

Data-Driven Prediction of Gas Hydrate Formation under Thermodynamic Inhibitor Conditions

by

Assain Kaldybayev

THESIS SUPERVISOR

Dr. Masoud Riazi

THESIS CO-SUPERVISORS

Dr. Shams Kalam

Dr. Ali Rasoolzadeh

Thesis submitted to the School of Mining and Geosciences of Nazarbayev
University in Partial Fulfillment of the Requirements for the Degree of
Master of Science in Petroleum/Mining Engineering/Geosciences

Nazarbayev University

14.04.2026

ORIGINALITY STATEMENT

I, **Assain Kaldybayev**, hereby declare that this submission is my own work and to the best of my knowledge it contains no materials previously published or written by another person, or substantial proportions of material which have been accepted for the award of any other degree or diploma at Nazarbayev University or any other educational institution, except where due acknowledgement is made in the thesis.

Any contribution made to the research by others, with whom I have worked at NU or elsewhere is explicitly acknowledged in the thesis.

I also declare that the intellectual content of this thesis is the product of my own work, except to the extent that assistance from others in the project's design and conception or in style, presentation and linguistic expression is acknowledged.

Signed on **14.04.2026**

Assain Kaldybayev

ABSTRACT

Gas hydrates are crystalline ice-like solids, which develop at high pressure and low temperature, and are one of the most severe flow assurance problems in natural gas pipelines. When they grow out of control, they could block pipelines and pose significant safety, financial and environmental risks. The most widespread method of prevention is using thermodynamic hydrate inhibitors (THIs), which can include inorganic salts of NaCl, and MgCl₂ or organic inhibitors like methanol and glycols. These inhibitors act by reducing the hydrate equilibrium temperature. Thus, the correct determination of hydrate formation temperature when using inhibitors is a major engineering requirement. Nevertheless, classical thermodynamic theories grounded in the van der Waals-Platteeuw framework necessitate large-scale parameterisation of the system under consideration, whereas all previous machine learning methods have not addressed the issue of whether the experimental data used to train the model is thermodynamically consistent.

This thesis constructs thermodynamically informed, data-based model of gas hydrate formation temperature prediction across a broad spectrum of inhibitor types, concentrations, gas compositions, and operating conditions. The methodology assumes two steps, the initial step is a strict thermodynamic consistency filtering of literature-based experimental data and the second step is a systematic creation of supervised machine learning models.

Dataset containing 1751 experimental data points of inhibited hydrate phase equilibrium was gathered out of published works. These data involved 11 types of inhibitors and 6 components of a gas system. As a compulsory data quality filter, the three-criterion thermodynamic consistency approach, a method that measures Clausius-Clapeyron linearity, heat of dissociation consistency, water activity consistency, was applied. This step left 1167 thermodynamically consistent data points. Following the preprocessing, the resulting training dataset had 922 data points under pressure between 0.23 and 73.28 MPa as well as temperatures between 234.03 and 291.4 K.

Six machine learning models were trained and compared, including Support Vector Regression, Artificial Neural Network, XGBoost, CatBoost, Gradient Boosting Regression, Random Forest, and Decision Tree. Their cross-validated train-test splits were used to evaluate their performance in terms of R^2 , RMSE, MAPE, AIC, and BIC. The best overall performance was obtained with SVR using a radial basis function kernel with a test R^2 of 0.989 and test RMSE

of 0.963 K., and importantly, it exhibited better generalisation when performing supervised interpolation at inhibitor concentrations absent in the training set thus giving physically realistic and monotonically ordered equilibrium curves which can be explained by the smoothness nature of the model.

The four major thermodynamic trends that were physically validated to be reproduced using the SVR model with no explicit thermodynamic constraints in the training phase were a positive slope of pressure-temperature equilibrium, the correct sequence of inhibition curves in the order of concentration, the additive temperature depression in mixed inhibitor systems, and the monotonic rise in hydrate suppression temperature with increasing inhibitor concentration. These results demonstrate that a combination of thermodynamic consistency screening with careful feature engineering and an appropriately regularised machine learning model can lead to a physically trustworthy and flow-relevant predictive device of inhibited gas hydrate systems.

DEDICATION

To everyone who contributed to this work, my family and friends.

ACKNOWLEDGMENT

I would like to express my acknowledgement and sincere gratitude to my supervisor, Dr. Masoud Riazi, my co-supervisors, Dr. Shams Kalam and Dr. Ali Rasoolzadeh for their vital guidance, continuous support. dedicated efforts throughout the course of this research. I am grateful for the time and commitment they invested in this research. Without their expert advice, constructive feedback and contribution this research would not have been possible.

TABLE OF CONTENTS

TABLE OF CONTENTS.....	vii
LIST OF FIGURES	viii
LIST OF TABLES	x
1. INTRODUCTION.....	1
1.1 Background/Problem Definition	1
1.2 Objectives of the Thesis	9
1.2.1 <i>Main Objectives</i>	9
1.2.2 <i>Specific Objectives</i>	9
1.5 Scope of Work	10
2. LITERATURE REVIEW.....	13
3. METHODOLOGY.....	34
4. RESULTS.....	66
5. DISCUSSION	87
6. CONCLUSIONS AND RECOMMENDATIONS	98
7. REFERENCES.....	103
8. APPENDICES	108

LIST OF FIGURES

Figure 1. Gas hydrate plug recovered from a subsea pipeline (Koh et al. 2011)	1
Figure 2. Three natural gas hydrate crystal structures: (a) sI, (b) sII, (c) sH (Hassanpouryouzband et al. 2020)	2
Figure 3. Schematic of three-stage pipeline blockage process (Wang et al. 2019).....	3
Figure 4. Thermodynamic consistency check framework (three-criteria schematic) (Sa et al. 2018)	8
Figure 5. P-T schematic of a subsea pipeline with hydrate stability region (Koh et al. 2011).....	15
Figure 6. Full methodological pipeline flowchart.....	34
Figure 7. Distribution of the dataset by inhibitor type	36
Figure 8. Distribution of the dataset by gas composition.....	36
Figure 9. EDA outputs: correlation heatmap after preprocessing	47
Figure 10. EDA outputs: box plots for Pressure input.	48
Figure 11. EDA outputs: box plots for Gas composition input.....	48
Figure 12. EDA outputs: box plots for Salts Inhibitors inputs	48
Figure 13. EDA outputs: box plots for Alcohols Inhibitors inputs	49
Figure 14. EDA outputs: box plots for Nitrogen-based Inhibitors inputs	49
Figure 15. EDA outputs: box plots for Complex Inhibitors inputs	49
Figure 16. Random Forest and gradient boosting schematic comparison.....	59
Figure 17. Schematic of SVR ϵ -tube (A) and kernel mapping (B)	61
Figure 18. Model comparison bar charts: R^2 , RMSE, MAPE (train vs test) for all 6 models.....	68
Figure 19. AIC and BIC comparison bar charts (train vs test) for all 6 models.....	69
Figure 20. SVR Test: Experimental vs Predicted Temperature (K) scatter plot	72
Figure 21. SVR Train: Experimental vs Predicted Temperature (K) scatter plot	73
Figure 22. SVR P-T Equilibrium Curve Comparison plots at 0 wt% MeOH	74
Figure 23. SVR P-T Equilibrium Curve Comparison plots at 5 wt% MeOH	75
Figure 24. SVR P-T Equilibrium Curve Comparison plots at 10 wt% MeOH	76
Figure 25. SVR P-T Equilibrium Curve Comparison plots at 20 wt% MeOH	77
Figure 26. Interpolation P-T plots at 0 wt% MeOH.....	79
Figure 27. Interpolation P-T plots at 5 wt% MeOH.....	80
Figure 28. Interpolation P-T plots at 10 wt% MeOH.....	80
Figure 29. Interpolation P-T plots at 20 wt% MeOH.....	81
Figure 30. ΔT vs MeOH concentration trend analysis	82
Figure 31. SHAP Summary Plot: Feature Importance (bar chart)	83
Figure 32. SHAP Summary Plot: Feature Effects (beeswarm).....	85
Figure 33. EDA outputs: correlation heatmap before preprocessing	108

