefficient, and the variance using this metric has been determined through use of cross-validation

3.2 Dataset Sampling

In order to train and validate a neural network, a vast quantity of data is required. Fortunately, a wide array of conditions for gas hydrate formation has been examined in various experimental studies. A large quantity of gas hydrate equilibrium measurements has been published. Due to the wide range of equilibrium measurements publicly available, all data sampled for the models developed has been obtained from literature. Creating the datasets used to model hydrate equilibria has been achieved through a rigorous data sampling campaign. It is worth noting that the compiled datasets do not cover an even distribution of the range of conditions investigated, which is largely due to the lack of experimental data for certain ranges. As such, the datasets are not considered to be Gaussian, which requires significant consideration when sampling data to be used in model training and testing. Much of the difficulty in performing gas hydrate equilibrium measurements is due to the time factor associated with experimental studies. As

discussed in Ward (2015), hydrate equilibria experiments require a significant time investment, largely due to the metastability effects requiring significant periods of waiting for hydrates to stabilize, and the slow rate of cooling required for hydrate dissociation so as to accurately obtain an equilibrium measurement. Significant differences in environment are often present between natural gas pipelines and laboratory experiments, notably the site at which hydrate formation initiates and agitation present in the pipeline. Ruffine et al. (2018) attributes much of the difference in hydrate formation between water oversaturated laboratory-scale experiments and natural gas equipment to a lack of time taken to gradually form stable hydrates, and insufficient time over which hydrates occupy a stable state.

Data has been extracted from several sources which include the following features listed in Table 3.1. All data sources used to develop the datasets used in this investigation have been listed in Appendix A. Note that a significant portion data sources were discovered through the works of Sloan & Koh (2007), which lists a great many hydrate equilibrium points from various sources. GPSA (2004) and Sloan (1998) identifies significant hydrate formers as methane, ethane, propane, iso-butane, n-butane, carbon dioxide, nitrogen dioxide and hydrogen sulphide. Due to the scope of this investigation and a lack of independent experimental studies publishing data for the component, the presence of hydrogen sulphide has not been included in this investigation. Thus, the study conducted is applicable to sweet natural gases. Furthermore, Carroll (2009) details heavy non-hydrate forming hydrocarbons as causing an azeotropic effect when mixed with pure methane in a system which forms hydrates. As such, the equilibrium conditions can possibly be affected by the presence of non-hydrate forming hydrocarbon molecules in the gas phase. The inclusion of heavy non-hydrate forming hydrocarbons is accounted for in the models developed in this investigation through the inclusion of an input consisting of a lumped sum of hydrocarbons present in the gas phase with a carbon number of five or greater. Finally, for the development of the model, both the extrinsic properties of the temperature and pressure under which hydrate equilibrium was attained are required.

$$P_{equilibrium} = f(T, C_1, C_2, C_3, iC_4, nC_4, C_{5+}, CO_2, N_2) \tag{3.1}$$

While GPSA (2004) details kinetics and mass transfer as additional factors influencing hydrate formation, it is assumed that data sources have adequately accounted for metastability effects when measuring equilibria. This could be achieved through allowing sufficient time for formed hydrates to stabilize, and a sufficiently low heating rate during dissociation up to equilibria (Ward, 2015; Tohidi et al., 2000). GPSA (2004) lists the dew point of the mixture as a factor

influencing formation conditions, in that the gas must be at a temperature lower than the dew point to facilitate hydrate formation. As hydrate equilibria is being measured, it is assumed that the temperature of the mixture being investigated is lower than the dew point of water in the system. While the exact quantity of water present in the system is not required as an input to the model, this model assumes that there is a sufficient quantity of water present, either in liquid phase or entrained in the gas such that hydrate formation is possible.

Table 3.1: Features influencing hydrate formation included in model development

Temperature (ºC)	C1 mol fraction	C2 mol fraction	C3 mol fraction	i-C4 mol fraction	n-C4 mol fraction	C5+ mol fraction	CO2 mol fraction	N2 mol fraction	Pressure (MPa)

While GPSA (2004) lists salinity as impacting hydrate formation, the effect of inhibitors and electrolytes on hydrate equilibria is beyond the scope this study, which is to establish confidence in the methodology as a whole. Detailed investigation into the modelling of Gas Hydrate formation in environments with inhibitors and electrolytes present has been conducted by means of a neural network in Chapoy et al. (2007) and Elgibaly & Elkamel (1999).

Collection of experimental data from literature has been achieved through a rigorous data sampling campaign. In order to promote confidence in the model, the inclusion of generated data of any kind has been avoided. As such, data reported by studies which involved using software or equations to predict the conditions of hydrate equilibrium have been excluded from this study. All data sampled has been obtained through experimental measurements of hydrate equilibrium. As a regression model is being developed, the inclusion of data outside of equilibrium biases the result and reduces the overall accuracy of the model. Thus, experimental data included in development of the model has been checked where possible to ensure that the reported values do indeed occur at equilibrium, rather than at some condition within the hydrate formation region. Several of the data sources used in this study have routinely been included in other prediction models in the field, notably Sloan & Koh (2007) which constitutes a significant number of equilibrium points used to compile the model datasets. While there is a significant overlap of data used in this study and others published in the field, several other equilibrium data sources which have been published in recent years have also been included to present an updated dataset which is less reliant on individual sources than other studies in the field. The overlap of data between this study and others may serve a basis for comparison between models.

3.3 Dataset Composition

Having collected a wide range of experimental data from various sources through a data sampling campaign, the data was compiled into datasets to be modelled. To standardize data, units of measurement for each feature have been converted to those in Table 3.1. Conversions between metric temperatures has been performed on the basis that that $0^{\circ}C = 273.15\,K$. Normalization has additionally been performed in setting the sum of molar fraction compositions to be unitary. The need for this arises from inconsistencies in the decimal places in several sources of reported data. Systems containing a relatively insignificant quantity of inert components which do not interfere with hydrate formation have similarly undergone normalization to fit within the features of the model.

Having collected a vast quantity of data, it became possible to develop two distinct datasets which would be used to train and test a neural network. A distinct lack of studies explicitly restricted to multicomponent systems containing three or more components is present in literature, particularly in terms of neural network models. Some past studies have included vast quantities of data irrelevant to the field of natural gas, such as equilibria of pure nitrogen or carbon dioxide, which often serve little use in modelling natural gas systems, and may result in biasing the reported regression score. Due to the lower complexity when compared with natural gas mixtures, pure components exhibit lesser difficulty to model, hence their inclusion in datasets used to train models could possibly result in a high regression score for a model that may not adequately predict natural gas hydrate equilibria. As such, a minimum of 50% methane concentration for the gas phase is imposed on the data included in all datasets utilized in modelling. This constraint serves to provide a model specifically designed to predict natural gas equilibria, and to avoid artificially inflating the regression score.

Data sampled consists of sI and sII hydrates. The work of Tohidi et al. (2001) indicates that the occurrence of structure H hydrates in natural gas systems at equilibria is unlikely for normal operating ranges, and as such there is no need to include this structure in the equilibrium model. It must be emphasized however, that with operating conditions well within the hydrate stability zone, it is possible that structure H hydrates could arise. For the sake of this model, it is assumed that in the case of a thermodynamic inhibition strategy, the hydrate stability zone is avoided, while for other inhibition strategies that the operating temperature is not low enough for significant periods of time that would lead to structure H hydrates. These assumptions do not

account for unexpectedly long pipeline shut-ins, where heat transfer results in the fluids in the pipeline gradually cooling to ambient temperature.

In order to gauge the performance of the model in terms of multicomponent data prediction, two datasets have been compiled to develop distinct neural network models. Through proving the ability of an accurate model to be developed from the exclusively multicomponent data set summarized in Table 3.2, a model trained from the complete dataset summarized in Table 3.3 could be expected to perform comparably as all multicomponent data is present in the complete dataset. The final model developed in this research will be trained and tested using the complete dataset.

The neural network developed by Soroush et al. (2015) was among first to model a neural network through means of a multicomponent exclusive dataset to yield a R^2 score of slightly over 0.998. This model is however limited through a dataset consisting of only 279 equilibrium points being used to train, test and validate the model. The limited range of the data in this case results in the model being highly dependent on a few individual data sources, and renders testing the model over a wide range of conditions difficult. Nevertheless, a successful neural network exclusively trained and tested on multicomponent data lends credibility to the methodology in the field of natural gas hydrate equilibria prediction. As such, through use of additional data sources, some of which have been published since 2016, a new dataset has been compiled in this investigation through use of 670 exclusively multicomponent samples with the intent of developing a model less reliant on individual data sources through the practice of grouping. A summary of the compiled dataset can be seen in Table 3.2. Through excluding pure and binary components, the model will provide a lower estimate of the prediction capabilities of neural networks in the field trained by published experimental data. This is a significant investigation as viability in the methodology can be assessed through the knowledge that the reported regression score and cross-validation performance of the model is reflective of a wide range of multicomponent data, rather than an abundance of high scoring pure or binary components.

The exclusion of pure and binary components from the dataset however is likely to yield poor predictions for ranges of data which are absent from the multicomponent exclusive dataset. Including a quantity of methane and binary components across the operating range of the model would allow for better results for conditions where multicomponent data is absent. As such, a separate dataset containing a total of 1209 experimental equilibrium data has been compiled.

This dataset includes the 670 multicomponent samples from the exclusive dataset, in addition to a large quantity of pure methane and binary component data, such as methane-propane and methane-isobutane equilibria. A summary of the complete dataset can be seen in Table 3.3. The inclusion of pure methane and binary component data results in the trained model being more likely to distinguish between hydrates of structure I and structure II. As such, the model trained on this complete dataset is likely to result in a higher regression score than the multicomponent exclusive model, in part due to the improved ability of the model to predict test samples for ranges significantly different from data used to train the model, and in part due to a slight inflation to the regression score due to some pure or binary component data present during model validation.

Table 3.2: Multicomponent Exclusive Dataset Summary

Feature	Temperature (˚C)	C1 mol fraction	C2 mol fraction	C3 mol fraction	i-C4 mol fraction	n-C4 mol fraction	C5+ mol fraction	CO2 mol fraction	N2 mol fraction	Pressure (MPa)
Sample Count	670	670	670	670	670	670	670	670	670	670
Mean	13.5732	0.8524	0.0651	0.0300	0.0038	0.0065	0.0024	0.0145	0.0253	8.1922
σ	6.9344	0.1026	0.0604	0.0304	0.0078	0.0115	0.0057	0.0350	0.0620	8.8865
Min	0.0000	0.5000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.4966
25%	7.4500	0.8025	0.0169	0.0096	0.0000	0.0000	0.0000	0.0000	0.0000	2.6164
50%	14.1550	0.8654	0.0543	0.0204	0.0001	0.0004	0.0000	0.0004	0.0004	5.1435
75%	19.2502	0.9280	0.0795	0.0357	0.0040	0.0085	0.0018	0.0143	0.0090	10.4623
Max	30.3735	0.9940	0.2500	0.1698	0.0461	0.0510	0.0340	0.3140	0.4000	68.2300

Table 3.3: Complete Dataset Summary

Feature	Temperature (˚C)	C1 mol fraction	C2 mol fraction	C3 mol fraction	i-C4 mol fraction	n-C4 mol fraction	C5+ mol fraction	CO2 mol fraction	N2 mol fraction	Pressure (MPa)
Sample Count	1209	1209	1209	1209	1209	1209	1209	1209	1209	1209
Mean	12.5068	0.8758	0.0413	0.0223	0.0066	0.0049	0.0014	0.0225	0.0252	9.2320
σ	7.3586	0.1168	0.0620	0.0445	0.0301	0.0109	0.0044	0.0667	0.0725	11.1019
Min	0.0000	0.5000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.1800
25%	6.1500	0.8375	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	2.8300
50%	12.1000	0.8950	0.0088	0.0019	0.0000	0.0000	0.0000	0.0000	0.0000	5.3778
75%	18.3500	0.9725	0.0647	0.0308	0.0029	0.0055	0.0000	0.0051	0.0068	10.8080
Max	31.8500	1.0000	0.4360	0.4980	0.5000	0.0582	0.0340	0.5000	0.4975	72.2600

3.4 Dataset Grouping

As will be detailed in model validation, the development of a validation set which is held-out of neural network training for the purpose of testing the model on unseen data necessitates the grouping of data by source so as to increase the likelihood that the training-validation split of data is representative of a wide range of equilibrium conditions. Furthermore, due to the

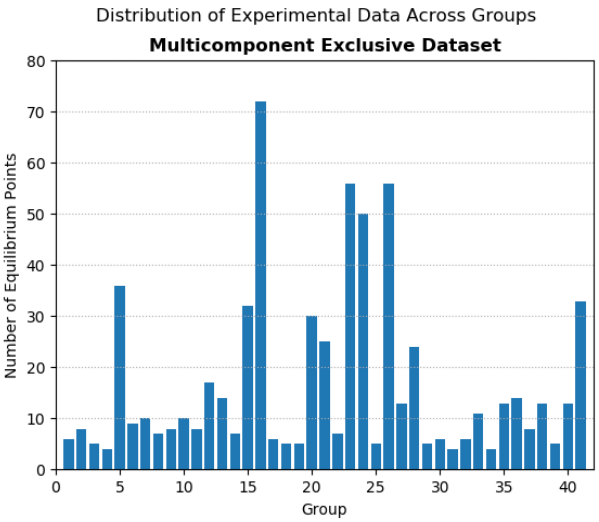

Figure 3.1a: Multicomponent Exclusive Dataset Grouping

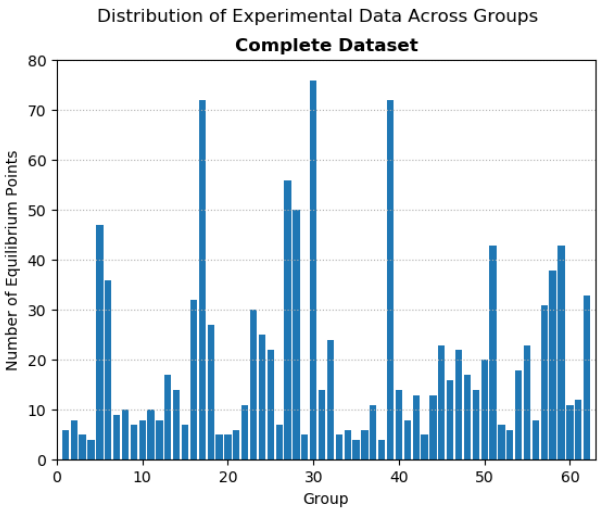

Figure 3.1b: Complete Dataset Grouping

inability to definitively determine as to whether or not metastability was adequately accounted for in several sources of data, it is expected that a few data sources report measurements that may not precisely report the conditions at equilibrium. Ward (2015) recommends new measurements to confirm equilibrium data for measurements reported in literature for certain components such as hydrogen sulphide. In order to reduce the potential impact that measurement errors may have on the model, and to additionally account for differences in mass transfer due to insufficient time taken between heating steps, a decision was made to group data according to experimental source. As such, studies measuring hydrate equilibria using the same apparatus, under the same ambient conditions are grouped. This grouping further has the benefit of providing a true measure of regression variability through the process of cross-validation displaying model results across a wide combination of groups. Several models have been developed in this study, and the effect of grouping on the overall regression and cross-validation folds is assessed. The grouping strategy is illustrated in Figure 3.1a and Figure 3.1b. Groups have been selected such that sources reporting less than four equilibrium points are excluded from the dataset due to difficulties achieving a converging model when small groups are present. Large groups of 50 or more equilibrium points are restricted to multicomponent sources, and large binary sources are divided into groups based on the components recorded. Finally, pure component sources have been grouped into two groups such that an even distribution of pure methane is present over 81 of the 123 pure methane equilibrium measurements. This has been done so as to ensure that pure methane sources are not heavily represented in the randomized train-test samplings during both cross and holdout validation. Overall the practice of grouping reduces the reported R^2 score of the model, while lending confidence to the reported result through indicating that the model is less dependent on

individual data sources covering specific ranges of conditions. While the purpose of the data sampling campaign is to provide a wide enough dataset such that statistical variations between samples are rendered insignificant, due to the non-Gaussian nature of the datasets and the limited number of independent experimental studies, it cannot be assumed that significant statistical differences between experimental studies examining similar conditions are not present. The development of grouped models will aid in reducing the impact of these factors.

3.5 Model Datasets

In order to assess the effectiveness of the developed model at predicting the hydrate equilibrium conditions for multicomponent gases, two distinct datasets have been compiled for model development. The first dataset bears similarity to the datasets used to develop various other neural networks in the field, consisting of a wide range of methane inclusive equilibrium data points. A constraint for this dataset is a minimum methane molar concentration in the gas phase of 50%. The dataset includes equilibrium points for pure component methane, in addition to binary and multicomponent gas mixtures consisting of the components listed in Table 3.1. A second dataset has been developed, which restricts equilibrium data to gases consisting of at least three components. Equilibrium data on these multicomponent gases is constrained by a minimum gas phase methane molar concentration of 50%, while excluding pure methane and binary methane gas mixtures. Both datasets, henceforth referred to as the Complete and Multicomponent Exclusive datasets, have been used to develop several models, as detailed in Table 3.4. The purpose for using different datasets for model development is to facilitate a comparison of model performance when examining the predictive capability of a model trained and tested exclusively using multicomponent data, with the results of a model developed using a wide dataset including pure methane and binary components. Due to the complexity associated with predicting multicomponent hydrate equilibrium conditions, a very large dataset is required. Due to the limited number of publicly available equilibrium data for multicomponent gas hydrates, the multicomponent exclusive dataset consists of significantly fewer samples than the complete dataset. Comparing the results of models developed using the multicomponent exclusive dataset likely provides a better indication as to the predictive capability of the model when examining multicomponent gases, notably natural gas. Models developed from the complete dataset are expected to achieve a more accurate regression score due to the increased number of data samples, and further assist the model in distinguishing between sI, sII or mixtures of hydrate structures due to the additional data, particularly methane-propane and methane-ethane mixtures, in ranges of conditions lacking

multicomponent equilibrium points. While the complete dataset reported accuracies include pure and binary component testing data, this score does serve to provide an upper estimate of model performance. Both datasets are used to develop models capable of being used to predict equilibrium conditions for multicomponent gases. As the complete dataset includes all equilibrium points from the multicomponent exclusive dataset, adequate cross-validation performance for both models will indicate that multicomponent data has a high likelihood of being accurately predicted by the complete dataset model.