Figure 34. EDA outputs: distribution analysis	109
---	-----

LIST OF TABLES

Table 1. Summary of experimental data sources for the compiled dataset	25
Table 2. Fitting parameters from Rasoolzadeh et al. (2025) used for single-component uninhibited equilibrium temperature calculations	38
Table 3. Cages specifications (Rasoolzadeh et al. 2025)	39
Table 4. The acentric factors and the critical properties of the hydrate formers (Rasoolzadeh et al. 2025)	40
Table 5. Kihara potential parameters for various hydrate formers (Rasoolzadeh et al. 2025)	42
Table 6. Classification of consistency criterion 1	44
Table 7. Classification of consistency criterion 2	45
Table 8. Classification of consistency criterion 3	45
Table 9. Hyperparameter search space — Decision Tree (DT)	53
Table 10. Hyperparameter search space — Random Forest (RF)	54
Table 11. Hyperparameter search space — Gradient Boosting Regressor (GBR)	56
Table 12. Hyperparameter search space — XGBoost (XGB)	57
Table 13. Hyperparameter search space — CatBoost	59
Table 14. Hyperparameter search space — Support Vector Regressor (SVR)	61
Table 15. Performance metrics for all six ML models: final tuned run (training and test sets)	66

1. INTRODUCTION

1.1 Background/Problem Definition

1.1.1 Gas Hydrates: Nature and Industrial Relevance

Gas hydrates, also known in the scientific literature as clathrate hydrates, are the non-stoichiometric ice-like crystalline compounds, which are formed when small gas molecules are trapped in cage-like structures composed of hydrogen-bonded water molecules. These solid inclusion compounds are formed at high pressure and low temperature wherever there are adequate amounts of both molecules of water and hydrate-forming gas (Schicks, 2022; Hassanpouryouzband et al., 2020). The encased guest molecule stabilises the cage-like structure of the clathrate hydrate by van der Waals forces, and the water framework by hydrogen bonding. Since gas hydrates are non-stoichiometric, the proportion of water to guest molecules, the hydration number, varies with the thermodynamic conditions of formation and the size of the encasing species (Nasrifar and Moshfeghian, 2001).



Figure 1. Gas hydrate plug recovered from a subsea pipeline (Koh et al. 2011)

It has been possible to identify three naturally occurring and industrially important hydrate crystal structures, structure I (sI), structure II (sII) and structure H (sH). Structure I is the type that generally occurs with smaller guest molecules between the size of 0.4 and 0.55 nm, including methane and carbon dioxide, and is the most common structure of gas hydrate in nature. Structure II is typically found with larger guest molecules of 0.6 to 0.7 nm like propane and iso-butane, and is most common in engineering applications. The structure H is

characterized by the simultaneous presence of both large and small molecules (0.8 -0.9 nm) and is only observed under special conditions (Sloan, 2003; Hassanpouryouzband et al., 2020). Of special interest to natural gas transportation systems, which contain mixtures of multicomponent gases, is the fact that co-presence of multicomponent gas mixtures containing methane and ethane can facilitate structural transitions, e.g. some binary mixtures of structure I formers can give rise to structure II hydrates, highlighting the thermodynamic complexity of multicomponent systems (Sloan, 2003).

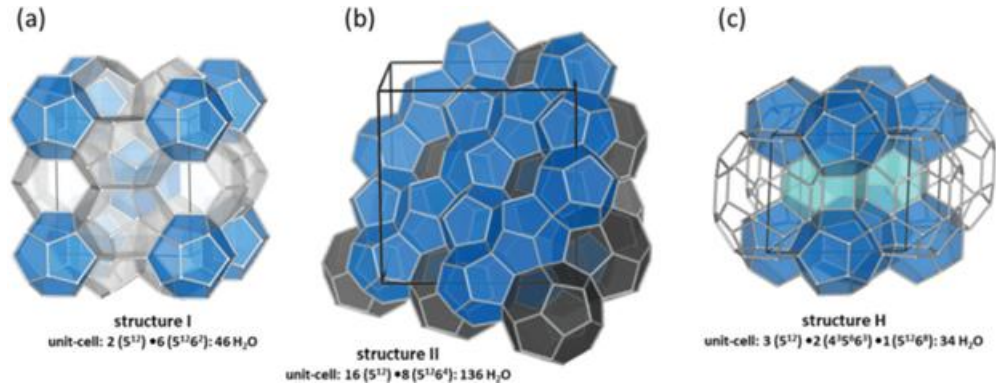


Figure 2. Three natural gas hydrate crystal structures: (a) sI, (b) sII, (c) sH (Hassanpouryouzband et al. 2020)

The initial period in gas hydrate research was preoccupied with industrial anxieties which arose due to the realization by Hammerschmidt in 1934 that the gas transmission pipelines were blocked not by ice as assumed but by natural gas hydrates (Koh et al., 2011; Schicks, 2022). This finding essentially shifted scientific interest towards the nature of conditions in which hydrates grow and how their development could be foreseeable and managed with some certainty. Gas hydrates have come to be set down as the main challenge of flow assurance which is described as the engineering field that is involved in making sure that the flow of hydrocarbon fluids that are found in the reservoir is continuous and economical to flow to the surface processing facilities (Koh et al., 2011).

Deepwater offshore is particularly dangerous to the risk of hydrate formation. Deepwater offshore oil fields are defined by increased high-pressure and low-temperature levels and the threat of gas hydrate blockage of pipelines and offshore plants is thus severe (Koh et al., 2011). Even the most conservative engineering evaluations admit that gas hydrates are a continuing threat during the lifetime of deepwater production systems. The effects of hydrate blockage are not just limited to interruptions in production. Gas hydrate formation was directly associated with the failure of a 100-tonne containment structure used in the emergency response operation in the Gulf of Mexico in 2010 in the aftermath of the Macondo well blowout, and led to the use

of subsequent containment strategies incorporating effective hydrate prevention measures (Koh et al., 2011). This incident demonstrates that the risks related to uncontrolled hydrate formation are not only financial in characteristics but have direct safety and environmental security implications.

A three-stage blockage mechanism is a characteristic that has been outlined in the literature in the context of natural gas pipeline systems. Gas and water move in the same direction in the pipeline, thus gas and water become separated in the phase resulting in hydrate and hydrate slurry. The formation of the hydrate uses water in the flowing stream and as the ratio of solid and liquid in the hydrate slurry exceeds a critical value, the pipeline will be blocked by the decrease in the liquid phase. Moreover, water-gas contact can lead to the formation of a hydrate film on the pipe wall that gradually gets thicker over time and gradually decreases an effective pipe diameter (Wang et al., 2019). The mechanism is different in oil-based pipelines; the blockage follows hydrate nucleation, crystal growth, and agglomeration, and finally hydrate plug formation that occurs due to the aggregates (Hassanpouryouzband et al., 2020). The prevention and control of hydrate formation in pipelines has been approximated to cost the industry over \$100 million annually (Kashish et al., 2024), not to mention the cost of production delays, tool damage, high labour costs, and imposition of serious safety hazards.

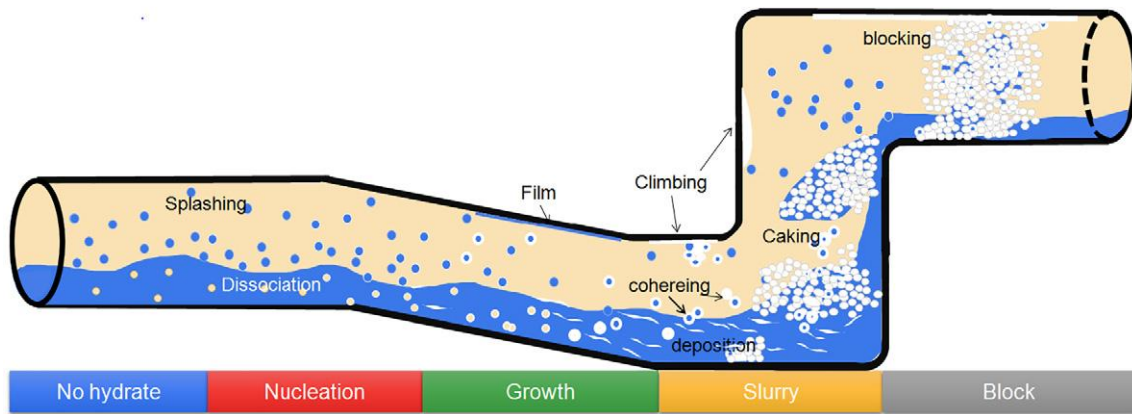


Figure 3. Schematic of three-stage pipeline blockage process (Wang et al. 2019)

1.1.2 Thermodynamic Hydrate Inhibitors: purpose and importance.

Due to the scale of the industrial hydrate-related operational risk, various mitigation measures have been established and implemented. There are physical methods like the insulation of the pipeline, heating, depressurisation, and dehydration of the stream of gases and the injection of chemical hydrate inhibitors (Wang et al., 2019; Kashish et al., 2024). In high-pressure and low-

temperature pipelines (over long distances), heating and insulation become more expensive, and full removal of water in the gas stream is not always technologically and economically feasible (Wang et al., 2019).