3.6 Neural Network Model Validation

The means of selecting and validating the model is perhaps the area of this research where the most significant contribution to the field can be made. The selected approach involves applying a 10-fold cross-validation which selects train-test indices based on a random selection to optimize parameters and perform model selection, while a hold-out validation set created before cross-validation tests the cross-validated model parameter selection by testing using data completely unseen during cross-validation, thus allowing the accuracy of the final model to be assessed. This approach has been selected so as to make the most use of a dataset with a limited sample size. An illustration of the validation approach is provided in Figure 3.2. A further benefit of performing cross-validation is that the bias and variance of the model may be assessed, and the standard deviation associated with reported accuracy of the model is provided.

Through the process of back-propagation, the neural network is trained by reducing the mean squared error resulting from training data. While the final reported mean square error for this training data may be reported as an exceptionally low figure, this does not provide an indication of the capability of the model to predict unseen data. A high training score may be achieved at the cost of overfitting the model, providing artificially favourable model results at the expense of reduced generalization capability and thus ability to predict unseen data (Cawley & Talbot, 2010). In order to test the predictive capability of the model, unseen data must be used. The validation strategy used in this investigation includes a randomized hold-out validation set which is separated from the dataset before training occurs, thus separating the dataset into a training and validation set. Having trained the model, the holdout validation set is run through the model, and the error resulting from the difference between experimental and predicted results is used to calculate the final predictive capability of the model. Reported metrics for this validation includes the R^2 score of predicted data defined in equation 4.1, and the mean

absolute percentage error. In addition to performing hold-out validation, cross-validation is performed. Cross-validation is employed to gauge whether overfitting has occurred, and provides an indication of how the model performs under a range of conditions. Cross-validation additionally allows for a variance to be attributed to the final score of the model. Under cross-validation, the training dataset which is used to develop the neural network itself is further split into training and testing data. Cross-validation is performed over 10 folds, and provides a final R^2 score in addition to a standard deviation which serves to highlight potential outliers or ranges of conditions which lack a significant quantity of data needed to make accurate predictions. While the hold-out validation score provides an upper average of the model's capability over a wide range of conditions, cross-validation provides an indication of potential shortcomings of the model regarding the ability to predict conditions over the entire range of data provided.

The hold-out validation set is created by sampling data from the dataset prior to training the model, and is completely unseen by the neural network during training. In order to ensure the hold-out validation set covers a wide range of conditions, creation of the hold-out set is achieved through means of stratified sampling. Due to the non-Gaussian nature of the both datasets required in model development, and the limited quantity of data available, randomly sampling a number of equilibrium points is unlikely to cover an adequate range of equilibrium conditions needed to assess the capability of the model to predict a universal range of conditions. While a degree of randomization when sampling data for the validation set is required to avoid biasing results, performing the traditional approach of randomly selecting equilibrium points to hold out is not viable; as such, groups are created in the dataset according to the criteria of equilibrium measurements obtained using the same experimental equipment, methodology and means of measuring the equilibrium conditions. Thus, groups are, for the most part, created according to equilibrium measurements reported per publication. Due to the long experimental times encountered for multicomponent gas hydrate equilibrium measurements as a result of metastability (Ward, 2015), many publications from which data was sampled for this investigation report a narrow range of conditions, thereby rendering the compiled datasets highly suited towards sampling data per group when developing holdout sets. A consequence of grouping data by source is a highly uneven number of equilibrium measurements per group, as illustrated for both datasets in Figure 3.1a and Figure 3.1b. This behaviour eliminates the possibility of using traditional sampling techniques which are centred on sampling an equal number of points from each group, or the approximate uniformity of group sizes. Therefore, the conclusion has been reached that the optimal approach to generating

a universal-approximating holdout set is through use of a randomized stratified split technique. Through stratified sampling, items from each group are sampled such that the original proportion of entries per group remains approximately constant. Hence, large groups will have more equilibrium points sampled than smaller groups. Furthermore, randomization is added to the sampling process through randomly selecting the entries which are sampled from each group. Overall this process allows a wide range of data to be tested for a limited dataset, without requiring a large number of samples to be split into the holdout set relative to the size of the overall compiled dataset. The grouping of data by source is a feature which is additionally utilized during cross-validation for several models developed, and groupings are carried over into the training dataset for the models which require this information. For the models developed in this investigation, all models employing a hold-out validation split follow a target of 10% of the original dataset being used to develop the hold-out set. The reason this figure is low is due to the employment of cross-validation in model development.

Having divided the dataset into a training set and a hold-out validation set, it is necessary to elaborate on the testing process of the model which is performed by cross-validation. While hold-out validation allows a definitive model to be developed from the entire training set, hold-out validation in prone to a high variance in reported accuracy. This is due to the random selection of data to be held out not necessarily being representative of the entire range of the dataset, caused by the datasets being non-Gaussian. While a stratified split assists in mitigating this factor, due to the relative size of the holdout set only being a fraction of the complete dataset, there is a significant variance associated with the reported accuracy of holdout validation when altering the randomization of the data split. The non-Gaussian nature of the compiled dataset further compounds this problem. While a larger validation set would allow for a more confident reported model validation accuracy, this would come at the expense of having less data available to train the model. In order to train the model with as much data as possible, while judging the variance in results, cross-validation is employed. Cross-validation allows the bias and variance of the model to be examined, and certainty that the model is not merely stable for a favourable data split can be provided. Hence, cross-validation additionally allows for the parameters of the neural network model to be tuned to improve the predictive capability of the model. Cawley & Talbot (2010) however warns that hyperparameter optimization paired with model selection does present the risk of overfitting occurring, whereby information on the validation set leaks into the network. In order to facilitate

hyperparameter tuning while attempting to reduce the risk and impact of overfitting, hold-out validation has been employed to further test the model after model selection.

While cross-validation does involve further dividing the training data according another train-test split, all training data may be used after cross-validation has been performed to generate a final model which tests held-out validation data and thus yields vital insight into the stability of the model without compromising the quantity of training data. In order to incorporate cross-validation into the model, the training set which is created by splitting the dataset into a training set and hold-out validation set, is further split into a train-test set. Unlike the hold-out validating set, cross-validation allows for multiple train-test splits to be performed. For this investigation, all models undergo 10-fold cross-validation, whereby 10 randomized train-test splits are generated, and a model is trained and tested for each. This yields 10 reported regression scores, which are then averaged and the standard deviation is calculated so as to assess the bias and variance of the model. Unlike the hold-out validation split, the train-test split does not follow a stratified sampling procedure. Rather, two different approaches have been applied for the various models developed in this investigation. The first approach which is used in models F and G involves simply selecting random samples to be divided into a train-test split according to a pre-defined ratio, in this case 30%. The second approach to generating the train-test split for cross-validation is achieved through a randomized group split, and is employed in models A, B, C, D and E. The group split is performed by separating entire groups of data, such that the same group is not used for both training and testing the model during cross-validation. This grouped approach allows the dependence of the model on individual data sources for certain ranges of conditions to be assessed, and serves to mitigate some of the bias of experimental or measurement errors from the equilibrium data on the reported accuracy through preventing experimental factors such as equilibrium conditions measured when inadequate time was given to account for metastability from achieving high prediction scores due to the presence of some of this group's data in the data used to train the model. An example of both cross-validation strategies is shown in Figures 4.3a-f, where for Figure 4.3e & Figure 4.3f, cross-validation is performed sampling random indices, and in Figure 4.3a-d, cross-validation has been performed through the group split technique. After having performed cross-validation, the entire original training set develops a model which is tested with the held-out set. This practice allows for the model parameters to be tuned to yield more effective predictions with a significantly reduced risk of overfitting the model, which would serve to reduce the reported hold-out validation score. Overfitting results when tuning a model that may even include cross-validation,

increases reported accuracy while reducing the accuracy of predicting unseen data, due to information on the testing set leaking into the neural network through parameter tuning and model selection (Cawley & Talbot, 2010). Through combining cross-validation and hold-out validation, confidence is granted that overfitting highly unlikely. Note that model selection based on hold-out validation results has been limited to avoid leaking hold-out validation data, as opposed to hyperparameter optimization whereby model selection is performed from a wide range of models.

In summary, the model parameters are selected by a process of 10-fold cross-validation with a 70%-30% train-test split (80%-20% in the case of model A), and then validated for performance on a truly unseen data set which was randomly selected prior to cross-validation in a 90%-10% train-validation split, where the 10% represents data which is complete held-out from the cross-validation procedure. An illustration of how validation and model selection is performed is provided in Figure 3.2. This process allows for extensive insight to be gained from a model trained and tested using a dataset with a limited number of samples available. While larger datasets would allow for more accurate models, the validation practice employed ensures that overfitting has not occurred and confirms that the model accurately predicts unseen data. Overall this method when combined with the variance yielded after cross-validation, and the seed-test results provided in Figure 4.2a-d allows for a highly transparent validation procedure, with a high degree of confidence that reported results can be expected to represent those of the practical application of the model.

3.7 Model Development

In this investigation, Neural Network models have been designed as multi-layer perceptrons. The models are trained in a supervised manner through use of datasets containing equilibrium data, which are trained by means of a backpropagation algorithm. Due to the significant non-linearity of the data, neural networks are well suited to model the conditions of gas hydrate equilibria. Leshno et al. (1993) details that a multilayer feedforward network is capable of acting as a universal approximator, which is the end goal of this research.

Models A through G have been developed as artificial neural networks designed to predict the hydrate equilibrium pressure in MPa for the inputs of gas phase composition as a molar fraction and temperature in degrees Celsius. All neural networks have been developed as a feedforward Multilayer Perceptron, wherein one or more hidden layers containing nodes linking model input to output are present. Supervised learning occurs, and model training is achieved by

means of a back-propagation. All neural network models have been developed using the Python infrastructure and libraries developed for use with Python, specifically the keras library running on a tensorflow backend. While each neural network developed for models A through G differs in terms of topology or parameters, the generic diagram in Figure B.1 provides an overview as to the relation between the input, hidden and output layers. In this investigation, several different network topologies are used, as indicated in Table 3.4. Deep learning has been implemented, with a minimum number of two hidden layers for neural network models. For each case, nine inputs are present corresponding to Table 3.1 excluding pressure, which is the output. Each of these input nodes connects to each neuron present in the subsequent hidden layer, each node of which connects to all nodes of the following layer, be it another hidden layer or the output. As a regression is being performed, the model output consists of a single node, and outputs the numerical value of gas phase pressure. Neuron training weights have been initialized by means of a normal distribution.

Table 3.4: Summary of Models and Results

Model	Dataset	Hold-out	Group	Hidden Layers	Neuron Count	Activation	R^2 (CV)	σ (CV)	R^2 (Holdout)
A	Multicomponent Exclusive	True	True	2	64	ReLU	0.90622	0.05410	0.95114
B	Complete	True	True	2	256	ReLU	0.96032	0.01811	0.98564
C	Multicomponent Exclusive	False	True	3	256	ReLU	0.91554	0.00450	-
D	Complete	False	True	3	256	ReLU	0.95219	0.03062	-
E	Complete	False	True	2	128	Sigmoid	0.94703	0.04610	-
F	Multicomponent Exclusive	True	False	3	352	ReLU	0.97067	0.01135	0.97870
G*	Complete	True	False	3	352	ReLU	0.98604	0.00352	0.99255
H**	Multicomponent Exclusive	True	False	0	-	-	0.60671	0.02469	0.68302
I**	Complete	False	False	0	-	-	0.87105	0.05564	-
J	Complete	True	True	3	352	ReLU	0.95814	0.02154	0.98918

*: End product of this research

**: Simple polynomial regression

As discussed, both hold-out validation and cross-validation has been implemented in model development. For models employing hold-out validation, a pre-defined split of 90% of the original dataset is used to compile a training set, while the remaining 10% of data forms the hold-out set used to test the model post-training. The hold-out validation split has been performed in a stratified manner such that the approximate proportion of indices per group present in the dataset is preserved across both the training and validation set. In order to ensure the replicability of results, the function generating the stratified split indices has been seeded such that the randomly selected indices present in the training and hold-out validation set are constant over successive models. Regarding cross-validation, in which the training set

following the split from hold-out validation is further split into train and test sets, all models incorporating cross-validation have been implemented through a 10-fold split. Under this practice, 10 different splits of training and test data are used to develop models and the final accuracy is taken as the mean of each of the 10 fold results, with the standard deviation being yielded. As to how the 10-fold split occurs, two different mechanisms have been developed for determining the train-test split for various models. Models A, B, C, D & E undergo a randomized group split, whereby data is divided into separate train and test sets such that data from the same group is only present in one of these sets. Under this approach, approximately 70% of the data forms the training set, while approximately 30% forms the test set, with the exception of Model A which utilizes a 20% test set due to the limited number of groups in the multicomponent exclusive dataset. Models F and G employ a standard randomized split, whereby exactly 30% of randomly selected indices from the sampling dataset is used to develop the test set, while the remaining 70% is used to train the model. In the case of the group split, as indicated in Figures 3.1a and 3.1b, only a limited number of groups of uneven size are present, as such an exact 70%-30% train-test split is unlikely, as opposed to the ungrouped randomized approach in which an exact 70%-30% split is ensured. In either case, 10 different random train-test split sets are used to develop 10 distinct models. For each model, training is performed using the train set, while the test set is run through the trained model and evaluates the overall disparity between predicted and experimental gas phase pressure. For each of the 10 folds, the testing score is recorded, and an overall score is yielded by averaging these results, and calculating the standard deviation. As with creating the hold-out validation set, both the randomized grouped and ungrouped split methods are randomly seeded, such that the same random segmentation of data into train and test sets for each fold is constant for repeated model executions. The seeding of model randomization allows the effect of different model parameters on results to be assessed, and is thus a foundation of hyperparameter optimization, the process by which a wide range of model parameters are investigated so as to achieve an improved result.

Having performed cross-validation, models which implement hold-out validation then train the neural network using the entire original training set which was obtained prior to performing cross-validation. This model is then tested using the hold-out validation set, to yield the overall accuracy of the model when testing completely unseen data. Unlike cross-validation, this is a single fold process and yields a single real number as a result. As such, several useful metrics are yielded which can be used to evaluate the accuracy of the model. From cross-validation the

variance of the model can be assessed by investigating the variation in results for various folds containing different combinations of data used for training and testing. From this, potential weak folds can be identified and a lower estimate of model performance can be obtained. A real number corresponding to the accuracy of each fold is provided, with the average of the 10 folds provided to assess the bias, while the standard deviation across the 10 fold results being used to assess the variance of the cross-validated model. The hold-out validation procedure yields a single real number representing the accuracy of the model across the stratified range of validation data, thus providing an indication of model performance across a wide range of data unseen during cross validation, and serves as the final accuracy of the model. Accuracy for a regression model can be assessed through a number of metrics. For this investigation, the R^2 regression score has been used as a baseline for model tuning, and is taken to represent the overall performance of the model. The R^2 score is calculated, and has been selected due to its penalizing of pressure predictions on test data both greater or less than the actual test value without reducing the error metric due to overshoot and undershoot balancing. The mean absolute percentage error has additionally been calculated for several models, to facilitate comparison results with other studies which have not adopted R^2 metrics. Mean absolute percentage error metrics do exhibit considerably more variation than R^2 scores for the models developed in this investigation, thus limiting the replicability of their results. As such, R^2 metrics have been used as the primary means of scoring the models developed. A variance plot of cross-validated results obtained for each model illustrates the effect of altering the random state of train-test split indices, while seed plots such as Figure 4.2a-d illustrate the fact that the model does not merely perform in the range of reported results for a favourable train-test split. The split of data for both grouped and ungrouped cross-validation train-test split configuration is illustrated in Figure 4.3 a-f. Plots of predicted pressure vs experimental pressure are additionally used to indicate potential outliers, while serving to illustrate potential weak ranges of the model, and are provided in Figure 4.5a-d.

Model development has been performed using Python (v. 3.6.8) and the Keras (v. 2.2.4) library running on a Tensorflow (v. 1.12.0) backend to develop the neural network and train the model. Other libraries utilized in model development include scikit-learn (v. 0.20.2), Pandas (v.0.24.1) and Numpy (v. 1.15.4). Matplotlib (v. 3.0.2) is a library, which has been used to develop the plots and illustrations in Figures 3.1a and 3.1b, all plots including and between Figures 4.1 to 4.5, and the appendix Figure B.1 and Figure C.1.

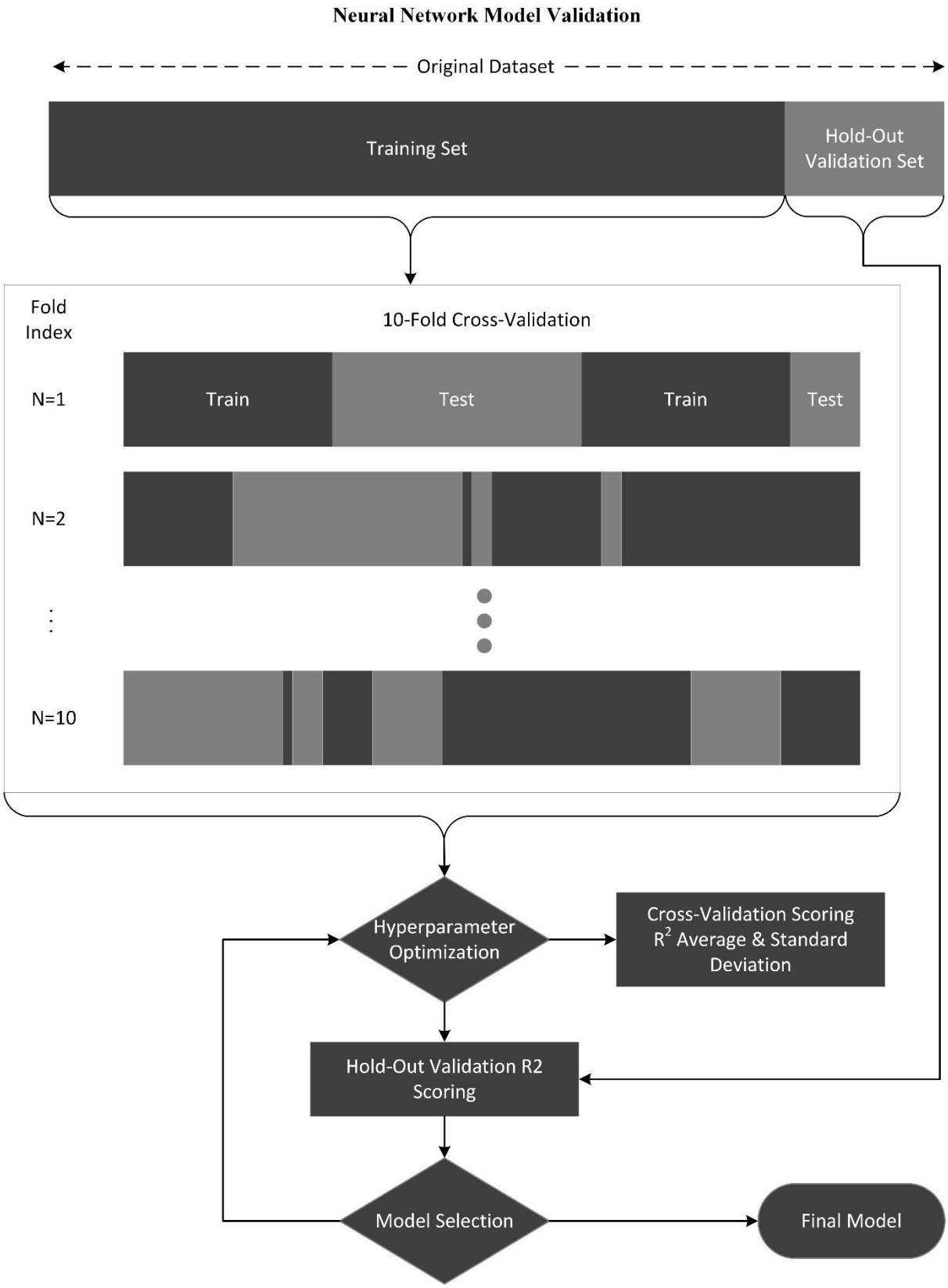

Figure 3.2: Illustration of the model validation strategy. The dataset is divided into separate indices used for training and hold-out validation in a stratified manner. The training indices are further sub-divided for cross-validation into train-test indices over 10 folds. The algorithm for splitting cross-validation indices depends on whether or not grouping is being investigated. In the absence of grouping, indices are divided randomly to form sets of pre-defined sizes. If grouping is investigated, entire groups are divided such that the same group cannot be used to both train and test a fold. Note limited selection has been performed based on hold-out validation results to avoid overfitting.

3.8 Hyperparameter Optimization

A significant risk is present when attempting to optimize model hyperparameters in that overfitting is possible. Cawley & Talbot (2010) elaborates that whenever model selection is performed over a limited dataset, overfitting is a risk, whereby iterating hyperparameters risks improving model results by providing a regression highly specific to the sample data while lowering generalization ability across unseen data. A significant advantage to the incorporation of both hold-out and cross-validation is that optimization of the model is facilitated while lowering the risk of overfitting. By separating the dataset into train and validation sets prior to cross-validation being performed allows for the model to be tested post-tuning to truly unseen data. Thus, the model parameters may be iterated to yield improved results for both cross-validation testing and hold-out validation without significant risk of overfitting occurring unnoticed. Seed tests being performed on models further lends confidence to the model results in that surety is provided that results are only valid for a favourable train-test split. Model parameters have been tuned in a hyperparameter optimization approach, using a grid-search iterative technique to achieve a model accuracy close to the optimal result. Due to the computational times involved in training highly non-linear, multi-layered neural network regression models, the optimal neural network could not be determined. As part of an iterative procedure, a wide range of conditions for each parameter has been specified before training, as such the true optimum cannot be obtained. Additionally, variance in the results obtained when training successive models may result in the highest scoring model alternating between several different parameter combinations. Nevertheless, an iterative approach has allowed for significant improvement to results, and likely yields a result close to optimal.

Hyperparameter tuning has been performed on several parameters, including: The number of hidden layers and neurons, the activation function of the hidden layers, the optimizer used to minimize the training loss function, the training batch sample size and the number of epochs over which the model was trained. The tuning was performed by means of a grid-search procedure, and iterating over a wide number of parameters by means of pre-defined values in a range. This procedure involved cross-validation yielding the final performance metric, and 10-fold validation was performed for each combination of parameters. The output of the grid-search was the mean cross-validated R^2 score and standard deviation. Models with the best R^2 score and a sufficiently low variance were selected as potentially the best model, and were further tested by performing hold-out validation in addition to cross-validation over a variety of different random seeds to ensure results were not specific to a favourable train-test split.

Finally, the best model was selected based on the criteria of adequately low cross-validation variance, and a high scoring hold-out validation test.

The hidden layers of the neural network serve to greatly increase computational time, while allowing more complex behaviour to be modelled through the activation functions of hidden layers. The presence of hidden layers allows non-linearity to be introduced into the model by means of activation functions (Leshno et al., 1993). As such, models have been tested using configurations with no hidden layers, thus yielding a baseline for minimum performance. A neural network comprised of no hidden layers, and only an input and output layers functions as a linear model. Models H & I are such models, bearing limited polynomial functions which poorly model the system behaviour, as indicated in Table 3.4. Optimal neural network configurations have been found to consist of either two or three hidden layers, and have the advantage of training in an acceptable time frame. It is expected that at least one hidden layer is required due to the high non-linearity of the dataset resulting from the significant differences in equilibrium conditions arising from sI and sII hydrates of similar composition that may be present in the training set. During the tuning of the number of neurons per hidden layer, due to training time a constraint has been specified such that the number of neurons is constant across each of the hidden layers. As such, results in Table 3.4 specify the number of neurons present in each hidden layer.

The role of the activation function in a neural network, is to govern whether or not a neuron is turned on and transforms incoming input during the training process, where a deactivated neuron does not alter an input, while an activated neuron transforms input according to a pre-defined function. Leshno et al. (1993) details that any continuous function can be modelled by a non-constant activation function, provided the activation function is not a polynomial. As a regression is being performed, each model constitutes a single neuron in the output layer which corresponds to the predicted pressure. In all neural network models with the exception of Model E, a linear activation function is used for the output layer, so as to provide a continuous, real output. Model E achieved better results using the softplus activation function as opposed to linear, due to its use of the sigmoid activation function governing hidden layer output. Regarding the activation function of the hidden layer, as the dataset is highly non-linear, a non-linear activation function is required. Utilizing a linear activation function in the hidden layers would merely result in a linear approximator being developed, unable to accurately predict the hydrate equilibrium conditions. As such, several non-linear activation functions were selected to be iterated as part of the grid-search procedure. Amongst these non-linear activation

functions, are the rectified linear unit (ReLU) and sigmoid activation functions. Glorot & Begio (2010) discusses the difficulty of implementing the sigmoid activation for deep-learning application. The ReLU hidden layer activation function successfully introduces non-linearity into the system and allows complex behaviour to be learned (Maas et al., 2013). A constraint on the selection of hidden layer activation functions has been made, in that the activation function selected is constant across all hidden layers. The following equations represent the activation functions used to develop the neural network models in this investigation. In most cases, ReLU activation is applied to hidden layer activation, while Linear activation is applied to the output layer as a continuous output is desired. With the exception of linear activation, which does not transform the input into a neuron, activation functions in a deep learning context are provided and discussed in Nwankpa et al. (2018).

$$Linear: f(x) = x \tag{3.2}$$

$$Softplus: f(x) = \ln(1 + e^{x}) \tag{3.3}$$

$$ReLU: f(x) = \max(0, x) \tag{3.4}$$

$$Sigmoid: f(x) = \frac{1}{1 + e^{-x}} \tag{3.5}$$

The purpose of training the model is to lower the loss function, thus ensuring an improved correlation between the model and training data. In order to achieve this, the model is trained according to an optimizer which serves to minimize the loss of the model when comparing to training data. The optimizer should be selected such that the global minima, rather than some local minima is obtained when seeking to minimize the loss. Optimizers function by minimizing a cost function during training (Kingma & Ba, 2014). Highly popular optimizers include the stochastic gradient descent (SGD) optimizer, which has achieved numerous successes in machine learning and is widely used in machine learning (Hsueh et al., 2019; Kingma & Ba, 2014). The SGD and similar approaches however do include the learning rate of the optimizer as an additional hyperparameter, which adds to the extensive list of parameter iterated during the grid-search procedure. Learning rate schedulers such as the step decay method, which adjusts the learning rate after a number of epochs, could serve to reduce the number of hyperparameters (Hsueh et al., 2019). An alternate to this approach is to utilize some of the newer optimizers which incorporate adaptive learning rates. Several optimizers have been investigated during model development, and in the majority of cases the best results during hyperparameter optimization has been achieved by means of the Adam optimizer, a

gradient-based optimizer which is discussed and introduced in Kingma & Ba (2014). As an adaptive learning rate optimizer, Adam allows for fewer hyperparameters to be iterated, and thus simplifies model selection. Due to success using this optimizer, and the reduced number of hyperparameters requiring tuning, all models listed in Table 3.4 have been developed using the Adam optimizer.

3.9 Model Description

During the course of model development, it became clear that multiple models would be required in order to satisfy the research objective. In order to judge the effect of excluding a large quantity of non-multicomponent data from model development, two neural networks have been developed. One neural network is trained and tested using exclusively multicomponent data, while another is trained and tested using a wide-ranging dataset including hydrate equilibrium data for pure methane and methane binary gas phase mixtures. Developing models from two different datasets provides an indication as to whether or not sufficient multicomponent data is available to reliably predict unseen data when comparing the results of both models. Additionally, proving the model yields adequate results when trained with a multicomponent exclusive data will indicate a lower-estimate for the multicomponent predictive capability of a model trained on the complete dataset. Having identified that two separate models will be required to accommodate the different datasets, the impact of including grouping in model training and testing is also to be assessed. As discussed, grouping of data by source has been performed so as to facilitate a stratified split for a separate hold-out validation set in order to ensure a wide range of data is validated. A benefit of this grouping configuration is that the dependency of the model on individual sources of experimental equilibrium data on certain ranges of conditions can be assessed. The heavy reliance of the model upon individual sources of data to cover specific ranges of data is often unwanted, particularly when considering experimental sources of data, which possibly used outdated measurement practices for identifying the equilibrium point, or may not have adequately accounted for metastability. While the goal of the data sampling campaign was to gather an abundance of equilibrium measurements over a wide range, thus rendering factors such as this insignificant, the limited number of independent groups results in the grouping practice being a necessary supplement to model development. As such, for each dataset two neural networks are developed, one where cross-validation is performed by grouping data, and another where grouping is not considered during cross-validation. It is expected that ungrouped cross-validation results will be significantly higher than grouped results, due to a relative shortage of

multicomponent gas hydrate equilibrium data over several ranges of conditions. Each of these 4 models discussed employ both hold-out validation and cross-validation. Models A, B, F and G cover these cases, and are summarized in Table 3.4. In addition to these 4 models, additional neural networks will be developed without validating with a hold-out dataset, and exclusively using cross-validation to assess model performance. Developing two models for each dataset where grouping is performed without a hold-out validation set allows for 10% more data to be included in the training set, and while parameter tuning cannot be performed so as to attempt to optimize these models, an indication of model cross-validation performance when the training set size increases by a relatively small amount. Models C and D cover these cases. Finally, due to the abundant use of the sigmoid activation function for hidden layer activation by numerous similar studies in the field, and a relative lack of studies in the field employing the Rectified Linear Unit (ReLU) activation function for hidden layer activation, a final additional model is trained using the Sigmoid activation function so as to indicate similar performance between the model employing sigmoid activation, and another employing ReLU activation. This model is developed in Model E. Finally, in order to facilitate comparison between Model B and G which have been developed with different parameters, Model J has been trained using all the same parameters as Model G with the exception of performing grouped cross-validation.

In addition to developing neural network models for the compiled datasets, simple polynomial regressions have additionally been developed for comparative purposes, and to serve as a lower baseline for model performance. These models have been developed by fitting a polynomial of a specific degree to the data provided. Naturally these linear models are expected to perform significantly worse than neural network models due to inability to distinguish between sI and sII hydrates in many cases. Models H and I cover these cases.

Several models have been developed so as to provide a comparative basis of results using the same sets of data. This approach has been selected due to the datasets being used in this investigation significantly differing from others in the field used to develop similar neural networks. In order to definitively assess model results, industrially applied methods tested under the same datasets used in this investigation will provide a good basis for comparison. While assessing industrially applied hydrate prediction methods, particularly computerized models, falls outside the scope of this research, a comparative study is certainly viable due to the large size of the datasets provided and could be investigated in future studies to further assess the viability of the neural network methodology on a wider scale in terms of regressions

performed in the chemical and petroleum industries. Several studies such as Ballard & Sloan (2004) have assessed several industrially popular software models using limited multicomponent and natural gas hydrate equilibrium data. Due to the significant differences between datasets used in this research and other studies, extrapolating model accuracies from other works cannot supplant actual testing industrial models with the datasets used in this research. Based on preliminary comparison with other studies such as Ballard & Sloan (2004), results do suggest that the neural network models developed in this research, Particularly Models F and G, do fall within an acceptable range of accuracy, and further investigation into comparison between popular software methods and these neural networks is worthwhile.

CHAPTER 4 RESULTS AND DISCUSSION OF MODELS

Results of neural network models have been assessed according to cross-validation bias and variance, in addition to the validation error of testing the final model using the hold-out validation set. A sensitivity analysis by means of an iterative grid-search procedure has been used to select model parameters, and the best performing model has been selected based on the criteria of hold-out validation error and cross-validation performance, specifically in terms of variance across 10 folds. The primary scoring metric used in this study is the coefficient of determination, or R^2 score. The results for all models developed are available in Table 3.4

$$R^2 = 1 - \frac{\sum\left(y_{true} - y_{predict}\right)^2}{\sum(y_{true} - \bar{y}_{true})^2} \tag{4.1}$$

As 10-fold cross-validation is performed, and cross-validation seed tests have been developed, it is possible to visualize the results. Figures 4.1a & 4.1b plot the 10-fold cross-validation variance in the form of a box-plot. As only 10 folds are tested, the error bars for the box plot extend to the minimum and maximum result, rather than the traditional approach of extending error bars up to the second standard deviation. This approach is taken due to the intention of the diagram to illustrate the variance between cross-validation folds, in an environment where the presence of poor-scoring folds is a highly significant statistic. Figures 4.2a-d plot the effect of altering the random seeding of the train-test split for cross-validation, a useful means of indicating the highly non-Gaussian nature of the dataset, while providing a means of ensuring that the model has not simply been trained to accurately model a convenient train-test split seeding. Seed tests have been included where cross-validation on the testing fold exhibited a low variance. Seed tests for models exhibiting a high cross-validation variance have not been included, as their purpose is to illustrate low variance models are not only valid for a favourable train-test split.

Cross-validation and hold-out validation results for all models are summarized in Table 3.4, while cross-validation variance is illustrated in Figure 4.1a & Figure 4.1b. Seed-Tests are presented in Figures 4.2a-d to indicate the non-Gaussian complex nature of the dataset, while Cross-validation indices have further been illustrated in Figures 4.3a-f to illustrate how train-test indices are randomized despite grouping occurring. Figures 4.4a & 4.4b provide the hold-out validation indices for models applying hold-out validation. The purpose for including these indices is to illustrate the stratified manner in which data is divided into training and validation sets.