Injection of thermodynamic hydrate inhibitors (THIs) is the most popular industrial method of controlling the conditions of hydrate formation in natural gas pipelines. THIs are chemical substances which change the thermodynamic conditions of hydrate formation by moving the hydrate equilibrium curve to lower temperatures and higher pressures - effectively expanding the safe operating conditions of a particular pipeline system (Wang et al., 2019). The inhibitory effect of THIs is possible mainly by direct contact with water molecules in the aqueous phase: organic inhibitors like methanol and glycols create hydrogen bonds with water molecules, decreasing the activity of liquid water, and inorganic electrolytes like sodium chloride, calcium chloride, and potassium chloride bind water molecules by electrostatic interactions in the solvation shell of dissolved (Sa et al., 2018; Wang et al., 2019).

Some of the most common organic THIs utilized in natural gas production are methanol (MeOH) and monoethylene glycol (MEG). Methanol's advantage lies in its relatively low cost and ability to inhibit hydrate formation at very low concentrations. However, methanol's high volatility, toxicity, and tendency to evaporate from the hydrate plug during operation severely limits its application to intermittent usage situations like well restarts and hydrate plugging (Wang et al., 2019). Compared to MeOH, monoethylene glycol and diethylene glycol are chemically stable, less toxic, and exhibit a much lower solubility in the hydrocarbon gas phase. Therefore, they are preferred for continuous injection systems in long distance pipeline (Wang et al., 2019). Triethylene glycol (TEG) has been used, mostly in applications requiring a larger temperature suppression window. In terms of inorganic inhibitors, the order of the inhibition efficiency of cations is: $Mg^{2+} > Ca^{2+} > Na^{+} > K^{+}$. Furthermore, the inhibition efficiency of the chloride anion is significantly higher compared to the sulphate anion (Wang et al., 2019). The extent of temperature depression achievable using a specific inhibitor depends solely upon the inhibitor's concentration within the aqueous phase. Thus, accurate determination of hydrate formation conditions as a function of inhibitor type and concentration is a necessary engineering consideration in designing flow assurance designs.

In addition to single-component inhibitor systems, many field operations utilize combined inhibitor systems, i.e. the combination of organic inhibitors with inorganic salts, or combinations of glycols with electrolytes, which are introduced intentionally or unintentionally

through formation waters. The thermodynamic description of such mixed inhibitor systems is significantly more complicated than the description of single-component inhibitor systems, mainly due to the complexities associated with describing the thermodynamics of the aqueous phase and due to the nonlinear interactions between multiple inhibitor species (Nasrifar and Moshfeghian, 2001; Cai et al., 2025).

1.1.3 Predictive Approaches for Hydrate Phase Equilibrium: From Classical Thermodynamics to Machine Learning

Hydrate formation conditions have been predicted using three major types of methodologies: classical thermodynamic models, empirical correlation models, and more recently machine learning methods.

The theoretical basis for classical hydrate thermodynamics was established by van der Waals and Platteeuw in 1959 with the development of a statistical thermodynamic model for clathrate hydrates (Cai et al., 2025; Rasoolzadeh et al., 2025) that has remained the most widely used model for hydrate phase equilibrium calculations in both academic research and industry (Cai et al., 2025; Rasoolzadeh et al., 2025). The van der Waals-Platteeuw (vdW-P) model assumes a number of simplifications: no deformation of the hydrate lattice by encapsulated molecules, only single occupancy of each cavity by guest molecules, no interaction among guest molecules occupying adjacent cavities, and no quantum mechanical effects (Cai et al., 2025). Within this framework, the chemical potential difference of water between the empty hydrate lattice and the aqueous phase is described as a function of temperature, pressure and Langmuir constants quantitatively describing interactions between guest molecules and cavities. For inhibited systems, the vdW-P model is linked with an aqueous phase activity model, usually an excess Gibbs energy model or an electrolyte equation of state to describe changes in water activity caused by the presence of THIs (Nasrifar and Moshfeghian, 2001). One important contribution made by employing the Kihara potential for evaluating Langmuir coefficients continues to be extensively used today. Additional contributions have included attempts to include lattice expansions/distortions, multiple layers/shells, and non-spherical shapes of guest molecules into Langmuir coefficient expressions (Medeiros et al., 2020).

While the vdW-P model is theoretically robust and has been widely accepted in academia and industry alike, it has certain inherent limitations that make it unsuitable for modeling hydrate formations under complex conditions experienced in actual natural gas systems. Specifically,

the vdW-P model is highly sensitive to values for Kihara potential parameters and reference properties (Rasoolzadeh et al., 2025), therefore sensitivity to these parameters is increased in predicting hydrate conditions for multi-component gas mixtures and for inhibited systems containing complex mixed inhibitor solutions. Additionally, the computational implementation of the full vdW-P model involves solving a set of coupled equations involving: (i) the vapor phase EOS; (ii) Langmuir coefficient integrals for the hydrate phase; and (iii) an activity coefficient model or electrolyte EOS for the aqueous phase, which is computationally expensive and requires careful management of model parameters for each system. Although new extensions have been proposed as being rigorous, it has been noted that because the vdW-P model has proven successful, but that there are still opportunities for improvement (Medeiros et al., 2020).

A second class of predictive tools includes empirical correlations that trade-off some thermodynamic rigor for ease-of-computation. An example is the Hammerschmidt equation for hydrate temperature depression. Another example includes a series of correlations developed by Østergaard et al. for predicting temperature depressions caused by a variety of THIs. These correlations have had widespread applicability in industry for making initial estimates quickly. However, as stated previously, empirical correlations are generally limited in their accuracy outside of the conditions for which they were originally developed. Also, empirical correlations do not easily allow the representation of complex multicomponent gas systems with various mixed inhibitors (Cai et al., 2025).

In this context, Machine Learning (ML) methods represent a new class of predictive tools. ML methods can predict hydrate formation conditions rapidly. They require no explicit thermodynamic parameterization. They can potentially capture any nonlinear relationship between experimental variables using training data representing any range of experimental conditions. As noted above, numerous ML models have been explored in the literature to predict hydrate phase equilibria. Most ML models are classified into categories including Tree-Based Algorithms, Neural Networks, Support Vector Machines and Reinforcement Learning (Cai et al., 2025). Several studies have demonstrated that ML models can predict hydrate phase equilibria accurately. Yarveicy and Ghiasi employed an Extremely Randomized Trees algorithm to predict hydrate phase equilibria in pure water and in solutions containing electrolytes/alcohols with an average absolute relative deviation of 0.04 (Cai et al., 2025). Eslamimanesh et al. demonstrated that LSSVM could produce small deviations when

estimating phase equilibria for hydrates formed by CH₄, CO₂, N₂ and H₂ with water soluble promoters (Cai et al., 2025). Song et al. used Gradient Boosted Regression Trees and Deep Neural Networks to predict clathrate hydrate phase equilibria (Cai et al., 2025). More recently, XGBoost was used to predict gas hydrate phase equilibria in solutions containing electrolytes with $R^2 = 0.9956$ over data sets having more than 1800 data points (Cai et al., 2025).

Similar to classical models however, machine learning models present significant limitations. Feature extraction plays a crucial role in machine learning models. Especially when gas mixtures contain multiple components and when complex chemical additives are used. Additionally, although machine learning models have been successful for predicting phase equilibria for multicomponent systems, the data driven nature of machine learning means that once training is complete these models cannot be extrapolated to new systems that have additional chemicals that inhibit or promote hydrate formation (Cai et al., 2025).

One critical issue associated with all machine learning-based hydrate prediction studies published to date is quality/thermodynamic consistency of experimental databases upon which models are trained. Since experimental measurements of hydrate phase equilibria are susceptible to uncertainty stemming from inaccuracies in pressure and temperature measurement techniques, timescales required for achieving equilibrium, sample preparation protocols and variability between research groups' experimental protocols, inaccurate or thermodynamically inconsistent experimental data points can result in systematic error in the mapping of features to output generated by machine learning models. Given that hydrate phase equilibria exist throughout literature as a vast database; no formal method currently exists to evaluate thermodynamic consistency of measured data; thus, a formal method would be beneficial (Sa et al., 2018). Regardless of the quality or accuracy of any predictive model whether classical or machine learning, the predictive capability is strictly limited by quality/experimental accuracy/consistency of experimental databases upon which it was built (Sa et al., 2018).