Before discussing model results, it is important to clarify that several factors have necessitated the development of multiple models. Model G is trained and tested using the complete dataset without grouping restrictions, and serves as the model which would be used to predict real natural gas hydrate equilibria for practical application. Information regarding performance for exclusively multicomponent data, and the dependence of the model on individual sources of experimental data however, cannot be obtained from model G alone. For this reason, separate models have been developed to assess these factors and to later be used to conclude that the results model G likely indicates actual performance on unseen data, and that the model could viably be used in practical application, after testing using specific cases applicable to the application. The following section will appear rather verbose due to the number of factors used to assess each model, and as such routinely refers to each model by letter. A summary of each model developed is provided in Table 3.4. For shorthand reference, models A and B assess model dependency on individual sources to cover equilibrium conditions where little overlap between independent sources is present, models C, D, and E supplement models A and B by excluding hold-out validation and thus injecting slightly more data into cross-validation. models F and G serve to provide final results for each dataset. Comparison between models in each category is performed to prove that complete dataset models accurately predict multicomponent data. In order to facilitate comparison between models B and G, model J was developed using the parameters and topology of G. Similar results between models J and B result in further discussion of unoptimized model J being unnecessary. The sole purpose of model J is to ensure that grouped validation does not experience overfitting when using the higher neuron count of G. Model G is regarded as the best model developed in this investigation, and serves as the proposed universal model. All other models presented have been designed to lend further credibility towards model G, and will be discussed extensively.

As expected, polynomial regression models listed in Table 3.4 prove vastly inferior to the developed machine learning models. This further serves to illustrate the highly non-linear and multimodal nature of the equilibrium datasets, which a linear model proves incapable of adequately accounting for.

As several models have been developed, it is necessary to first compare model results according to dataset before discussing trends. Observations arising when comparing multicomponent exclusive models A & F are similar to the those between results of Complete Dataset Models B & G, and as such parallels can be drawn between models. As the purpose of this investigation is to develop a natural gas hydrate prediction model, detailed discussion regarding the

differences between multicomponent exclusive models as a result of grouping is held, and the similar behaviour between Complete dataset models will be noted by drawing on this discussion.

As discussed, it is expected that grouped cross-validation results be worse than ungrouped results. On examining the results of neural network models A, C & F, which are trained using the multicomponent exclusive dataset, it can be seen that there is a significant difference in results between grouped and ungrouped models. Model A has been developed to incorporate a 10% hold-out validation split, and a 20% cross-validation train-test split, thus facilitating a grid-search sensitivity analysis. As detailed in Table 3.4, the model yields a coefficient of determination of 0.9062, a standard deviation of 0.0541, and a hold-out validation R^2 score of 0.9511. The cross-validation results prove unexpectedly low, likely due to the randomized group split practice being implemented for 10-fold cross-validation. This result starkly contrasts the results Model F, similarly incorporating a hold-out validation split of 10% and a cross-validation train-test split of 30%, but performs cross-validation without regard to grouping. Model F achieves a cross-validation R^2 score of 0.9707, a low variance of 0.0114 and a hold-out validation score of 0.9787. As such, it can be seen that model A, through the practice of grouping cross-validation test indices, achieves a significantly worse R^2 score, indicated by a cross-validation R^2 score lower than model F by 0.064, with a significantly higher variance. As the final validation score from the hold-out dataset is calculated by using the same indices across models A, C & F, one may expect the hold-out validation R^2 score to be approximately the same across model. This however is not the case, as Models A and F have undergone a grid-search procedure to yield close to optimal parameters based on both hold-out and cross-validation scores, thus the disparity between validation scores is expected. Thus it can be concluded Model F is clearly the best model of those developed using the multicomponent exclusive dataset. As such a significant disparity of cross-validation results is present between models A and F, the cause is worth investigating so as to gain further insight from the grouping of cross-validation indices. Cross-validation for Model A is performed according to a grouped practice, which randomly selects entire groups to be used for testing a specific cross-validation practice. Unlike Model F, developing a convergent model A using a cross-validation train-test split of 30% proved unsuccessful, with average cross-validation R^2 scores below 0.9, the bare minimum considered by this investigation. As such, a 20% train-test split was imposed for selecting groups being separated into train and test sets for this model. While lowering the amount of testing data allowed a marginally larger training set, it is likely

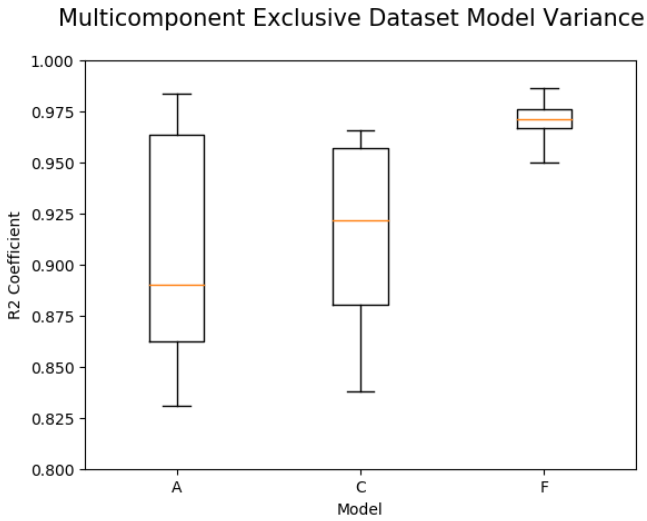

Figure 4.1a: Multicomponent exclusive dataset model cross-validation variance

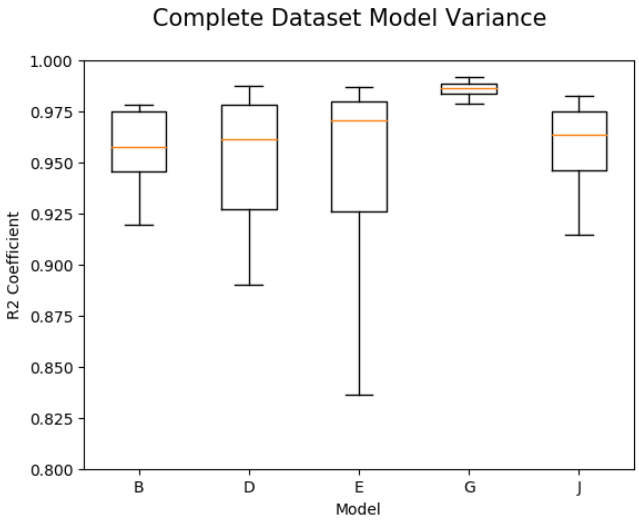

Figure 4.1b: Complete dataset model cross-validation variances: Model G selected as optimal model in this investigation due to low variance and mean R^2 score

that insufficient independent data groups are present to adequately execute the grouped-randomization split, due to a lack of multiple sources covering overlapping ranges of equilibrium conditions. This indicates that it is possible that the model is highly dependent on individual data groups to cover wide ranges of conditions. In order to test the validity of this claim, Model C was developed to provide a slightly larger amount of data available to cross-validation by not implementing a hold-out split. As such, the entirety of groups provided in the multicomponent dataset are present. As a hold-out validation set is not present, this model cannot undergo an optimization attempt, instead a few widely different parameter values have been selected to determine a favourable model as opposed to the highly encompassing grid-search. Using a train-test split of 30%, grouped cross-validation yields highly similar results to model A. Model C results in a minor improvement to the R^2 Score of Model A, a 0.0093 increase, with a slight reduction in the standard deviation by 0.0091. This improvement in results is rather insignificant, despite the 30% train-test split being facilitated. While an extensive grid-search could not be performed for model C, even an unoptimized configuration for model F provides significantly improved results.

A lack of improvement to the cross-validation results of model A through the slightly greater in size dataset of Model C can be seen. This allows the elimination of the possibility that model A yields poor results simply because a very unfavourable hold-out validation split has been made. Thus, it can be seen that establishing a model capable of predicting the gas hydrate equilibrium pressure with a high accuracy is very unlikely when using the multicomponent exclusive dataset and testing the model by a randomized grouping cross-validation procedure.

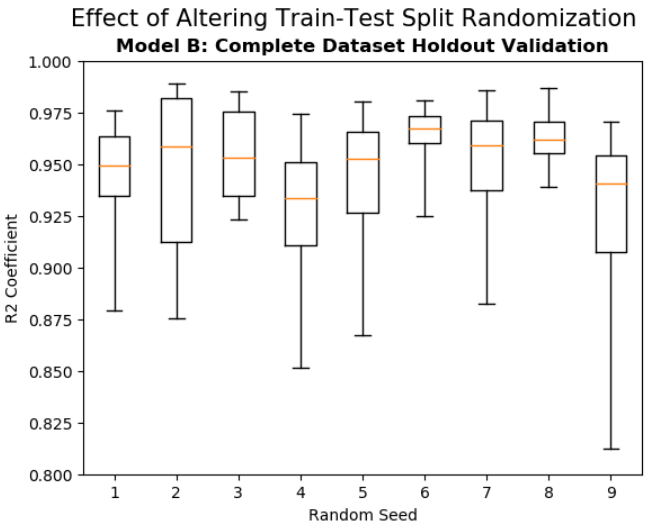

Figure 4.2a: Cross Validation Seed Tests for Model B: Note reasonable interquartile ranges. Individual weak folds due to lack of overlapping data groups

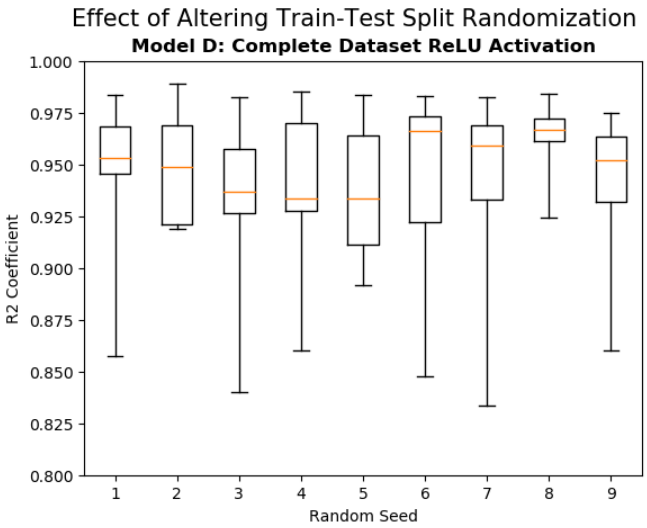

Figure 4.2b: Cross Validation Seed Tests for Model D: Unoptimized version of Model B with additional data passed to cross-validation through lack of a hold-out set

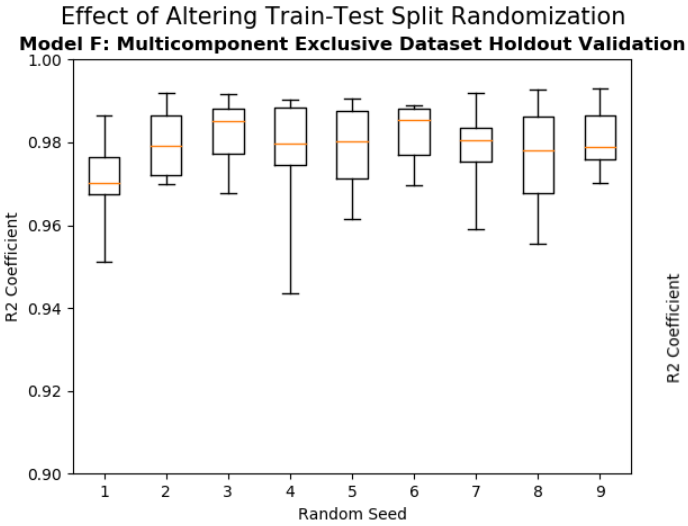

Figure 4.2c: Cross Validation Seed Tests for Model F: Proves low variance solutions are possible with a model trained and tested only using multicomponent data

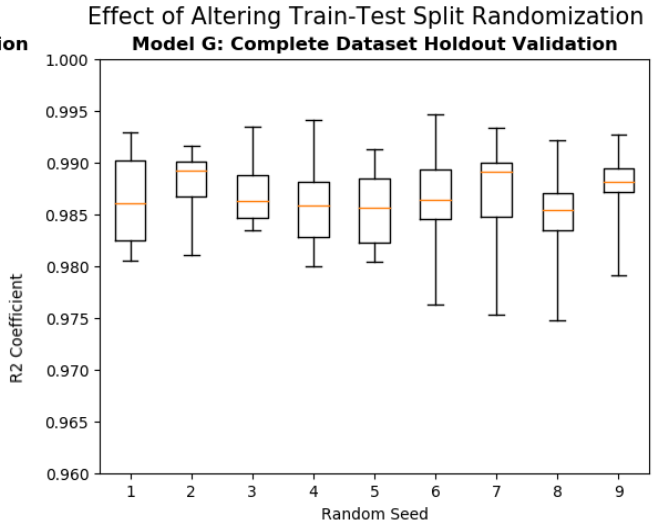

Figure 4.2d: Cross Validation Seed Tests for Model G: Primary model selected in this investigation, yields the lowest variation with randomization changes.

As the possibility that model A is simply underperforming due to an unfavourable hold-out validation split, or an unfavourable cross-validation train-test split has been eliminated, it can be concluded that the superior performance of model F is due to the ungrouped cross-validation being performed. Model F applies cross-validation by randomly selecting 30% of the training data after the original dataset is split into a training set and a held-out validation set, to be used in each fold to test the model. As data is divided between train and test sets in a completely random manner, there is expected to be a significant degree of overlap between the train-test splits of different folds, however as 10 folds are being implemented, it is exceedingly unlikely that the majority of folds will exhibit a high degree of similarity. Figure 4.3e provides the indices of the training set, which is split during cross-validation, and indicates that highly

similar folds are very improbable. Nevertheless, in order to prove that the results model F are not high simply due to a favourable randomization across the 10 cross-validation folds, seed tests have been performed and are indicated in Figure 4.2c. Figure 4.2c indicates that average cross-validation performance is approximately constant across several different randomization seedings, with the lowest single-fold R^2 score exceeding 0.94. Figure 4.2c additionally serves to prove that the model is not overfitted, while the hold-out validation R^2 score of 0.9771 lends further legitimacy to this claim. As model F employs random selection of indices, and the results are a significant improvement to model A, it can be concluded that the higher accuracy of model F is due to the completely randomized train-test split of cross-validation better allowing for a more balanced distribution of equilibrium conditions across the training and test sets. This behaviour is contrasted by model A, which prevented entire groups from being used in both training and testing sets, thus withholding wide ranges of data from the training sets of various folds. This behaviour is confirmed in Figure 4.1a, where model A can be seen providing highly accurate ($R^2 > 0.98$) predictions for some folds, while a relatively large number of folds provide poor predictions of less than 0.9 and extremely poor predictions of $R^2 \sim 0.825$, while Model F displays a relatively low variance in results, and the lowest fold score exceeding 0.945. As such, in order to achieve a high scoring model developed from the multicomponent exclusive dataset while applying grouped cross-validation, a larger dataset is required.

While model F clearly is more likely to yield accurate predictions regarding the hydrate equilibrium pressure for a gas of a specified composition at a certain temperature, there are certain factors which must be considered concerning the reported model accuracy for ungrouped cross-validation. The disparity in cross-validation results between models A and F proves that the model is highly reliant upon certain sources to cover specific ranges of conditions where little overlap is prevalent between data groups. The predictions of the ungrouped trained model F are thus highly susceptible to any experimental or measurement errors, which are present in data used to train the model. As discussed, the presence of experimental or measurement error is of great concern when considering both datasets used to train models, due to the high variability in the methodologies of experiments measuring the hydrate equilibrium conditions for specific gases, and a lengthy history of data being published in the field. Verifying as to whether or not metastability effects were adequately accounted for in data sources in the field occasionally proves impossible, often due to outdated techniques of measuring the equilibrium point and the lengthy time periods over which an acceptably accurate equilibrium measurement occurs. Ward (2015) does state that equilibrium data for

components such as hydrogen sulphide reported in literature may need revision to confirm accuracy. The same may well hold true for sources of other components or mixtures using inadequate measurement techniques or dissociation times. Inadequately accounting for metastability effects does significantly alter the nature mass transfer and kinetics when considering the equilibria of a sample, thus likely altering the equilibrium conditions significantly. Due to a notable lack of multicomponent gas equilibrium measurements for a wide range of conditions, it is not possible to merely accept verifiably perfect measurements as worthy of inclusion into the datasets. Neural networks inherently require a vast amount of data to train and test, the result of an inadequate range of data being clear from comparisons between model A and F. While care has been taken to prioritise including data, which reportedly occurs at equilibrium, and not somewhere within the metastable region, it is expected that potentially a non-insignificant number of equilibrium measurements located in both model datasets could bias the model for certain ranges of conditions, resulting in equilibrium pressure predictions outside the expected margins of error. Nevertheless, model A through practice of grouping serves to provide a lower-estimate of model performance while satisfying many of the concerns raised as to the impact of experimental or measurement errors on the predictive capability of the model. With a lower estimate exceeding a cross-validated average R^2 score of 0.90, when accounting for the fact that the lower score is largely due to a lack of data covering certain ranges, it can be seen that the results model F are feasible, and the true model performance if tested with further data would likely lead to a similar result. A conclusion, which may be drawn from models A, C and F, is that a convergent, accurate model can be developed using a multicomponent dataset. This can be seen through model F yielding for hold-out validation a mean absolute percentage error of 7.563% when predicting the equilibrium pressure of a specific gas mixture for a certain temperature, a figure that is within the expected performance when considering certain established models of gas hydrate prediction. This figure is particularly significant, as the testing set exclusively contains multicomponent gases, and results are thus free from potentially being inflated by single component gases or certain binary component mixtures which may prove simpler to model than multicomponent mixtures. Having discussed the neural network models trained using the Multicomponent Exclusive dataset, parallels may now be drawn from the observations on the disparity between the multicomponent exclusive grouped and ungrouped cross-validation neural network models A, C & F, and the complete dataset models of B & G. As with the grouped cross-validated model A and the ungrouped cross-validation model F, grouped cross-validated model B reports a significantly lower coefficient of determination than ungrouped

cross-validated model G. However, the difference between results of models B & G are significantly smaller than those between the multicomponent exclusive models. Unlike the multicomponent exclusive models, notably model A, all complete dataset models undergo cross-validation using a train-test split where 30% of training data is used to form the cross-validation test set for each fold, following the hold-out train-validate split, if applied. Model B reports an average cross-validation R^2 score of 0.9603, and a standard deviation of 0.0181 for 10-fold grouped cross-validation. Hold-out validation for Model B yields an R^2 score of 0.9856. Model G reports an average cross-validation R^2 score of 0.9860, with a standard deviation of 0.0035 across 10 ungrouped folds. The hold-out validation R^2 score for model G is 0.9926. For comparative basis, the mean absolute percentage error for model G is given as 6.877%. An immediate observation arising from these results is that the models trained on the complete dataset models perform significantly better than the multicomponent exclusive dataset models in terms of cross-validation performance. As discussed, this behaviour is expected due to the significant increase in dataset size, the multicomponent dataset consisting of 670 experimental equilibrium measurements as opposed to the complete dataset including 1209 experimental equilibrium measurements. Furthermore, as pure and binary components usually prove less complex to model than multicomponent gases, model performance for the complete dataset is expected to be slightly increased due to the likelihood that the model will better predict the equilibrium conditions of pure of binary gases than complex multicomponent gases. Comparisons between Models A, C & F revealed that grouped cross-validation models were significantly affected by a lack of data covering certain ranges of conditions. Through the inclusion of pure and binary component equilibrium data, it is possible that the model is more capable of predicting multicomponent ranges where only a few if more than one source contains overlapping conditions.

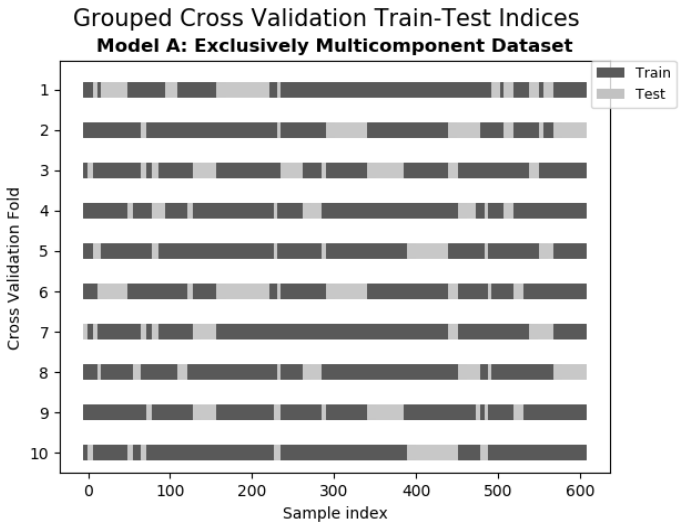

Figure 4.3a

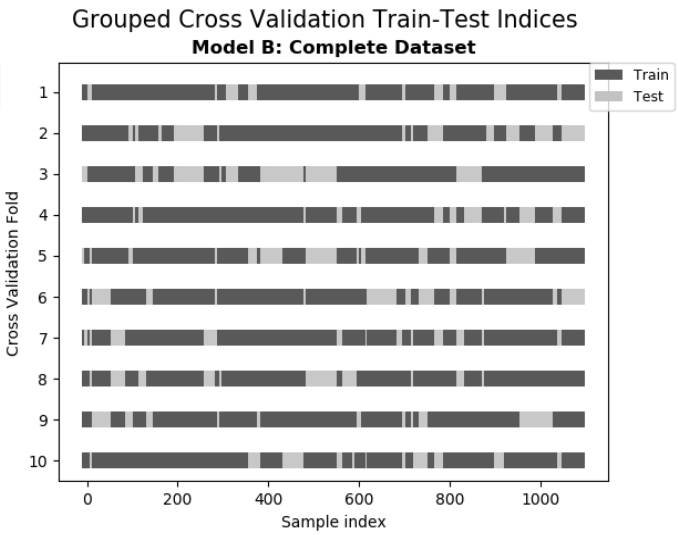

Figure 4.3b

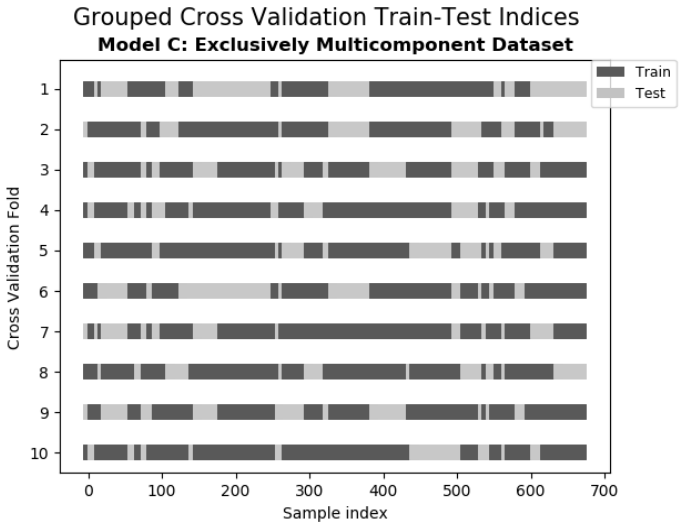

Figure 4.3c

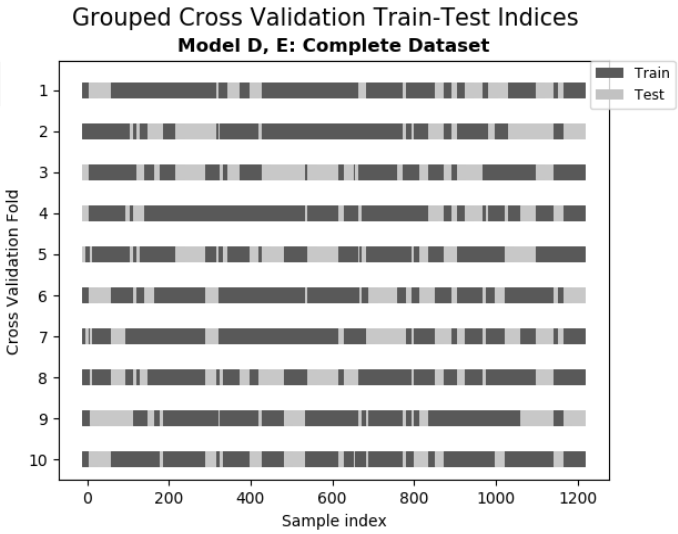

Figure 4.3d

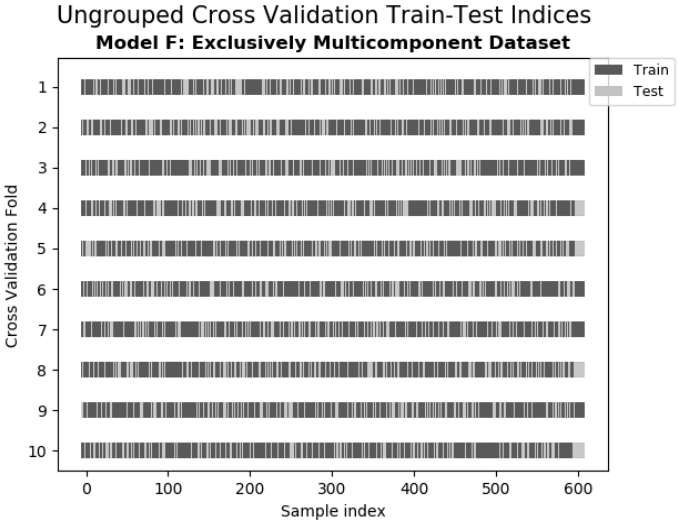

Figure 4.3e

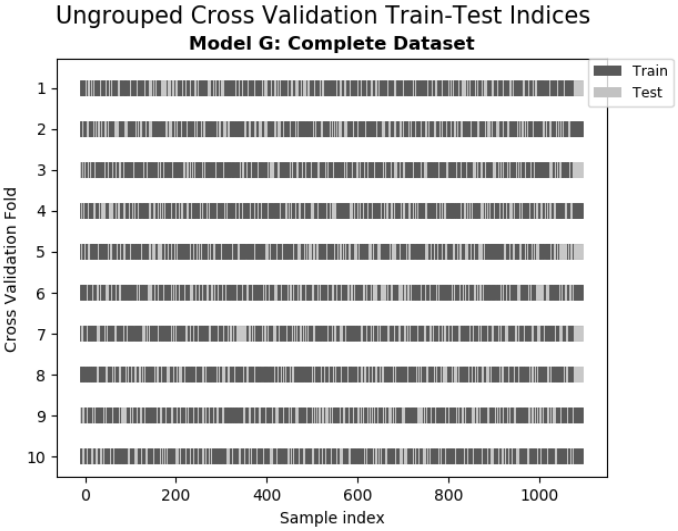

Figure 4.3f

Figures 4.3a-f: Cross-validation train-test indices. Illustration of indices used in the training and testing of the various neural network models developed. Included to illustrate the variability of train-test sets across the 10 cross-validation folds for the various models developed.

This behaviour is due to binary mixtures such as methane-propane being capable of exhibiting sII hydrate formation, thus providing a basis for predictions for multicomponent gases under conditions where few, if any multicomponent data points were present during training. Due to this effect, the inclusion of pure and binary methane gas mixtures introduces to the model in predicting conditions where training data is lacking, it is very likely that not all of the increased model performance is due to simply the presence of easier to predict equilibrium points when comparing the multicomponent exclusive grouped model A with complete dataset model B, and ungrouped multicomponent model F with complete dataset model G. As such, it can be seen that the inclusion of pure and binary component data into the training dataset does not hamper the model or provide grounds for dismissing model results. Furthermore, the results of model J being similar to B while using the same parameters and topology as G facilitates comparison of model B with G.

A significant result of these complete dataset models are the seed tests illustrated in Figure 4.2d, where it can be seen that the weakest fold for model G is R^2 >0.97, a major improvement over even model F. As discussed, this observation is due in part to the increased size of the dataset with more data which proves simpler to model than multicomponent data, however this also indicates the possibility of a damping effect, whereby the presence of pure and binary methane gases serves to reduce the effect bad data points may have on the model. The damping effect further may serve to improve predictive capabilities of the model in addition to this negative factor of increasing the number easily predicted samples in the test set. This will be further elaborated after discussing the variance of results.

A notable observation when considering models B & G is that the hold-out validation score is approximately the same, with a relatively minor deviation of 0.0069 in terms of the coefficient of determination. This indicates that sufficient independent data groups are present in the dataset to facilitate grouped cross-validation practices, as the tuned model parameters both lead to a similar result when tested using the same hold-out validation set. These hold-out indices are illustrated in Figure 4.4b, which depicts the hold-out split indices being constant between models B & G. Model B has been developed using the complete dataset, a stratified 10% hold-out validation set, and a randomly selected 30% group cross-validation procedure across 10 Folds. Unlike model A, model B exhibits a relatively low standard deviation of 0.0181, thus indicating that most cross-validation folds present similar R^2 scores. Figure 4.1b illustrates the relatively low variance of model B, achieving a minimum fold R^2 score of approximately 0.92. Although the variance of the cross-validated model appears low, seed tests are performed in

Figure 4.2a to ensure that results are not simply valid for a favourable train-test split random seeding. Figure 4.2a indicates that, box-plots do correlate to a certain degree with the Interquartile range for model B in Figure 4.1b. A variance in results when altering the random seed used to determine which groups are used to develop the training and testing sets for cross validation is expected due to the highly non-Gaussian nature of the dataset and the limited number of equilibrium data samples, which further aggravates variance between results of a different random seeding. Thus, as the model has undergone a grid-search optimization procedure concentrated on a certain random seeding, it is expected that other random seeds will display varied results, but similar interquartile ranges over 10-fold cross-validation. While the interquartile ranges indicated in Figure 4.2a largely correlate for most folds with the variance box plot in Figure 4.1b, the seed test indicates that for certain individual folds, unexpectedly poor R^2 scores are yielded, with values as low as 0.805 resulting for single folds. Unlike model A, where very poor predictions are noted across the interquartile range, these weak folds are the result of singular poor folds. Due to the adequate interquartile ranges, it is likely that these weak folds present in Figure 4.2a are from the model attempting to predict a certain range multicomponent equilibrium conditions, which is represented only by a single source where pure and binary component data proves insufficient to adequately model the complex multicomponent nature of the gas in question. This significant outlier only appears in a few random seedings, thus indicating that the issue is likely due to a single, completely isolated group or two, which are present in the test set. As all training data is used to train the model tested on the hold-out set, this weak fold is covered in the final model by the singular problematic group or two guaranteed being present in the training set. Another possibility is that there are several data points present in the dataset, likely from the same groups, which do not truly report equilibrium conditions, perhaps due to the metastability effect not being adequately accounted for, or due to measurement error. As experimental sources of data have been checked where possible to ensure obvious errors have been excluded, it is however unlikely that these poor-quality points would have such a significant effect on individual folds. Additionally, as is demonstrated by the consistent performance of model G across seed tests as illustrated in Figure 4.2d, no exceedingly weak folds are present. As such, weak folds present in Figure 4.2a for model B are most likely due to isolated ranges of data.

As with multicomponent exclusive model C being developed in the absence of a hold-out validation set, models D and E have been developed for the complete dataset to assess potential improvements to model B by providing a marginally larger cross-validation dataset. Just as no

significant improvement to model A is obtained through model C, models D & E fail to provide a significant improvement to the results of model B. This is expected, as a cross-validation performed without a hold-out validation set required to attempt optimization is unlikely to compete with a model which has undergone an extensive grid-search to yield parameters near the optimal configuration. Useful information can however be garnered from models D & E due to the varied activation functions through which these models are trained. Many similar studies have made extensive use of the sigmoid hidden layer activation function due to its ability to introduce non-linearity to the system, and illustrating how the Rectified Linear Unit (ReLU) activation function is able to successfully introduce non-linearity to the system while yielding similar results to a model trained using the sigmoid activation function. Model D has been developed using the sigmoid activation function for hidden layer activation, while output layer activation is achieved via the softplus activation function. Model E has been developed using the ReLU activation function for hidden layer activation, while a linear activation function determines the pressure value provided by the output layer. The results for both models and their respective topologies are provided in Table 3.4. As can be seen, results from both models are largely similar, with the cross-validation average and standard deviation displaying slight differences, which can be explained by variance due to slight differences in the results of replicated models, and observations of the bias-variance trade-off. This is further compounded as Models D & E have been developed with a different number of hidden layers. As expected from observations in Glorot & Bengio (2010) and Maas et al. (2013), training times were longer for using the sigmoid activation function for hidden layers than the ReLU function, while ReLU models achieved slight improvements in regression performance. The factor of training time proves significant considering the number of models developed during hyperparameter optimization, and is the reason all other models in this investigation have been developed using ReLU activation for hidden layers. Models D & E indicate that for the datasets used in this investigation, the ReLU activation function performs just as well as the sigmoid activation function for hidden layer activation. This is a significant observation, as all other neural networks in this research have been developed using a ReLU activation function for hidden layer activation, and a linear activation function for output layer activation. The reason that ReLU activation was applied so prolifically in the models developed in this research, is that a slightly faster training time was observed over sigmoid activation, and allowed the time-span over which parameter grid-searches were conducted to be shortened.

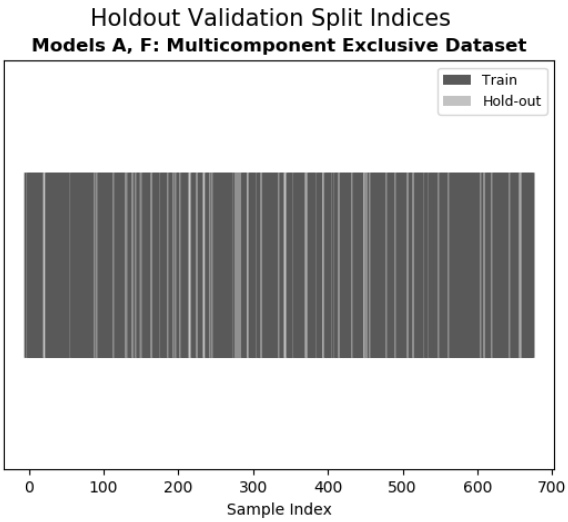

Figure 4.4a: Multicomponent Exclusive Hold-out validation set indices. Stratified selection of indices

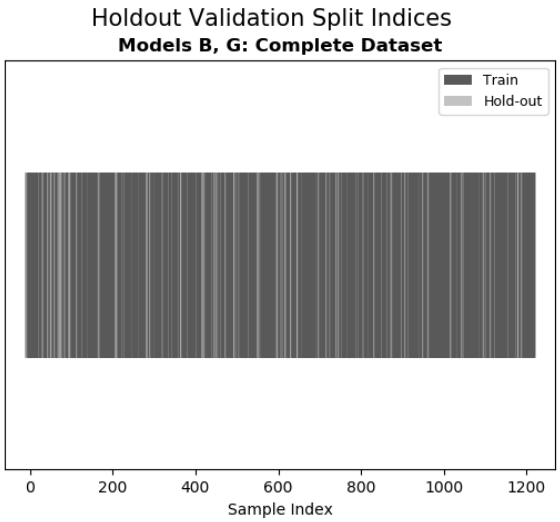

Figure 4.4b: Complete Dataset Hold-out Validation set indices. Stratified selection of indices

The results obtained by models B, D, E & G are obtained through cross-validation and, where applicable, hold-out test sets, which includes pure methane, binary methane mixtures and multicomponent mixtures up to and including natural gas mixtures. Ideally, these models would be tested using exclusively multicomponent data, even though training included pure and binary components. This however is not possible due to the limited availability of data. It is essential to test model performance on pure methane and binary methane mixtures even though these are not relevant to natural gas mixtures, as ensuring the model has not been overfitted is of paramount importance. A more prudent final validation practice, for those seeking to apply a model such as the one developed in this investigation, would be to test the performance of these models on completely independent, unseen equilibrium natural gas mixtures, which have not been included in this investigation. This could be performed when practical application is considered, where the model is tested on gas mixtures and conditions relevant to those likely encountered during operation. The act of not validating the potential variance of the model, even for pure or binary mixtures, which were used to train the model, would reduce confidence in the final reported accuracy, as there would be the potential for unforeseen weak ranges of multicomponent data which lie outside the range of conditions which have been tested by the multicomponent validation set.