To address this critical shortcoming in ML-based hydrate prediction studies, Sa et al. (2018) proposed a rigorous three-criteria methodology for assessing thermodynamic consistency of experimentally determined hydrate phase equilibrium data. The three criteria are based on fundamental thermodynamic relations as follows: (i) Linear regression consistency in phase equilibrium curves plotted in $\ln P$ vs. $1/T$ space per the Clausius Clapeyron equation; (ii) Consistency of the heat of dissociation (ΔH_d), where ΔH_d should be essentially independent of

the type/concentration of THI since THIs affect the aqueous phase rather than causing structural deformation of hydrate crystals and (iii) Consistency of the water activity (a_w) which should be monotonic increasing with increasing inhibitor concentration as temperature depression increases relative to pure hydrate equilibrium conditions (Sa et al., 2018). These three criteria collectively establish a basis for assigning "pass," "acceptable," or "fail" labels to individual data sets based on user defined tolerance ranges. It is demonstrated that this screening process provides a meaningful benefit in terms of improving model performance — specifically for uninhibited systems — as shown by Rasoolzadeh et al. (2025) who demonstrated that removing inconsistent data points from a set of phase equilibrium data resulted in an improvement from 0.61 K & 6.52 % to 0.20 K & 2.62 % AAD/AARD respectively when predicting dissociation temperatures for C_2H_6 hydrates.

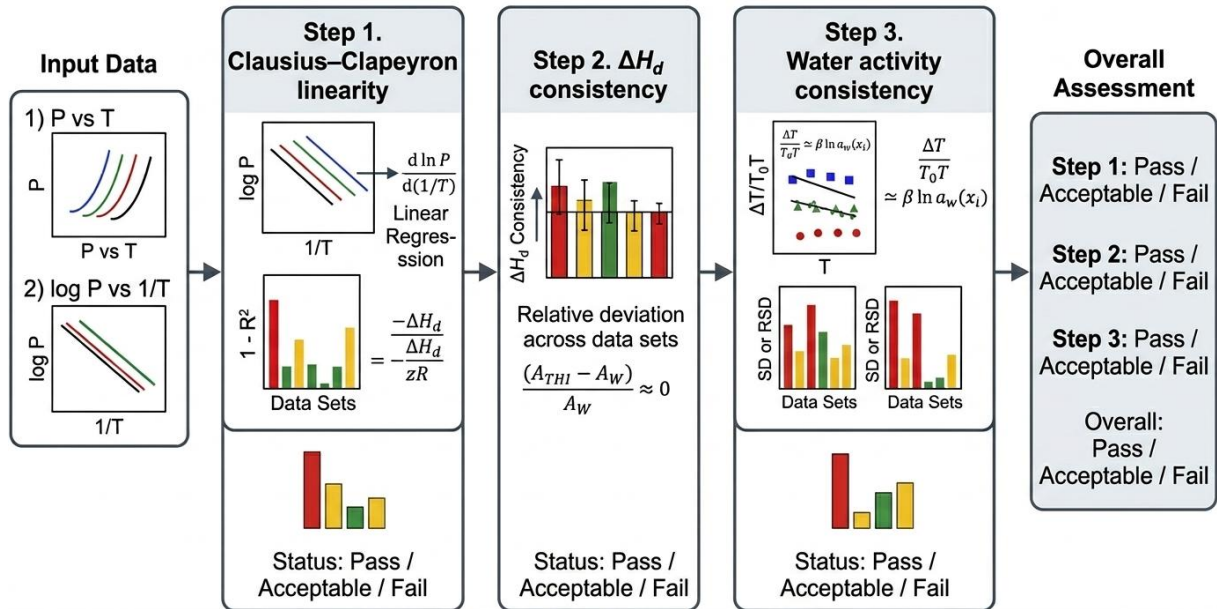


Figure 4. Thermodynamic consistency check framework (three-criteria schematic) (Sa et al. 2018)

An added complication to performing this type of screening is determining uninhibited hydrate equilibrium temperatures as a reference point for assessing consistency. Rasoolzadeh et al. (2025) provided a straightforward thermodynamic model for determining dissociation conditions for simple gas hydrate formers such as CH_4 , C_2H_6 , C_3H_8 , CO_2 , N_2 , H_2S & O_2 based on slight variations in enthalpies of dissociation with respect to temperature as expressed through Equation 10 provided in Rasoolzadeh et al. (2025) as a means for determining dissociation temperatures as functions of pressures. Van der Waals—Platteeuw Model provided

by Rasoolzadeh et al. (2025) can be used to determine uninhibited equilibrium temperatures as inputs for performing the consistency evaluation.

Although the Sa et al. (2018) consistency methodology has been available for several years along with an ever-growing body of literature related to ML-based hydrate prediction; no prior publication has performed systematic screening of thermodynamic consistency as a preliminary step to ensure quality/data consistency before developing ML models to predict inhibited hydrate phase equilibria. This represents a substantial methodological gap that will be addressed directly by this dissertation effort.

1.2 Objectives of the Thesis

1.2.1 Main Objectives

The main objective of this thesis is to develop a robust, thermodynamically informed, data-driven predictive pipeline for estimating gas hydrate formation temperature under thermodynamic inhibitor conditions. This pipeline integrates rigorous thermodynamic consistency screening of experimental literature data as a prerequisite data quality step, followed by supervised machine learning model development and evaluation using a comprehensively curated dataset covering multiple inhibitor types, inhibitor concentrations, and both single-component and multicomponent gas hydrate formers.

1.2.2 Specific Objectives

To achieve the main objective, the following specific objectives are defined:

Objective 1: Data collection and consolidation. To systematically collect experimental gas hydrate phase equilibrium data from the published literature for inhibited systems, encompassing a broad range of thermodynamic inhibitors including inorganic salts (NH_4Cl , KCl , MgCl_2 , NaCl , CaCl_2) and organic inhibitors (methanol, propanol, ethylene glycol, monoethylene glycol, diethylene glycol, and DMSO), across diverse gas compositions including CH_4 , C_2H_6 , C_3H_8 , C_4H_{10} hydrocarbons, N_2 , and CO_2 , in both single-component and multicomponent configurations.

Objective 2: Thermodynamic consistency screening. To apply the three-criterion thermodynamic consistency methodology of Sa et al. (2018), encompassing Clausius-Clapeyron linearity regression, heat of dissociation consistency, and water activity consistency

to all collected data, employing the uninhibited equilibrium temperature calculation procedure of Rasoolzadeh et al. (2025) as the baseline reference for single-component gas hydrate formers, and the van der Waals–Platteeuw model for multicomponent systems, in order to retain only thermodynamically reliable data points for model training.

Objective 3: Machine learning model development and comparison. To develop, train, and systematically evaluate a suite of six machine learning models: Support Vector Regression (SVR), XGBoost, CatBoost, Gradient Boosting Regression (GBR), Random Forest (RF), and Decision Tree (DT) for the prediction of gas hydrate formation temperature, employing appropriate hyperparameter optimisation, cross-validation, and a robust evaluation framework based on R^2 , RMSE, MAPE, AIC, and BIC metrics.

Objective 4: Best model identification and validation. To identify the best-performing model through rigorous statistical evaluation and to validate its predictions against independent experimental data from the literature, assessing whether model predictions are physically and thermodynamically consistent with expected behaviour across the full range of inhibitor types, concentrations, pressures, and gas compositions represented in the dataset.

Objective 5: Trend analysis and physical interpretation. To conduct systematic trend analysis of model predictions as a function of inhibitor type, inhibitor concentration, pressure, and gas composition, and to assess whether the predicted trends are directionally consistent with established thermodynamic expectations for the effect of each inhibitor class on hydrate formation temperature.

Objective 6: Supervised interpolation. To perform supervised interpolation within the verified dataset range, generating physically meaningful intermediate predictions in under-sampled regions of the feature space, and to use these results to demonstrate the practical utility of the trained model for engineering estimation tasks beyond the directly measured conditions.

1.3 Scope of Work

This thesis focuses on the prediction of gas hydrate formation temperatures using data-based methods with thermodynamic hydrate inhibitors, and under conditions relevant to oil and gas flow assurance. The dataset that was used for developing ML models was obtained from existing published literature. No additional experimental data was generated during the course of this research project. Prior to consistency checking, there were 1751 data points in the

original dataset. Following application of the thermodynamic consistency check developed by Sa et al. (2018) (with respect to hydrate inhibitor types), 1167 data points remained. Application of subsequent data cleaning procedures, i.e., duplicate removal, encoding, and feature engineering reduced the number of entries in the final cleaned dataset to 922 entries.