Overall, the positive results of model F reveal that a neural network based on multicomponent hydrate equilibrium experimental data is indeed viable. While the prediction of multicomponent gas hydrate equilibrium pressure could possibly be slightly less accurate than the reported results of model G, the actual results when tested with real fluids are most likely significantly more accurate than the results of models A and F. Thus, the addition of pure and

binary components into the training dataset likely provides a significant increase in potential model accuracy in systems where multicomponent data is lacking over a number of ranges of conditions. These additional components further serve to dampen the effect of experimental error on model accuracy. As the results of model B are not in themselves unacceptable, and within the margins of error, and model G achieves a highly accurate regression model, it can be seen that complete dataset neural network models are less likely than the multicomponent exclusive dataset neural network models to be heavily affected by potential experimental or measurement errors which are present for various equilibrium samples in the dataset. Thus, the presence of pure methane and binary methane mixture data will serve to dampen the error resulting from inaccurate data. In conclusion, as the complete dataset contains all elements of the multicomponent dataset, it is thus possible to conclude that the complete dataset neural network models, particularly model G, are indeed capable of accurately predicting multicomponent data. This claim is further reinforced through the lack of weak folds for model G indicated in Figure 4.1b and Figure 4.2d. As models F and G have been developed using the same neural network topology, and as the complete dataset contains all data from the multicomponent dataset, it can be concluded that the complete dataset neural network accurately models multicomponent data on account of the positive results of both multicomponent exclusive model F and complete dataset model G. The final model developed on the complete dataset is less likely to yield poor predictions in ranges where limited multicomponent data is available than for multicomponent exclusive dataset models.

Finally, it is worth emphasizing that cross-validation is merely a tool, which allows further insight as to the model's ability to predict unseen data, and to check whether or not overfitting has been performed. Hold-out validation is a further metric which indicates the model's ability to predict truly unseen data which has been excluded from cross-validation, and provides a basis for optimizing the model parameters. In order to utilize the neural networks developed in this research, the entire dataset could be used to train the model using the network topology and parameters for the desired configuration, listed in Table 3.4. The entire dataset would be used to train a neural network which would be applied to any gas-hydrate system within the constraints listed in Tables 3.2 & 3.3, depending on whether the multicomponent exclusive or complete dataset is used. In doing so, the models listed in Table 3.4 can easily be converted into practically applicable models. Thus, in effect, models B & G complement each other, although network topologies vary, the effect of altering the network topology is merely performed to attempt optimization. Model J confirms this, having been developed as a means

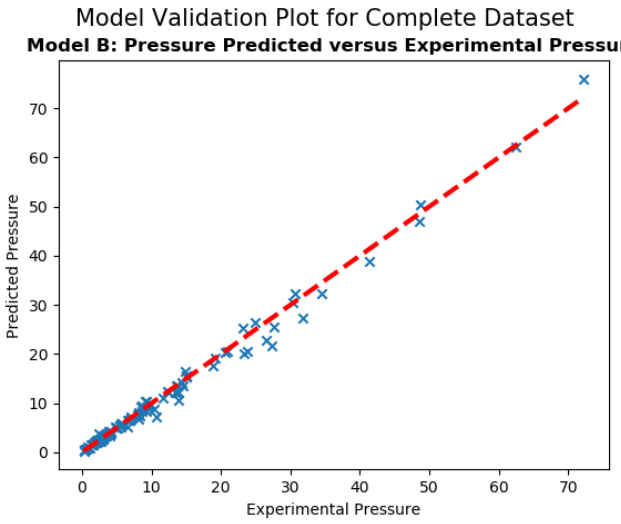

Figure 4.5a

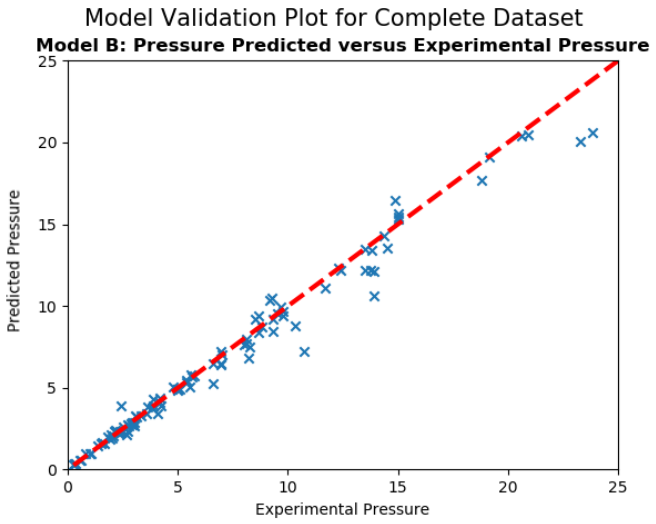

Figure 4.5b

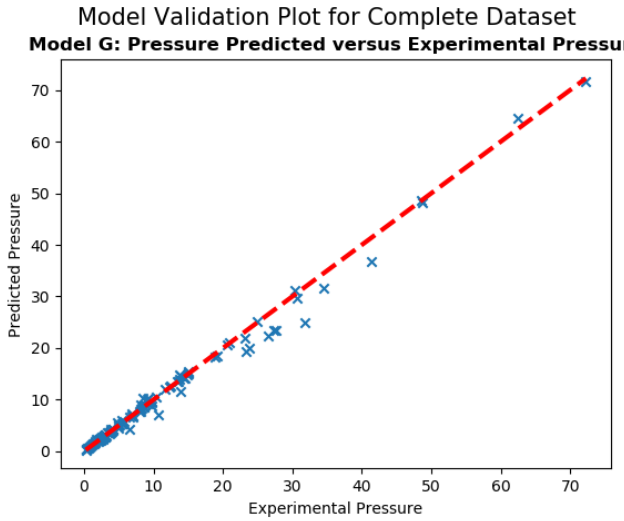

Figure 4.5c

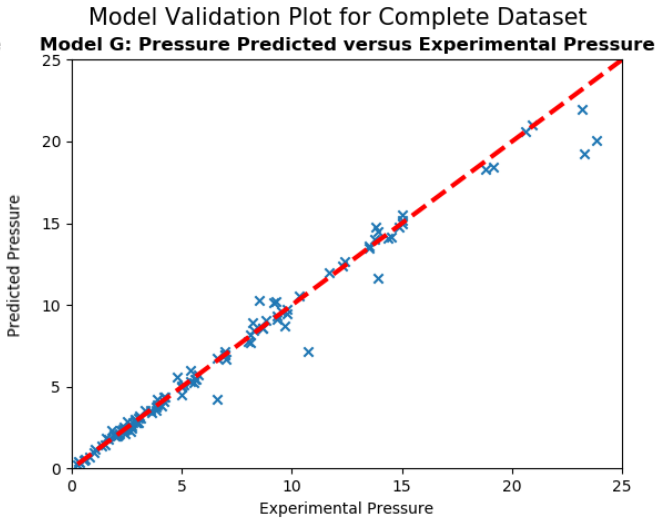

Figure 4.5d

Figures 4.5a-d: Regression plots of predicted vs experimental pressure: Developed using hold-out testing data for model B and model G. Overall deviations are within expected performance.

of mirroring the parameters of model G while performing grouped-cross validation, and still achieving similar results to model B. So long as deep learning is performed, with at least two hidden layers being used, and with the number of neurons per hidden layer being at least the number of inputs, which in this case is nine, results for any model trained under these criteria would not be expected to differ dramatically from the desired output listed in Table 3.4. A recommendation as to the exact parameters used to train a practically applicable model is provided as Model G. Note that alterations to the training dataset would likely alter certain model parameters leading to near-optimal solutions significantly, notably the number of epochs over which the model is trained – the number of times the entire training set undergoes back-propagation. The network topology and parameters such as ReLU hidden layer and Linear

Output Layer activation are unlikely to significantly change as long as most of the dataset remains constant. As such, providing the exact parameters such as number of epochs and training batch-size is an unfruitful exercise, as these can be readily obtained by a grid-search procedure, or simply performing a few wide iterations to gain insight into the potential accuracy of the model. It is important to note in closing however, that training a model with the desired parameters obtained from model G over the entire dataset, including hold-out validation data could introduce the possibility of the model overfitting which may go undetected due to a lack of separate validating data. Thus, in the absence of further testing data, it is recommended that the final model exclude hold-out data from training.

These results, particularly those of models performing cross-validation grouping (models A and B), indicate that additional experimental data from independent sources incorporated into existing datasets would result in further improved models. Additional data tailored toward application, such as samples from a specific gas pipeline seeking to apply this model, could be used to provide a testing basis for training model G on the entirety of the complete dataset, and provide a final indication of model performance for a specific application. In the event of significantly more data being published, this research could be revised and would likely lead to improved cross-validation results for grouped models A and B. As discussed, the inclusion of hydrogen sulphide as a feature of the model was not possible due to the lack of independent experimental studies. Several attempts were made to incorporate hydrogen sulphide into the datasets and to establish a convergent model, however performing grouped cross-validation resulted in folds which were simply too inaccurate to extract any meaningful information. It is possible to develop a model in the absence of performing grouping, however due to the highly limited number of independent sources covering a few narrow ranges of data including these models in this investigation would not prove meaningful when compared to the wide range of conditions covered in the multicomponent exclusive and complete datasets. More publications of equilibrium data considering natural gases containing hydrogen sulphide over a wide range of conditions would facilitate the inclusion of this component as a feature of the model. Including hydrogen sulphide as a feature of this model would serve as a significant step towards the development of a truly universal model. Such data could easily be incorporated into the datasets provided in this investigation, and yield results rapidly due to the established topology of the neural network through means of a grid-search procedure.

Overall for the purposes of this investigation, several aspects of the model have been assessed: The ability of the model to predict equilibrium data over a wide range of conditions, the ability

of the model to predict multicomponent data, and the dependency of the model on individual sources of data used in training and thus the potential for experimental or measurement errors to influence the final neural network. This investigation has proven through means of an extensive 10-fold cross-validation procedure, and stratified hold-out validation approach, whereby validation data is proportionally sampled from each independent experimental source of data, that the model does indeed accurately predict equilibrium pressure over a wide range of conditions. Seed-tests for various models further prove that model results are not artificially tailored through means of a favourable random seeding during cross-validation. The ability of the model to predict multicomponent data has been assessed through developing neural networks trained and tested using exclusively multicomponent data, all of, which is present in the complete dataset used to train and test the final model G.

Dependency of the model on individual sources of experimental data has been investigated, and variance plots of grouped cross-validation do indicate certain ranges of data where little to no independent source overlap is present. Analysis of the model incorporating pure and binary methane gases does however reveal a damping effect provided by the inclusion of this data, thus reducing the potential impact experimental or measurement errors of equilibrium data would have on the final model. This experimental error could take the form of incorrect methodology where inadequate time was provided between dissociation steps thus failing to account for metastability and thus reporting a value within the hydrate stability zone as opposed to equilibrium and outdated means of judging whether equilibrium has been reached. Thus, it can be concluded that the research objectives have been achieved, and a highly accurate model has been developed which is capable of predicting the equilibrium pressure for gas hydrates of a specific gas-phase concentration at a specified temperature, over a wide range of conditions. It must be emphasized that the model is designed to predict sweet natural gases, which are free of hydrogen sulphide.

CHAPTER 5: CONCLUSION AND FUTURE RESEARCH

5.1 Conclusion

In this investigation, deep learning has been applied to the prediction of gas hydrate equilibrium conditions over a wide range of data through means of a Multi-Layer Perceptron regression. Neural networks have been developed using two to three hidden layers with a high hidden-layer neuron count. Several artificial neural networks have been developed in this research for the purpose of proving the ability of the neural network model to accurately model multicomponent data, and to assess the dependence of the model on individual experimental sources of data to cover certain ranges of conditions. A combination of 10-fold cross-validation and hold-out validation has been performed to assess the variance of the model, while facilitating parameter optimization. Validation practices ensure that the results of the model reflect a wide range of conditions. Through developing models according to two separate datasets, one compiled of exclusively multicomponent data, and another including pure and binary methane gases, it has been proven that multicomponent data for both datasets has been adequately predicted by the neural networks. An emphasis has been placed on the collection of experimentally obtained natural gas equilibrium data. Additional models do reveal a dependence on individual experimental sources to provide equilibrium data for various ranges of conditions where little overlap between other independent sources exists. While cross-validation performed on grouped datasets where data cannot be used to both train and test in the same fold results in significantly worse predictions than ungrouped validation, results are generally acceptable in terms of average R^2 score and variance, barring the performance of poor folds where there is an evident lack of overlap in data from independent sources. The presence of pure and binary methane in the training set has been shown to provide a damping effect on potential errors of experimental data. Seed-tests, where the effect of the random selection of data for training and testing purposes is investigated, have been performed and assurance is provided that models are not merely accurate for a favourable randomization. A lack of independent experimental sources of data prevented the inclusion of hydrogen sulphide as a model feature, and thus the model is restricted in application to sweet natural gas flow-lines. Overall, results indicate a highly accurate model has been developed, with the neural network accurately modelling a highly non-linear, multimodal phenomenon. Cross-validation of the final model yields a coefficient of determination R^2 score of 0.98604, with a standard deviation of 0.00352, while an R^2 score of 0.99255 is obtained for a 10% stratified hold-out

validation set after cross-validation. These results prove that the neural network methodology is well suited to predicting gas hydrate equilibria. Further testing of model performance compared with thermodynamic models, which have gained industrial acceptance, would serve to prove the ability of the neural network model to compete at an industrial level with established methodologies, especially when considering real natural gas equilibrium

5.2 Future Research

This research has largely demonstrated the viability of the neural network methodology across a wide dataset from numerous independent sources by illustrating a low cross-validation 10-fold variance. In order to further assess model accuracy, comparison with existing methods must be performed. A significant indicator of model performance aside from the validation practices employed in this investigation would be a comparison with industrially accepted and applied models. There is a wide range of software programs available, based on statistical thermodynamics, which could be used as a comparative basis outside of corporate developed models. Furthermore, comparison may also be performed via older graphical methods such as gas-gravity charts, and empirical equations which have experienced some degree of industrial application. Performing pressure predictions for the entire dataset included in this research would allow coefficients of determination (R^2 Score) to be compared, and neural network effectiveness could thus be further assessed. Furthermore, closer investigation into the exact predictions made by the neural network would allow for individual ranges where equilibrium pressure was poorly predicted to be analysed, and corrections may be made by performing further outlier elimination from the dataset where measurement error is suspected.

As has been emphasized in this report, the developed models do not account for the presence of hydrogen sulphide. In order to develop a comprehensive, universal model, the concentration of hydrogen sulphide in the produced gas should be included as a neural network model feature. As discussed, insufficient independent data sources were present in the dataset to facilitate the development of an H_2S inclusive model while considering grouping effects during cross validation. Gathering significantly more equilibrium data including this feature and combining the new data with current datasets may allow a sufficiently accurate model to be developed. As thermodynamic inhibitors act by altering equilibrium conditions, it is also possible to include the concentration of thermodynamic inhibitors such as methanol or ethylene glycol in the flow-line as a feature in the model. A significant amount of data is publicly available in various publication, and other models such as the neural network developed by (Chapoy et al., 2007) includes thermodynamic inhibitors as a feature.

REFERENCES

Ahmed, T. & McKinney, P. (2011). *Advanced Reservoir Engineering*, Gulf Professional Publishing, United States of America. pp.271-272.

Antunes, C.M.M, Kakitani, C., Neto, M.A.M, Morales, R.E.M. & Sum, A.K. (2018). 'An Examination of the Prediction of Hydrate Formation Conditions in the Presence of Thermodynamic Inhibitors'. *Brazillian Journal of Chemical Engineering*, 35(1), pp. 265-274.

Austvik, T., Li, X. & Gjertsen, L.H. (2006). 'Hydrate Plug Properties: Formation and Removal of Plugs'. *Annals of the New York Academy of Sciences*, 912(1), pp.294-303.

Ballard, A.L. (2002). 'A Non-Ideal Hydrate Solid Solution Model for a Multi-phase Equilibria Program'. PhD thesis, Colorado School of Mines, Golden, Colorado.