The resulting dataset represents a total pressure range of 0.23-73.28 MPa and a temperature range of 234.03-291.4 K. The types of hydrate inhibitors present in the final dataset are comprised of inorganic salts, e.g., NH_4Cl , KCl , MgCl_2 , NaCl and CaCl_2 , organic inhibitors, e.g., methanol, propanol, ethylene glycol, monoethyleneglycol, diethyleneglycol and dimethyl sulfoxide (DMSO); and gas mixtures consisting of methane (CH_4), ethane (C_2H_6), propane (C_3H_8), butane (C_4H_{10}), nitrogen (N_2) and carbon dioxide (CO_2). The gas compositions can be either single-component mixtures or multicomponent mixtures composed of all of the above components, which are representative of natural gas. Hydrate formation temperature is the target variable being predicted in this work based upon input variables of pressure, type and concentration of inhibitor, and composition of the gas mixture.

In contrast, kinetic hydrate inhibitors (KHIs) and anti-agglomerant additives (AAs), hydrates formed in porous media or sedimentary environments, and systems that contain significant amounts of highly soluble acidic gases at sufficiently high concentrations to cause substantial modification of the aqueous solution phase chemistry beyond the range where the selected hydrate inhibition models are applicable are outside the scope of this research project. In addition, the machine learning algorithms are intended to only provide predictions through interpolation over the domain spanned by the training data set. Therefore, it should not be assumed that their use for making predictions that require extrapolation beyond the range of the training data will produce accurate results without first obtaining some form of experimental verification.

2. LITERATURE REVIEW

2.1 Gas Hydrate Fundamentals and Phase Equilibrium

2.1.1 *Structural Features and Characteristics of Gas Hydrates*

Gas hydrates, also known as clathrate hydrates, are a group of non-stoichiometric crystalline solids made up of water molecules arranged in a hydrogen-bonded polyhedral cage-like framework, that include small gas molecules under high-pressure and low-temperature conditions as described by Schicks (2022). Unlike chemical hydrates, where gas molecules are chemically bonded to water molecules, gas hydrates have the ability to maintain the integrity of the hydrate lattice through van der Waals forces acting between the gas molecules trapped inside the cage-like frameworks and the surrounding water molecules (Schicks, 2022). This unique structural feature allows for the creation of three distinct crystal types that have been characterized in both laboratory-created environments and natural environments.

Hydrate Type I (sI) includes a cubic unit cell consisting of 46 water molecules organized into two small pentagonal dodecahedra (5^{12}) and six large tetrakaidecahedra ($5^{12}6^2$). Typically, hydrate type I forms with guest molecules having diameters less than 0.55 nm such as methane (CH_4), carbon dioxide (CO_2), and ethane (C_2H_6) described by Hassanpouryouzband et al. (2020). Hydrate Type II (sII) contains a cubic unit cell made up of 136 water molecules organized into 16 small 5^{12} cages and eight large hexakaidecahedral cages ($5^{12}6^4$), and thus accommodates larger guest molecules in the diameter range of 0.6 – 0.7 nm such as propane (C_3H_8) and isobutane ($\text{i-C}_4\text{H}_{10}$) (Koh et al., 2011; Sloan, 2003). Hydrate Type H (sH) includes a hexagonal unit cell made up of 34 water molecules organized into three small 5^{12} cages, two medium-sized $4^35^66^3$ cages, and one large icosahedral cage ($5^{12}6^8$), and therefore requires the presence of both small and large guest molecules (diameter = 0.8 – 0.9 nm) to stabilize the hydrate (Hassanpouryouzband et al., 2020). For example, sII hydrates are dominant in many natural gas pipelines due to the presence of methane along with smaller amounts of other larger hydrocarbon molecules like propane and isobutane. In fact, as little as 0.01 mol% propane can promote sII hydrate formation in a methane rich gas mixture (Koh et al., 2011).

Another important structural characteristic with direct thermodynamic implications is that gas hydrates are non-stoichiometric. That means that some cages within the hydrate lattice will remain unfilled, especially the smaller cages. As described in Koh et al. (2011) the degree of

cage filling is highly dependent on temperature, pressure, and size of the guest molecule. In particular, large cages throughout all hydrate structures tend to have about 100% occupation levels however, the occupation level of smaller cages can vary considerably depending on the specific type of guest molecules and can be either close to zero or close to full in certain situations (Hassanpouryouzband et al., 2020). Due to this lack of stoichiometry, the hydrate phase boundary is not simply a straight-line but rather a function of temperature, pressure, and composition of both the gas and aqueous phases, thereby necessitating thermodynamic modeling that accounts for cage occupations through Langmuir-type adsorption theories.

The physical properties of gas hydrates distinguish them quite obviously from those of both liquid water and ice, although they are primarily composed of water. The thermal conductivity of hydrate type I (about $0.57 \text{ W m}^{-1} \text{ K}^{-1}$ at 263 K) is significantly less than that of ice Ih ($2.21 \text{ W m}^{-1} \text{ K}^{-1}$), whereas the heat capacities, densities, and compressional wave velocities of both sI and sII are generally similar, indicating that hydrates are mechanically robust and acoustically detectable in both pipeline and sedimentary environments (Koh et al., 2011). These physical characteristics represent the underlying engineering challenge. Once formed, hydrate plugs in pipelines are dense and difficult to dissociate quickly, requiring active intervention through thermodynamic inhibitors to prevent their initial formation.

In situ characterization experiments conducted over the last twenty years have greatly enhanced our understanding of gas hydrate structural features. Techniques used for these experiments included X-ray diffraction (XRD), Raman spectroscopy, nuclear magnetic resonance spectroscopy (NMR), and neutron diffraction at high pressure (Hassanpouryouzband et al., 2020). These techniques allowed researchers to observe in real-time the evolution of cage fillings over time as a function of thermodynamic conditions and guest molecule properties. These structural understandings provide the physical foundation for the development of thermodynamic models that support hydrate predictions.

2.1.2 Thermodynamic Conditions Necessary for Hydrate Formation

Gas hydrates exist wherever there are sufficient pressure and low enough temperature to create a stable hydrate phase in conjunction with the presence of water and a suitable guest molecule. The thermodynamic stability region of a gas hydrate is defined by its three-phase equilibrium boundary: the locus of pressure-temperature (P-T) conditions at which the hydrate (H), liquid water (Lw), and vapor (V) phases coexist (Cai et al., 2025). At pressures/temperatures below

this boundary (lower T/higher P) the hydrate phase exists stably. Above this boundary (higher T/lower P) the hydrate phase decomposes into its constituent gas and liquid phases.

As demonstrated in Koh et al. (2011), the shape and location of the hydrate stability zone depend heavily on the composition of the gas. For instance, at a temperature of 277.6 K, the formation pressure for pure propane hydrate (sII) is approximately 0.43 MPa while that for pure methane hydrate (sI) is about 4.06 MPa. Similarly, at a composition ratio of methane to propane of 50:50, the hydrate formation pressure at 277.6 K is approximately 0.8 MPa. This demonstrates how even a relatively small proportion of a larger hydrocarbon can significantly decrease the formation pressure for a gas hydrate and enlarge the stability zone.

These structural promotional effects are utilized in various industrial applications, but they complicate predictive modeling of multicomponent natural gas systems. Figure 5 provided by Hassanpouryouzband et al. (2020) shows examples of hydrate stability zones for CO₂, N₂, CH₄, flue gas, and their mixtures as functions of temperature illustrating how adding gases that have different hydrate stability pressures changes the position of the phase boundary and modifies the overall geometry of the stability zone.

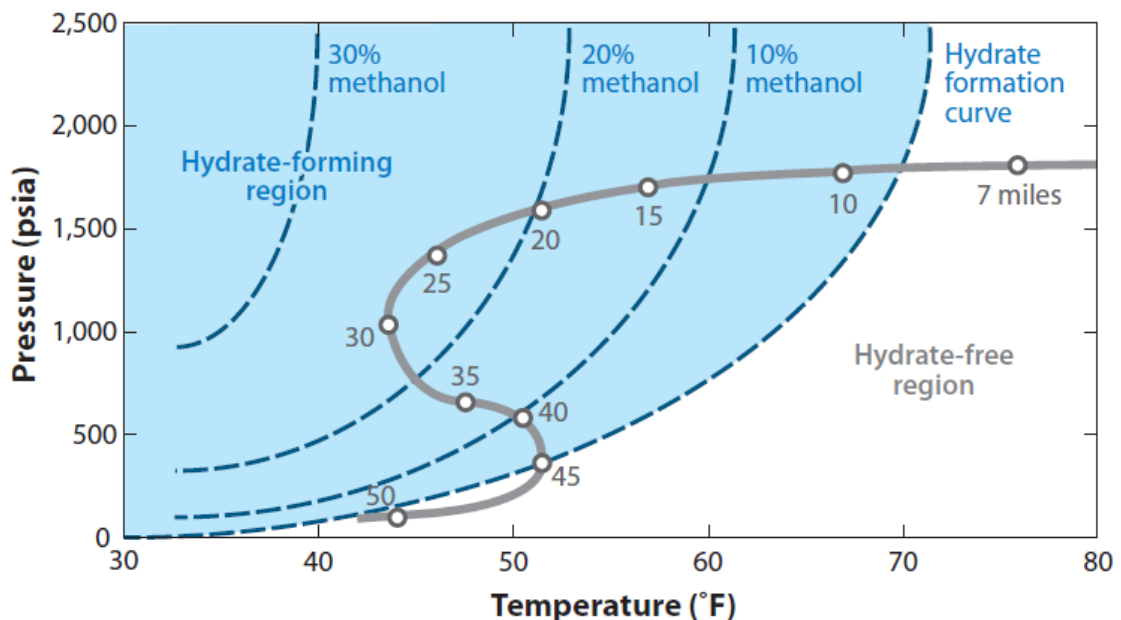


Figure 5. P-T schematic of a subsea pipeline with hydrate stability region (Koh et al. 2011)

For engineering purposes related to flow assurance, specifically for determining whether or not hydrates will form in a subsea pipeline transporting natural gas from a deepwater well to a surface processing facility at any given set of operational conditions (e.g., operating pressure,

gas composition, and aqueous phase composition), we need to determine if the fluid flow falls within the hydrate stability zone at any point along its path. Figure 5. depicts schematically the pressure-temperature profile of fluids flowing through a subsea pipeline from a deepwater wellhead to the processing facility with the hydrate stability envelope clearly indicated. As demonstrated in Koh et al. (2011), it is evident that considerable lengths of typical subsea pipelines lie entirely within the hydrate stability zone unless active intervention is taken to prevent hydrate formation by utilizing thermodynamic inhibitors.

2.1.3 Laboratory Experiments Determining Hydrate Phase Equilibrium Conditions

Accurate determination of hydrate phase equilibrium conditions serves as a foundation for developing all thermodynamic models and machine learning predictions related to gas hydrates. A total of two principal categories of laboratory experimentation has been employed in establishing experimental determinations of hydrate phase equilibria: laboratory experiments employing a pressure-volume-temperature (PVT) technique; laboratory experiments employing differential scanning calorimetry (DSC).

Both categories employ multiple variations. Isochoric step-heating experiments, currently employed as one of the most popular methodologies in open literature, involve loading a gas-water system into a fixed volume cell, cooling the contents to create a hydrated state followed by slowly heating the sample while simultaneously recording the pressure-temperature history (Cai et al., 2025). The equilibrium dissociation point occurs at the intersection point on the P-T diagram corresponding to the cooling curve and slow heating curves (Cai et al., 2025). The accuracy of isochoric experiments relies heavily on careful control of heating rates, quality of mixing during experiments, and availability of adequate equilibration times; an optimal heating rate should be less than 0.5 K/h (Cai et al., 2025), otherwise isolated crystals may form within the cell which exhibit greater stability than equilibrium crystals leading to inaccuracies in predicting dissociation conditions (Cai et al., 2025). Isothermal pressurization experiments involve maintaining a constant system temperature while increasing/decreasing system pressure in incremental steps until a visually identifiable three-phase equilibrium condition occurs via viewing through a transparent autoclave cell. Isobaric temperature increase/dissociation experiments involve changing system temperature at constant pressure. Isothermal and isobaric experiments allow for direct visualization of hydrate formation and dissociation processes; however, both types may lead to inaccuracies when dealing with multi-component systems or

inhibited aqueous solutions where direct visualization may be ambiguous or inconclusive (Cai et al., 2025).

Differential Scanning Calorimeter (DSC) measurements involve monitoring changes in heat flow resulting from changes occurring during temperature programmed scans associated with hydrate dissociation or formation; hence allowing simultaneous measurement of dissociation temperatures and dissociation enthalpies. DSC generates useful thermodynamic data, however, P-T conditions determined via DSC have exhibited increased variability compared to P-V-T conditions measured using traditional P-V-T methodology; especially in cases where DSC data were collected using inhibited systems due to non-equilibrium dissociation processes caused by insufficient agitation/mixing (Cai et al., 2025).

Regardless of experimentally employed methodologies, whether employing PVT or DSC, all experimental data must be free from internal inconsistencies. Even though both PVT and DSC methodologies are accepted as standard laboratory methodologies, many sources indicate that the data generated by both methods can vary greatly depending on numerous experimental variables, thus, inconsistent data may be produced among laboratories studying identical systems. Although previous studies have had varying degrees of success with these methodologies, a quantitative tool capable of evaluating consistently accurate thermodynamic data has yet to be developed; and recent studies have applied formal consistency evaluation methodologies toward addressing this issue (Cai et al., 2025; Sa et al., 2018).

2.2 Thermodynamic Models for Hydrate Phase Equilibrium Prediction

2.2.1 The van der Waals–Platteeuw Model

Statistical thermodynamics modelling for describing gas hydrate phase equilibria was pioneered by Van Der Waals & Platteeuw in their seminal paper published in 1959 entitled “A Theory for Polyuclear Molecules”, wherein Van Der Waals & Platteeuw presented what would become recognized as a landmark statistical thermodynamic treatment for clathrate hydrate phase behavior (van der Waals & Platteeuw, 1959). Their statistical mechanical treatment represents arguably one of the greatest successes in applying statistical mechanics to engineering problems involving clathrate hydrate systems (Medeiros et al., 2020; Cai et al., 2025).

Van Der Waals & Platteeuw treated the gas hydrate phase as a diluted solid-solution where water acts as both solvent lattice and guest molecules act as solutes adsorbed into cage-cavities according to Langmuir-type adsorption behavior [Medeiros et al., 2020]. The major result derived from Van Der Waals & Platteeuw's theoretical framework is an equation relating differences in chemical potentials of water between an empty hydrate lattice β -phase and a completely filled hydrate phase:

$$\frac{\Delta\mu_w^H}{RT} = \sum_m v_m \ln \left(1 + \sum_i C_{mi} f_i \right)$$

where v_m represents number-of-type- m cages per water molecule within a unit-cell; C_{mi} represents Langmuir constants for adsorbing guest species i onto cages-of-type- m ; and f_i represents fugacities for guest species i dissolved in vapor phase (Koh et al., 2011; Rasoolzadeh et al., 2025). Fractional occupancy θ_{mi} is directly obtainable from this relation using Langmuir-type expressions involving Langmuir constants and fugacity values. At phase equilibrium, chemical potentials for water dissolved in hydrate-phase must match chemical potentials for water dissolved in co-existing aqueous-phase; this constraint is used to establish an equilibrium criterion permitting calculation of hydrate forming conditions based upon thermodynamic properties.

The original framework formulated by Van Der Waals & Platteeuw rested upon several fundamental premises: no distortion of the water lattice by encaging gases, single occupancy of each cage with at most one guest species, no interactions between different guest molecules, and negligible quantum effects (Cai et al., 2025). In order to apply this framework to complex multicomponent gas systems Parrish & Prausnitz first extended the Van Der Waals & Platteeuw model in their paper in October 1972 through incorporation of Kihara cell potential for computation of Langmuir coefficients together with tuning parameters using empirical data (Hassanpouryouzband et al., 2020). Since that time all subsequent developments in hydrate thermodynamics have been founded upon this integration of vdW-P model and Kihara cell potential.