Ballard, A.L. & Sloan Jr, E.D. (2004). 'The Next Generation of Hydrate Prediction IV a Comparison of Available Hydrate Prediction Programs'. *Fluid Phase Equilibria*, 216(2), pp. 257-270.

Bassani, C.L., Barbuto, F.A.A., Sum, A.K. & Morales, R.E.M. (2017). 'Hydrate Formation Effects on Slug Flow Hydrodynamics and Heat Transfer: Wall Deposition vs Dispersion Formation'. *Journeys in Multiphase Flows IV (JEM 2017): March 27-31, Sao Paulo, Brazil, 2017*.

Bishnoi, P.R. & Dholabhai, P.D. (1998). 'Equillibrium Conditions for Hydrate Formation for a Ternary Gas Mixture of Methane, Ethane, Propane and a Natural Gas Mixture in the Presence of Electrolytes and Methanol'. *Fluid Phase Equilibria*, 158-160, pp. 821-827

Carroll, J. (2009). *Natural Gas Hydrates: A Guide for Engineers*. 2nd ed. Gulf Professional Publishing, United States of America. pp. 17-42

Cawley, G.C. & Talbot, N.L.C. (2010). 'On Over-fitting in Model Selection and Subsequent Selection Bias in Performance Evaluation'. *Journal of Machine Learning Research*, 11 (July), pp. 2079-2107

Chapoy, A., Mohammadi, A.H. & Richon, D. (2007). 'Predicting the Hydrate Stability Zones of Natural Gas Using Artificial Neural Networks'. *Oil & Gas Science and Technology – Rev. IFP*, 62(5), pp. 701-706.

Chong, Z.R., Yang, S.H.B., Babu, P., Linga, P. & Li, X. (2016). 'Review of Natural Gas Hydrates as an Energy Resource: Prospects and Challenges'. *Applied Energy*. 162, pp. 1633-1652.

Elgibaly, A.A. & Elkamel, A.M. (1998). ‘A New Correlation for Predicting Hydrate Formation Conditions for Various Gas Mixtures and Inhibitors’. *Fluid Phase Equilibria*, 152(1), pp. 23-42.

Elgibaly, A.A. & Elkamel, A.M. (1999). ‘Optimal Hydrate Inhibition Policies with the Aid of Neural Networks’. *Energy & Fuels*, 13(1), pp. 105-113.

Frostman, L.M. (2000). ‘Anti-Agglomerant Hydrate Inhibitors for Prevention of Hydrate Plugs in Deepwater Systems’. *Proceedings of the Society of Petroleum Engineers Annual Technical Conference and Exhibition: 1-4 October, Dallas, Texas, 2000*.

Gas Processors Suppliers Association [GPSA]. (2004). *Engineering Data Book*. 12th ed. Tulsa, Oklahoma.

Ghavipour, M., Ghavipour, M., Chitsazan, M., Najibi, S.H. & Ghidary, S.S. (2013). 'Experimental Study of Natural Gas Hydrates and a Novel Use of Neural Network to Predict Hydrate Formation Conditions'. *Chemical Engineering Research and Design*, 91(2), pp. 264-273

Ghiasi, M.M., Yarveicy, H., Arabloo, M., Mohammadi, A.H. & Behbahani, R.M. (2016). ‘Modeling of Stability Conditions of Natural Gas Clathrate Hydrates Using Least Squares Support Vector Machine Approach’. *Journal of Molecular Liquids*, 223, pp. 1081-1092.

Giavarini, C. & Hester, K. (2011). *Gas Hydrates: Immense Energy Potential and Environmental Challenges*. Springer-Verac, London, pp. 52-54

Glorot, X. & Bengio, Y. (2010). ‘Understanding the Difficulty of Training Deep Feedforward Neural Networks’. *Proceedings of the 13th International Conference on Artificial Intelligence and Statistics (AISTATS) 2010, Chia Laguna Resort, Sardinia, Italy*, 9, pp. 249-256.

Hammerschmidt, E.G. (1934). ‘Formation of Gas Hydrates in Natural Gas Transmission Lines’. *Industrial and Engineering Chemistry*, 26(8), pp. 851-855

Hawtin, R.W., Quigley, D. & Rodger, P.M. (2008). ‘Gas Hydrate Nucleation and Cage Formation at a Water/Methane Interface’. *Physical Chemistry Chemical Physics*, 10(32), pp.4853-4864

Heydari, A., Shayesteh, L. & Kamalzadeh, L. (2006). ‘Prediction of Hydrate Formation Temperature for Natural Gas Using Artificial Neural Network’. *Oil and Gas Business*, 2 (2016).

Hesami, S.M., Dehghani, M., Kamali, Z. & Bakyani, A.E. (2017). 'Developing a Simple-to-use Predictive Model for Prediction of Hydrate Formation Temperature'. *International Journal of Ambient Energy*, 38(4), pp. 380-388.

Hsueh, B., Li, W. & Wu, I. (2018). 'Stochastic Gradient Descent with Hyperbolic-Tangent Decay on Classification'. To be published in *2019 Winter Conference on Applications of Computer Vision (WACV)* [Preprint]. Available at: arXiv:1806.01593 (Accessed 3 May 2019).

Ikoku, C. U. (1980). *Natural Gas Engineering: A Systems Approach*. Tulsa, Oklahoma: PennWell Books.

Katz, D. (1945). 'Prediction of Conditions for Hydrate Formation in Natural Gases'. *Transactions of the AIME,* 160(1), pp. 140-149.

Kingma, D. & Ba, J. (2014). 'Adam: A Method for Stochastic Optimization'. *Proceedings of the 3rd International Conference on Learning Representations (ICLR), San Diego, 2015*. Available at: arXiv:1412.6980 (Accessed 3 May 2019).

Leshno, M., Lin, V.Y., Pinkus, A. & Schocken, S. (1993). 'Multilayer Feedforward Networks with a Nonpolynomial Activation Function can Approximate any function'. *Neural Networks*, 6(6), pp. 861-867.

Maas, A. L., Hannun, A.Y. & Ng, A.Y. (2013). 'Rectifier Nonlinearities Improve Neural Network Acoustic Models'. *Proceedings of the International Conference on Machine Learning (proc. IMCL 2013)*. Available at: robotics.stanford.edu/~amaas/papers/relu_hybrid_icml2013_final.pdf (Accessed 3 May 2019).

Makogon, Y.F. (1997). *Hydrates of Hydrocarbons*. Tulsa, Oklahoma: PennWell Books

Makogon, Y.F., Holditch, S.A., Makogon, T.Y. (2007). 'Natural Gas-Hydrates – A Potential Energy Sources for the 21st Century'. *Journal of Petroleum Science & Engineering*, 56(1), pp. 14-31.

Rajnauth, J., Barrufet, M. & Falcone, G. (2012). 'Hydrate Formation: Considering the Effects of Pressure, Temperature, Composition of Water'. *Energy Science and Technology*, 4(1), pp. 60-67.

Nwankpa, C., Ijomah, W., Gachagan, A. & Marshall, S. (2018). 'Activation Functions: Comparison of Trends in Practice and Research for Deep Learning'. Available at: arXiv:1811.03378 [cs.LG] (Accessed 3 May 2019).

Perrin, A., Musa, M.M. & Steed, J.W. (2013). 'The Chemistry of Low Dosage Clathrate Hydrate Inhibitors'. *Chemical Society Reviews*, 42(5), pp.1996-2015

Rodger, P.M. (1990). 'Stability of Gas Hydrates'. *Journal of Physical Chemistry*, 94(15), pp. 6080-6089

Ruffine, L., Broseta, D. & Desmedt, A. (2018). *Gas Hydrates 2: Geoscience Issues and Potential Industrial Applications*. 2nd ed. ISTE- Wiley, p.124

Shahnazar, S. & Hasan, N. (2014). 'Gas Hydrate Formation Condition: Review on Experimental and Modeling Approaches'. *Fluid Phase Equilibria*, 379, pp. 72-85.

Sloan Jr, E.D. (1998). 'Gas Hydrates: Review of Physical/Chemical Properties', *Energy & Fuels*, 12(2), pp.191-196

Sloan Jr, E.D. & Koh, C.A. (2007). *Clathrate Hydrates of Natural Gases*. 12th ed. CRC Press.

Soroush, E., Mesbah, M., Shokrollahi, A., Rozyn, J., Lee, M., Kashiwao, T. & Bahadori, A. (2015). 'Evolving a Robust Modeling Tool for Prediction of Natural Gas Hydrate Formation Conditions'. *Journal of Unconventional Oil and Gas Resources*, 12, pp. 45-55.

South Africa Government. (2010). *Integrated Resource Plan for Electricity [IRP]*, Department of Energy, viewed 11 May 2019, Available at: www.energy.gov.za/files/irp_overview.html (Accessed 11 May 2019).

South Africa Government. *Strategic Energy Plan 2015-2020*, Department of Energy, viewed 11 May 2019, Available at: www.energy.gov.za/files/aboutus/DoE-Strategic-Plan-2015-2020.pdf

Stringari, P., Valtz, A. & Chapoy, A. (2014). 'Study of Factors Influencing Equilibrium and Uncertainty in Isochoric Hydrate Dissociation Measurements'. *Proceedings of the 8th International Conference on Gas Hydrates (ICGH8-2014): 28 July – 1 August, Beijing, China, 2014*.

Tohidi, B., Burgass, R.W., Danesh, A., Østergaard, K.K & Todd, A.C. (2000). 'Improving the Accuracy of Gas Hydrate Dissociation Point Measurements'. *Annals of the New York Academy of Sciences*, 912(1), pp. 924-931.

Tohidi, B., Østergaard, K.K., Danesh, A., Todd, A.C. & Burgass, R.W. (2001). 'Structure-H Gas Hydrates in Petroleum Reservoir Fluids'. *The Canadian Journal of Chemical Engineering*, 79(3), pp. 384-391

Turner, D., Dubois, J., Bass, R., Hamilton, T., Howlett, J. & Greaves, D. (2014). 'Electric Heating for Hydrate Prevention in an Arctic, Single-Line Tieback'. *Journal of Chemical & Engineering Data*, 60(2), pp. 356-361

Union of Concerned Scientists. (2019*). Environmental Impacts of Natural Gas*. [Online], Available at: www.ucsusa.org/clean-energy/coal-and-other-fossil-fuels/environmental-impacts-of-natural-gas (Accessed 11 May 2019).

U.S. Energy Information Administration. (2013). 'Natural gas turbines are generally used to meet peak energy load'. *Today in Energy*, 1 October 2013, [Online], Available at: www.eia.gov/todayinenergy/detail.php?id=13191 (Accessed 11 May 2019).

Van der Waals, J.H. & Platteeuw, J.C. (1959). 'Clathrate Solutions'. *Advances in Chemical Physics*, 2, pp. 1-57.

Veluswamy, H.P., Kumar, A., Seo, Y., Lee, J.D. & Linga, P. (2018). 'A Review of Solidified Natural Gas (SNG) Technology for Gas Storage via Clathrate Hydrates. *Applied Energy*, 216, pp. 262-285

Ward, Z. (2015). 'Phase Equilibria of Gas Hydrates Containing Hydrogen Sulphide and Carbon Dioxide'. PhD thesis, Colorado School of Mines, Golden, Colorado.

Wu, Q. & Zhang, B. (2010). 'Memory Effect on the Pressure-Temperature Condition and Induction Time of Gas Hydrate Nucleation'. *Journal of Natural Gas Chemistry*, 19(4), pp. 446-451

Zenali, N., Saber, M. & Ameri, A. (2012). 'Comparative Analysis of Hydrate Formation Pressure Applying Cubic Equations of State (EoS), Artificial Neural Network (ANN) and Adaptive Neuro-Fuzzy Interface System (ANFIS)'. *International Journal of Thermodynamics (IJoT)*, 15(2), pp. 91-101.

Zerpa, L.E. (2013). 'A Practical Model to Predict Gas Hydrate Formation, Dissociation and Transportability in Oil and Gas Flowlines'. PhD thesis, Colorado School of Mines, Golden, Colorado

APPENDICES

Appendix A: Data Sources

Note that a significant number of sources were discovered from the works of Sloan & Koh (2007), which records a large number of equilibrium samples for a wide range of conditions from various experimental studies.

For a limited number of cases, where data is provided in a graphical format, equilibrium conditions have been obtained by manually measuring these points. This process involved determining the mid-point of indicators along a curve, and has been applied to cases where equilibrium condition plots have been made where gas composition has been explicitly specified. The associated error due to these manual measurements is expected to be low due to the precision with which manual measurements were performed. The dampening factor associated with the complete dataset further reduces the impact minor inaccuracies due to manual measurements may have had on model results.

Adisasmito, S., Frank, R.J. & Sloan, E.D. (1991). 'Hydrates of Carbon Dioxide and Methane Mixtures'. *Journal of Chemical Engineering Data*, 36 (1), pp. 68-71

Adisasmito, S. and Sloan, E.D. (1992). 'Hydrate of Hydrocarbon Gases Containing Carbon Dioxide'. *Journal of Chemical Engineering Data*, 37(3), pp. 343-349

AlHarooni, K., Gubner, R., Iglauer, S., Pack,D. & Barifcanim A. (2017). 'Influence of Regenerated Monoethylene Glycol on Natural Gas Hydrate Formation'. *Energy & Fuels*, 31(11), pp. 12914-12931

Babakhani, S.M., Bouillot, B., Douzet, J. Ho-Van, S. & Herri, J. (2018). 'PVTx Measurements of Mixed Gas Hydrates in Batch Conditions Under Different Crystallization Rates: Influence on Equilibrium'. *Journal of Chemical Thermodynamics*, 122(C), pp. 73-84

Bishnoi, P.R. & Dholabhai, P.D. (1998). 'Equillibrium Conditions for Hydrate Formation for a Ternary Gas Mixture of Methane, Ethane, Propane and a Natural Gas Mixture in the Presence of Electrolytes and Methanol'. *Fluid Phase Equilibria*, 158-160, pp. 821-827

Carson, D.B. & Katz, D.L. (1941). 'Natural Gas Hydrates'. *Transactions of the American Institute of Mining and Metallugrical Engineers*, Dallas Meeting, October 1941, 146, pp. 150-158

Chen, L., Sun, C., Chen, G., Nie, Y., Sun, Z. & Yantao, L. (2009). 'Measurements of Hydrate Equilibrium Conditions for CH4, CO2, and CH4 + C2H6 + C3H8 in Various Systems by Step-heating Method'. *Chinese Journal of Chemical Engineering*, 17 (4), pp. 635-641

Chen, L., Sun, C., Nie, Y., Sun, Z., Yang, L. & Chen. G. (2009). 'Hydrate Equilibrium Conditions of (CH4 + C2H6 + C3H8) Gas Mixtures in Sodium Dodecyl Sulfate Aqueous Solutions'. *Journal of Chemical & Engineering Data*, 54 (5), pp. 1500-1503

de Roo, J.L., Peters, C.J., Lichtenhaler, R.N. & Diepen, G.A.M. (1983). 'Occurrence of Methane Hydrate in Saturated and Unsaturated Solutions of Sodium Chloride and Water in Dependence of Temperature and Pressure'. *AIChE Journal*, 29 (4), pp. 651-657

Deaton, W.M. and Frost, E.M. (1946). *Gas Hydrates and Their Relation to the Operation of Natural-Gas Pipe Lines*. U.S. Bureau of Mines.

Ghavipour, M., Ghavipour, M., Chitsazan, M., Najibi, S.H. & Ghidary, S.S. (2013). 'Experimental Study of Natural Gas Hydrates and a Novel Use of Neural Network to Predict Hydrate Formation Conditions'. *Chemical Engineering Research and Design*, 91(2), pp. 264-273

Jager, M.D. & Sloan, E.D. (2001). 'The Effect of Pressure on Methane Hydration in Pure Water and Sodium Chloride Solutions'. *Fluid Phase Equilibria*, 185(1), pp. 89-99

Jager, M.D. and Sloan, E.D. (2002). 'Structural Transition of Clathrate Hydrates Formed from a Natural Gas'. *Proceedings of the Fourth International Conference on Gas Hydrates, May 19-23*, pp. 575-580

Jhaveri, J. & Robinson, D.B. (1965). 'Hydrate in the Methane-Nitrogen System'. *The Canadian Journal of Chemical Engineering*, 43 (2), pp. 75-78

Karaaslan, U. & Parlaktuna, M. (2000). 'Surfactants as Hydrate Promoters?'. *Energy & Fuels*, 14(5), pp. 1103-1107

Kobayashi, R., Withrow, H.J., Williams, G.B. and Katz, D.L. (1951). 'Gas Hydrate Formation with Brine and Ethanol Solutions'. *Proceedings of the 30th Annual Convention, Natural Gas Association of America*, pp. 27-31

Le Quang, D., Le Quang, D., Bouillot, B. Herri, J., Glenat, P. & Duchet-Suchaux, P. (2016). 'Experimental Procedure and Results to Measure the Composition of Gas Hydrate, during

crystalization and at equilibrium, from N2-CO2-C2H6-C3H8-C4H10 Gas Mixtures'. *Fluid Phase Equilibria*, 413 (Apr), pp. 10-21.