To determine if hydrate systems in the presence of thermodynamic inhibitors meet the vdW-P equilibrium criterion, the vdW-P criterion is generalized through the water activity term a_w in the aqueous phase chemical potential expression (Nasrifar and Moshfeghian, 2001; Rasoolzadeh et al., 2025). The complete equilibrium condition can then be expressed as follows:

$$\sum_m v_m \ln \left(1 + \sum_i C_{mi} f_i \right) = \frac{\Delta \mu_w^0}{RT_0} - \int_{T_0}^T \frac{\Delta h_w^{\beta-L}}{RT^2} dT + \int_0^P \frac{\Delta v_w^{\beta-L}}{RT} dP - \ln(a_w)$$

Where $\Delta \mu_w^0$ is the chemical potential difference of water between the empty hydrate and the liquid phase at reference conditions, $\Delta h_w^{\beta-L}$ is the corresponding enthalpy difference, $\Delta v_w^{\beta-L}$ is the molar volume difference, and a_w is the activity of water in the aqueous phase (Rasoolzadeh et al., 2025; Nasrifar and Moshfeghian, 2001). The water activity term is the key mechanism through which thermodynamic inhibitors influence the equilibrium calculation. When electrolyte inhibitors, like NaCl and CaCl₂, are introduced into the system, they decrease the activity of water through electrostatic binding of water molecules to the ions in their solvation shells, thus shifting the hydrate stability region towards lower temperatures. When organic inhibitors, such as methanol and glycols, are added to the system, a_w decreases through hydrogen bonding interactions between the inhibitor molecules and water. The enthalpy difference term $\Delta h_w^{\beta-L}$ is defined as a function of temperature based on (Rasoolzadeh et al., 2025):

$$\Delta h_w^{\beta-L} = \Delta h_{w0} + \int_{T_0}^T [-38.12 + 0.141(T - T_0)] dT$$

The necessary thermodynamic reference properties ($\Delta \mu_w^0$, Δh_w^0 , Δv_w) for both structure I and structure II cages, along with cage specifications, are provided in Rasoolzadeh et al. (2025) (Table 1) and are considered to be fundamental inputs to any vdW-P-based calculation. The accuracy of these reference values will directly affect the quality of the predicted equilibrium conditions.

2.2.2 Langmuir Constants and the Kihara Potential

Langmuir Constant $C_{mi}(T)$ measures the degree of interaction between a guest molecule and a water cage, and it is determined from the configuration integral of the cell potential:

$$C_{hj}(T) = \frac{4\pi}{kT} \int_0^{R'-a} \exp \left(-\frac{w(r)}{kT} \right) r^2 dr$$

where k is Boltzmann's constant, R' is the average radius of the cavity, a is the hard-core radius of the guest molecule, and $w(r)$ is a spherically symmetric cell potential function (Rasoolzadeh et al., 2025). The Kihara potential represents the contribution to the cell potential due to the finite hard-core size of the guest molecule and is represented by:

$$w(r) = 4\epsilon \left[\left(\frac{\sigma}{r - 2a} \right)^{12} - \left(\frac{\sigma}{r - 2a} \right)^6 \right]$$

McKoy and Sinanoglu developed this extension to represent non-spherical guest molecules (like ethane and CO₂) since spherical potentials do not accurately capture the size and shape characteristics of these molecules (Hassanpouryouzband et al., 2020). The Kihara parameters for all of the major natural gas hydrate formers (CH₄, C₂H₆, C₃H₈, CO₂, N₂, H₂S) used in this study were selected from Rasoolzadeh et al. (2025), which optimized these parameters against a large body of experimental data and ensured consistency among those data. The Langmuir constant calculation using these parameters and expressions for them is explained in detail in Sloan and Koh (2008) These equations were utilized within this study for consistency screening purposes.

2.2.3 Fugacities Using the Peng-Robinson Equation of State

The fugacity f_i of each guest species forming hydrates within the vapor phase needed as an input to the Langmuir occupancy equation is calculated using the Peng-Robinson equation of state. The PR EoS produces the compressibility factor and subsequently calculates the fugacity coefficient ϕ_i for each component in the gas mixture. From there, the fugacity can be determined by $f_i = \phi_i \times y_i \times P$, where y_i is the vapor-phase mole fraction and P is the total pressure. The overall fugacity calculation procedure including the mixing rules and derivation of the fugacity coefficient expression from the PR-EOS are described in detail in the fugacity document prepared for this project. The Peng-Robinson equation of state is commonly used within the oil-and-gas industry to predict fugacities for hydrate systems due to its performance regarding hydrocarbons and other small molecules across a wide variety of pressures and temperatures found during pipeline transportation (Rasoolzadeh et al., 2025).

When highly soluble acid gases (CO₂ and H₂S) are present within a hydrate system, their solubility in the aqueous phase can significantly impact both water activity and must be accounted for within any equilibrium calculation. Under such circumstances, estimates of CO₂ and H₂S solubility within the aqueous phase are made via the Krichevsky-Kasarnovsky equation. Furthermore, by applying the Lewis-Randall rule ($a_{vv} = x_{vv}$) to compute the water activity from this new mole-fraction value for water. The specific equation utilized within this study to perform this task described in Nasrifar and Moshfeghian (2001).

2.2.4 A Simplified Methodology for Evaluating Un-inhibited Hydrate Equilibrium Conditions

Although a fully accurate vdW-P methodology is generally required to evaluate multicomponent hydrate systems, a much simpler method for evaluating hydrate equilibrium conditions for single-component hydrate formers has been proposed by Rasoolzadeh et al. (2025). By recognizing that hydrate dissociation enthalpy ΔH_d exhibits minimal variation with temperature, these researchers created a simple analytical expression relating hydrate dissociation pressure P to temperature T :

$$P = \exp\left(\frac{A}{T} + \frac{B}{2T^2} + C\right)$$

and conversely an inverse relationship between hydrate dissociation temperature T and pressure P (Rasoolzadeh et al., 2025):

$$T = \frac{2A \mp \sqrt{4A^2 + 8B(\ln P - C)}}{4(\ln P - C)}$$

A , B , and C are empirical constants fit to experimental data for each hydrate former. This simplified model yields an average absolute error in temperature of 0.25 K and an average absolute relative error in pressure of 2.82% on 1879 experimental data points across seven hydrate formers (CH_4 , C_2H_6 , C_3H_8 , CO_2 , N_2 , H_2S , O_2) spanning a pressure range of 0.093 to 493 MPa (Rasoolzadeh et al., 2025). Therefore, this equation will be used to calculate uninhibited hydrate formation temperatures for individual single-component hydrate formers as a reference point for checking for thermodynamic consistency throughout this study, while the full vdW-P model will be used to evaluate multi-component hydrate systems.

2.2.5 Shortcomings in Classical Thermodynamic Models

Although classical vdW-P-type models have performed quite well for predicting hydrate equilibrium conditions, recent studies demonstrate that there exists additional room for improving upon these models (Medeiros et al., 2020). There are multiple types of shortcomings exhibited by these models which are particularly relevant to this thesis' focus on flow assurance applications. Firstly, classical models typically provide good predictions for inhibited systems that contain either little or no inhibitor. However, when high levels of inhibitors are used in combination with hydrate systems, there is evidence that classical models produce increasingly inaccurate predictions (Hassanpouryouzband et al., 2020). Secondly, classical models also

experience significant inaccuracies when determining hydrate equilibrium conditions at high pressures or with very small guest molecules. Additionally, when CO₂ or H₂S are present as one of two guest molecules within a system and other second guest molecules are also present, significant uncertainty may arise within these classical models. Lastly, classical models require extensive computational resources to solve iteratively due to their need to couple various different equations-of-states for both phases involved (the vapor phase equation-of-state describing how fugacities evolve with pressure and composition; an equation representing how the Langmuir integral describes hydrate phase behaviour and an activity coefficient model describing how aqueous phase chemistry evolves with inhibitor addition), especially when conducting engineering screening exercises assessing multiple operating conditions quickly.

Thus, classical models motivated developing complementary predictive methods, specifically, machine learning based data-driven predictive models which comprise a central aspect of this thesis.

2.3 Thermodynamic Inhibitors: Categories, Mechanisms and Experimental Results in Literature

2.3.1 Classification of Hydrate Inhibitors

Thermodynamic hydrate inhibitors (THIs) are categorized into three general classifications: THIs, Kinetic Hydrate Inhibitors (KHIs), and Anti-Agglomerants (AAs) (Wang et al., 2019; Kashish et al., 2024). THIs expands the operable conditions of a pipe system by lowering its equilibrium boundaries of hydrate stability toward lower temperatures and higher pressures. KHIs inhibit hydrate crystal nucleation rates without affecting equilibrium conditions; AAs inhibit hydrate particle aggregation. The scope of this thesis focuses solely on THIs since all experimental datasets collected herein include only THI systems.

Various strategies exist to hinder hydrate formation including thermal conditioning, eliminating free-water content within liquids being transported via piping, reducing operational pressure conditions of piping systems transporting fluids prone to hydrate formation and chemically inhibiting hydrate formation processes. Of these options, injecting chemical inhibitors is currently the most popular method to mitigate hydrate formation problems within long distance pipelines carrying fluids that operate at elevated pressures and low temperatures where

economic constraints limit heat-exchange technology usage, insulation thicknesses and dehydration technologies (Wang et al., 2019).