Lee, J. & Kang, S. (2011). 'Phase Equilibria of Natural Gas Hydrates in the Presence of Methanol, Ethylene Glycol, and NaCl Aqueous Solutions'. *Industrial & Engineering Chemistry Research*, 50 (14), pp. 8750-8755

Lim, D., Ro, H., Seo, Y., Seo, Y., Lee, J.Y., Kim, S., Lee, J. & Lee, H. (2017). 'Thermodynamic Stability and Guest Distribution of CH4/N2/CO2 Mixed Hydrates for Methane Hydrate Production Using N2/CO2 Injection'. *Journal of Chemical Thermodynamics*, 106, pp. 16-21

Mahabadian, M.A., Chapoy, A., Burgass, R. & Tohidi, B. (2016). 'Development of a Multiphase Flash in the Presence of Hydrates: Experimental Measurements and Validation with the CPA equation of State'. *Fluid Phase Equilibria*, 414, pp. 117-132

McLeod, H.O. and Campbell, J.M. (1961). 'Natural gas Hydrates at Pressures to 10,000 psia'. *Journal of Petroleum Technology*, 13(06), pp. 590-594

Mei, D., Liao, J., Yang, J. & Guo, T. (1996). 'Experimental and Modeling Studies on the Hydrate Formation of a Methane + Nitrogen Gas Mixture in the Presence of Aqueous Electrolyte Solutions'. *Industrial & Engineering Chemistry Research*, 35 (11), pp. 4342 – 4347

Mei, D., Liao, J., Yang, J. and Guo, T. (1998). 'Hydrate Formation of a Synthetic Natural Gas Mixture in Aqueous Solutions Containing Electrolyte, Methanol, and (Electrolyte + Methanol)'. *Journal of Chemical Engineering & Data*, 43(2), pp. 178-182

Moehebbi, V. & Behbahani, R.M. (2014). 'Measurement of Mass Transfer Coefficients of Natural Gas Mixture during Gas Hydrate Formation'. *Iranian Journal of Oil & Gas Science and Technology*, 4 (1), pp.66-80

Mogbolu, P.O. & Madu, J. (2014). 'Prediction of Onset of Gas Hydrate Formation in Offshore Operations'. *Society of Petroleum Engineers Nigeria Annual International Conference and Exhibition, 5-7 August, Lagos, Nigeria*.

Munck, J. Skjold-Jørgensen, S. & Rasmussen, P. (1988). 'Computations of the Formation of Gas Hydrates'. *Chemical Engineering Science*, 43(10), pp. 2661-2672. Citing: Ng & Robinson (1984).

Nakamura, T., Makino, T., Sugahara, T & Ohgaki, K. (2003). 'Stability Boundaries of Gas Hydrates Helped by Methane – Structure-H Hydrates of Methylcyclohexane and cis-1,2-dimethylcyclohexane'. *Chemical Engineering Science*, 58(2), pp. 269-273

Ng, H. & Robinson, D.B. (1976). 'The Role of n-Butane in Hydrate Formation'. *AIChE Journal*, 22 (4), pp. 656-661

Nixdorf, J. & Oellrich, L.R. (1997). 'Experimental Determination of Hydrate Equillibrium Conditions for Pure Gases, Binary and Ternary Mixtures and Natural Gases'. *Fluid Phase Equilibria*, 139 (1997), pp. 325-333

Notz, P.K. & Burke, N.E. (1991). 'Measurement and Predictionof Hydrate Formation Conditions for Dry Gas, Gas Condensate, and Black Oil Reservoirs'. *Offshore Technology Conference, 6-9 May, Houston, Texas*, pp. 409-419

Qureshi, M.F., Atilhan, M., Altamash, T., Tariq, M., Khraisheh, M., Aparicio, S. & Tohidi, B. (2016). 'Gas Hydrate Prevention and Flow Assurance by Using Mixtures of Ionic Liquids and Synergent Compounds: Combined Kinetics and Thermodynamic Approach'. *Energy & Fuels*, 30(4), pp. 3541-3548

Saberi, A., Alamdari, A., Shariati, A. & Mohammadi, A.H. (2017). 'Experimental Measurement and Thermodynamic Modeling of Equilibrium Condition for Natural Gas Hydrate in MEG Aqueous Solution', *Fluid Phase Equilibria*. 459, p. 110-118

Sadeq, D., Iglauer, S., Lebedev, M. & Smith, C. (2017). 'Experimental Determination of Hydrate Phase Equilibrium for Different Gas Mixtures Containing Methane, Carbon Dioxide and Nitrogen with Motor Current Measurements'. *Journal of Natural Gas Science & Engineering*, 38, p. 59-73

Semenov, M.E., Kalacheva, L.P., Yu,E. & Rozhin, I.I. (2010). 'Natural Gas Hydrate Decomposition in the Presence of Methanol'. *Chemistry for Sustainable Development*, 18, 147-151

Seo, Y. & Lee, H. (2001). 'Multiple-Phase Hydrate Equilibria of the Ternary Carbon Dioxide, Methane and Water Mixtures'. *Journal of Physical Chemistry B*, 105 (41), pp. 10084-10090

Seo, Y., Kang, S. & Jang, W. (2009). 'Structure and Composition Analysis of Natural Gas Hydrates: 13C NMR Spectroscopic and Gas Uptake Measurements of Mixed Gas Hydrates' . *Journal of Physical Chemistry A*, 113(35), pp. 9641-9649

Seo, Y., Kang, S., Lee, H. Lee, C. & Sung, W. (2000). 'Hydrate Phase Equilibria for Gas Mixtrures Containing Carbon Dioxide: A Proof-of-Concept to Carbon Dioxide Recovery from Multicomponent Gas Stream'. *Korean Journal of Chemical Engineering*, 17 (6), pp. 659-667

Sloan Jr, E.D. & Koh, C.A. (2007). *Clathrate Hydrates of Natural Gases*. 12th ed. CRC Press. Citing: Verma, V.K. (1974). 'Gas Hydrates from Liquid Hydrocarbon-Water Systems'. Thesis: University of Michigan.

Smith, C., Pack, D. & Barifcani, A. (2017). 'Propane, n-butane and i-butane stabilization effects on methane gas hydrates'. *Journal of Chemical Thermodynamics*, 115, pp. 293-301

Smith,C., Barifcani,A. And Pack,D. (2016). 'Helium substitution of natural gas hydrocarbons in the analysis of their hydrate'. *Journal of Natural Gas Science and Engineering*, 35(September 2016), pp. 1293-1300

Subramanian, S., Kini, R.A., Dec, S.F. & Sloan, E.D. 'Structural Transition Studies in Methane + Ethane Hydrates Using Raman and NMR'. *Annals of the New York Academy of Sciences*, 912(1), pp. 873-886

Tariq, M., Atilhan, M., Kharisheh, M., Othman, E., Castier, M.,Garcia, G., Aparicio, S. & Tohidi, B. (2016). 'Experimental and DFT Approach on the Determination of Natural Gas Hydrate Equilibrium with the Use of Excess N2 and Choline Chloride Ionic Liquid as an Inhibitor'. *Energy & Fuels*, 30(4), pp. 2821-2832

Taylor, C. & Kwan, J.T. (2004). *Advances in the Study of Gas Hydrates*. New York: Kluwer Academic/Plenum Publishers. Citing: Shukla, Khokhar & Kalpakci (2004).

Thakore, J.L. & Holder, G.D. (1987). 'Solid-Vapor Azeotropes in Hydrate-Formaing Sytems'. *Industrial & Engineering Chemistry Research*, 26 (3), pp. 462-469

Tohidi, B., Anderson, R., Clennell, M.B. and Biderkab, A.B. (2001). 'Visual Observation of Gas-Hydrate Formation and Dissociation in Synthetic Porous Media by Means of Glass Micromodels'. *Geology*, 29(9), pp. 867-870

Tohidi, B., Danesh, A. and Burgass, R.W. (1994). Hydrates Formed in Unprocessed Wellstreams. *Proceedings of the 69th SPE Annual Technical Conference*, September 25-28, pp. 157

Tohidi, B., Danesh, A., Burgass, R.W. & Todd, A.C. (1996). 'Effect of Heavy Hydrate Formers on the Hydrate Free Zone of Real Reservoir Fluids'. *European Production Operations Conference and Exhibition, 16-17 April, Stavanger, Norway 1996*, pp. 257-261

Tohidi, B., Danesh, A., Todd, C. and Burgass, R.W. (1997). 'Hydrate-Free Zone for Synthetic and Real Reservoir Fluids in the Presence of Saline Water'. *Chemical Engineering Science*, 52(19), pp. 3257-3263

Ward, Z. (2015). 'Phase Equilibria of Gas Hydrates Containing Hydrogen Sulphide and Carbon Dioxide'. PhD thesis, Colorado School of Mines, Golden, Colorado.

Wilcox, W.I., Carson, D.B. and Katz, D.L. (1939). 'Natural Gas Hydrates'. *Industrial and Engineering Chemistry*, 33(5), pp. 662-665

Wu, B., Robinson, D.B. & Ng, H. (1976). 'Three and Four-Phase Hydrate Forming Conditions in Methane + Isobutane + Water Correction of an Error in a Previous Paper'. *Journal of Chemical Thermodynamics*, 9 (2), pp. 461-469

Xu, S., F, S., Wang, Y. & Lang, X. (2015). 'An Investigation of Kinetic Hydrate Inhibitors on the Natural Gas from the South China Sea', *Journal of Chemical Engineering & Data*, 60(2), pp. 311-318

Yang, M., Song, Y., Liu, Y. & Jiang, L. (2012). 'Effects of Gas Component on Hydrate Equilibrium in Porous Medium'. *Proceedings of the Twenty-second (2012) International Offshore and Polar Engineering Conference, Rhodes, Greece, June 17-22, 2012*, pp. 41-44

Appendix B: Neural Network Topology Diagram

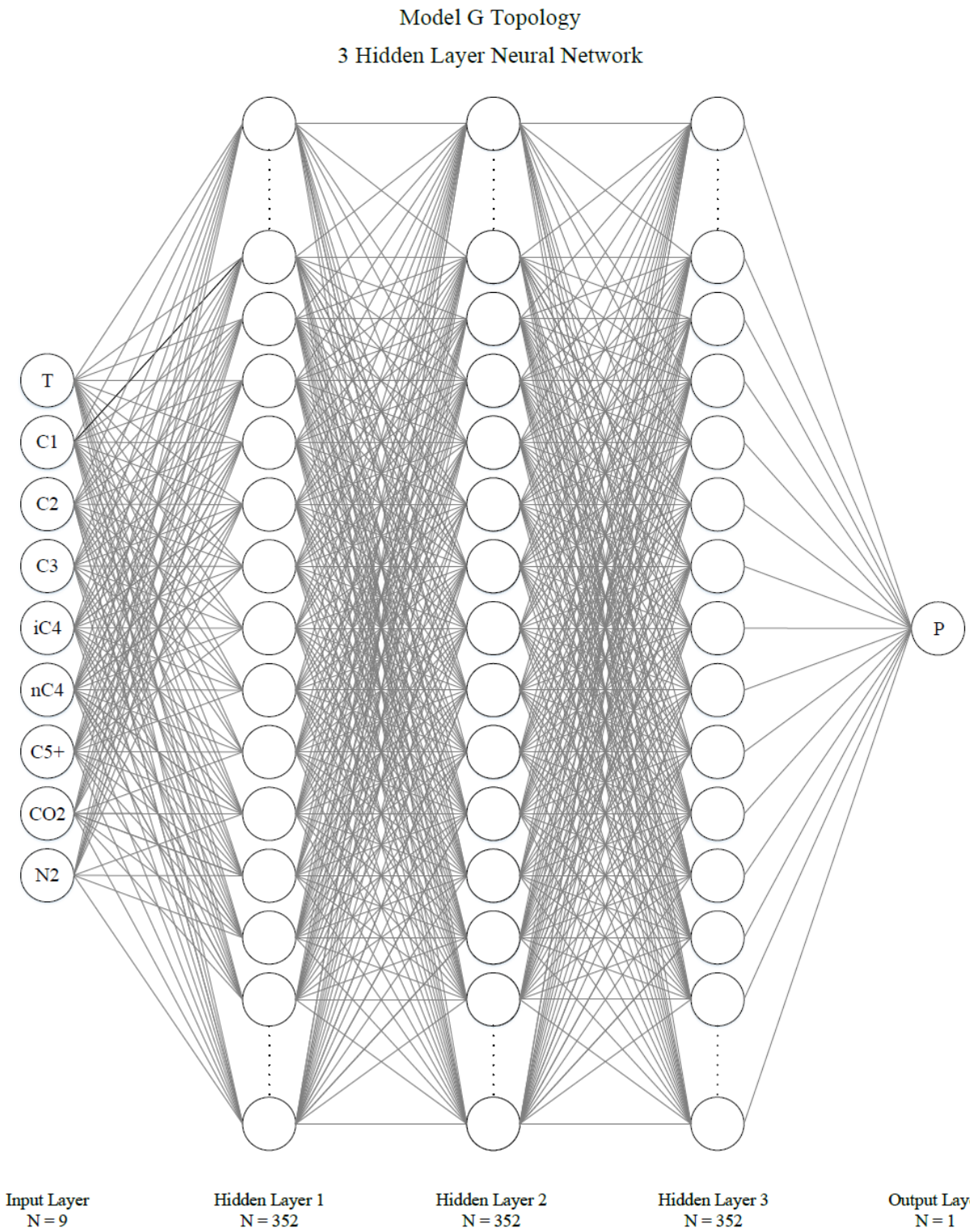

Figure B.1: Neural network topology diagram for Model G.

Appendix C: Distribution of Complete Dataset Data

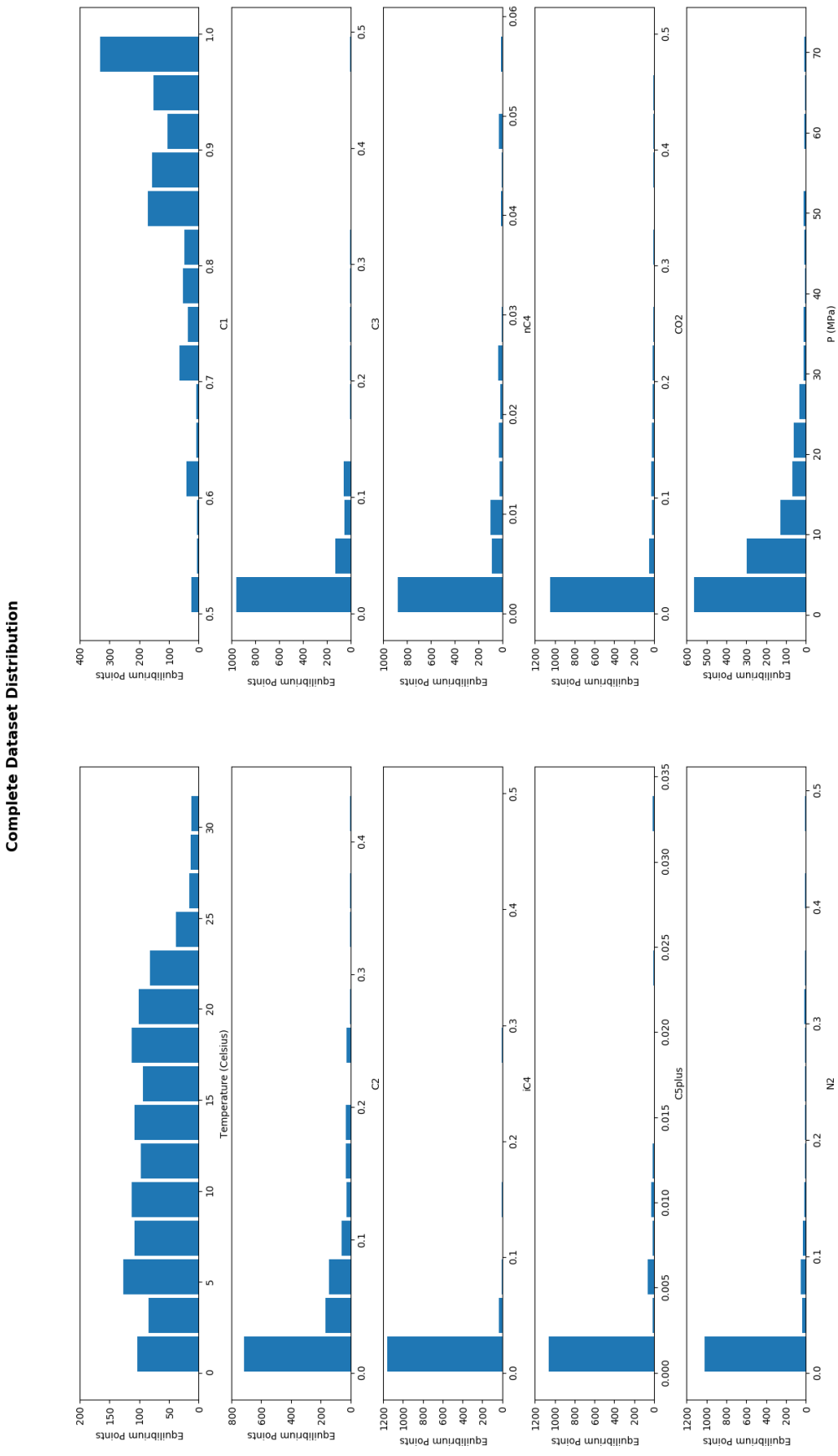

Figure C1: Distribution of data for the complete dataset. Note that the distribution merely indicates the range of individual dataset features, as visualizing the entire range in 10 Dimensions not within the scope of this report. Furthermore, note the scaling for compositions ranging between C2 and N2, a small number of high concentrations provided in the form of binary methane mixtures have been included for the sake of assisting the model in predicting sII hydrate behaviour. Most data for these ranges are well below 1% in terms of molar composition, so as to account for natural gas mixtures.