2.3.2 Organic Thermodynamic Inhibitors

Organic THIs primarily used within industrial flow assurance applications include methanol (MeOH), monoethylene glycol (MEG), diethylene glycol (DEG), triethylene glycol (TEG), ethylene glycol (EG). All four aforementioned compounds depress liquid water activities via hydrogen-bonding between their respective hydroxyl (-OH-) functional groups with water molecules thereby suppressing hydrate formation at any fixed pressure condition (Sa et al., 2018; Wang et al., 2019).

Amongst organic THIs evaluated above; methanol is favoured for its relatively low cost and high efficacy at low doses; particularly due to its ability to diffuse quickly into areas prone to hydrate formation due to its high diffusivity. Methanol loses considerable amounts to vaporization within flowing gas streams due to its high vapor pressure, limiting its suitability for continuous injection within pipeline systems, however, methanol remains viable as an inhibitor in short-term applications such as well restarts or emergency hydrate mitigation scenarios. Monoethylene Glycol possesses favourable properties compared to methanol including improved chemical stability, lower toxic effects and lower gas phase solubility; thus, MEG is often selected as an inhibitor for continuous injection applications within long-haul offshore pipelines (Wang et al., 2019). Conversely with respect to methanol, increasing concentrations of MEG result in increasingly non-linear temperature reductions, whereas methanol demonstrates linear temperature depressions as functions of concentration at a specified pressure increase, consequently, having direct implications for inhibitor dose rate calculations (Wang et al., 2019). DEG & TEG possess analogous physical/chemical properties but are chosen depending upon specific requirements of a particular system configuration; especially when recovery/regeneration of injected inhibitor becomes feasible.

Dimethyl Sulfoxide (DMSO) is another lesser-known organic THI candidate that has generated interest as a more effective and environmentally-friendly alternative to EG & MeOH (Semenov et al., 2021; Semenov et al., 2022); Recent experimental studies by Semenov et al. have measured hydrate phase equilibrium data for CH₄ & CO₂ hydrates in DMSO across a broad concentration range (10–55 wt%) and these data sets are included within the present collection.

Therefore, inclusion of DMSO data extends representation of THI-types previously underrepresented in previous machine-learning based hydrate-prediction studies.

2.3.3 Inorganic Salt Inhibitors

Inorganic electrolytes, specifically halide salts are very powerful THIs because electrolytic binding tightens water molecules around individual ions in aqueous solutions via electrostatic forces, therefore, greatly decreasing water activity (Wang et al., 2019). The order of effectiveness for cations in terms of THI potency follows: $Mg^{2+} > Ca^{2+} > Na^+ > K^+ > NH_4^+$; Anions are similarly ordered with Cl^- exhibiting significantly greater inhibitory capacity than SO_4^{2-} because Cl^- exhibits smaller ionic radii & larger charge densities (Wang et al., 2019). Ionic radius plays a greater role than does ionic charge in altering hydrate equilibrium because smaller ionic radii lead to tighter binding of surrounding water molecules about each ion, therefore, leading to greater reduction in activity coefficients associated with water molecules.

Pure salt injection is rarely practiced as a systematic means of deliberately introducing THIs within natural gas pipelines because dissolved salt promotes corrosion. Dissolved salts exist ubiquitously within production wells as part of co-produced waters accompanying hydrocarbon flows, therefore, precise determination of hydrate formation conditions in brine environments is crucial for assessing flow assurance risks regardless of whether or not intentional salt injection practices occur. The datasets prepared for this thesis contains data for NaCl, KCl, $CaCl_2$, $MgCl_2$ & NH_4Cl systems providing thorough representation of commercially relevant salt types.

2.3.4 Mixed Inhibitors Systems

Most of the time in real-world applications of gas production processes, the aqueous phase contains a mixture of two or more of the following types of inhibitors — the intentionally added organic inhibitors along with the naturally occurring dissolved salts in the formation water. Due to the nonlinear interactions between the various inhibitor species, the description of thermodynamics of mixed-inhibitor systems is much more complicated than it is for single-inhibitor systems. Wang et al. (2019) demonstrated using research that, in general, the total inhibiting action of mixed electrolyte and organic inhibitors is not the sum of the actions of the individual inhibitors, and the nature of the synergy/antagonism between the two will depend upon the specifics of the inhibitor types and their concentrations. The dataset collected for this thesis contains numerous examples of systems containing mixtures of inhibitors. The examples

include MeOH + NaCl (Jager et al., 2002; Lafond et al., 2014; Fu et al., 2024), KCl + NaCl + MeOH (Fu et al., 2024), MgCl₂ + MeOH (Semenov et al., 2024), NaCl + CaCl₂ + MeOH + MEG (Qazi Nasir et al., 2014) and NH₄Cl + MeOH + MEG (Hun Park et al., 2019).

2.3.5 Experimental Data Sources for This Thesis

The total amount of data contained within final dataset consists of 922 data points that were produced using the results from 30 experimental studies. The dataset was pre-processed and screened for thermodynamic consistency prior to use for training machine learning models. All of the data sources used in this study were published during a period spanning from 1987 until 2026. Examples of gas compositions, inhibitor types and operating conditions can be found throughout all of the studies. Table 2.1 lists all data sources, along with their major characteristics.

Table 1. Summary of experimental data sources for the compiled dataset

Source	Year	Gas System	Inhibitor(s)	Conc. (wt%)	P (MPa)	T (K)	N
Ng, Chen & Saeterstad	1987	Natural gas	MeOH, MEG	13–24	3.91–20.0	274.35–288.15	19
Song & Kobayashi	1989	CH ₄ + C ₃ H ₈	MeOH, EG	5–35	0.433–17.98	264.98–291.4	10
Fan & Guo	1999	CO ₂ +CH ₄ +C ₂ H ₆	NaCl	9.41–10	1.10–12.87	266.9–283.6	37
Fan & Guo	1999	CO ₂ +CH ₄ +N ₂	NaCl	9.45	1.12–3.16	266.9–276.2	13
Fan & Guo	2000	CH ₄ +CO ₂	EG	10	1.14–3.22	268.7–278.0	4
Dholabhai & Bishnoi	1994	CH ₄ + CO ₂	NaCl, KCl, CaCl ₂	4–20	1.38–10.61	264.14–281.57	31
Dholabhai, Parent & Bishnoi	1997	CH ₄ + CO ₂	MeOH, EG	10–20	1.46–10.35	262.24–281.51	24
Bishnoi & Dholabhai	1999	CH ₄ +C ₃ H ₈ +CO ₂	MeOH	5–20	2.52–10.09	273.81–288.03	33
Tohidi et al.	1997	CO ₂ , C ₂ H ₆ , C ₃ H ₈	MeOH	0.5–20	0.23–3.613	263.2–280.4	37
Tohidi et al.	2001	Natural gas	MCH	21.9–49.3	2.103–13.789	278.75–290.75	21
Jager & Sloan	2001	CH ₄	NaCl	2–8	4.0–71.56	270.66–290.06	39
Jager et al.	2002	CH ₄	MeOH + NaCl	2–40	2.05–73.28	249.9–290.9	57
Maekawa	1998	CH ₄ +C ₂ H ₆	NaCl	3	3.8–9.5	276.4–287.3	21
Maekawa	2001	CH ₄ +C ₂ H ₆	NaCl	3	3.7–9.5	276.4–288.1	29

Source	Year	Gas System	Inhibitor(s)	Conc. (wt%)	P (MPa)	T (K)	N
Jianfen Du	2019	CH ₄	NaCl, CaCl ₂ , MgCl ₂	1.5–14.52	2.71–15.61	271.35–287.75	36
Afzal et al.	2008	CH ₄ , C ₂ H ₆ , C ₃ H ₈ , CO ₂	DEG	6.6–16.4	0.48–8.02	272.1–285.7	32
Mohammadi & Richon	2010	CH ₄	MeOH, EG	4.2–50	2.46–10.79	254.6–285.1	38
Semenov et al.	2021	CH ₄	DMSO	10–55	2.9–12.99	243.03–286.28	30
Semenov et al.	2022	CH ₄	MeOH, MEG, DEG	10–30	4.95–8.24	269.87–282.14	35
Semenov et al.	2022	CO ₂	DMSO	10–40	1.02–4.0	254.92–280.35	21
Semenov et al.	2024	CH ₄	MgCl ₂ + MeOH	5–26.68	2.99–13.06	248.98–286.01	85
Zahra Rizi et al.	2026	Natural gas	MeOH, MEG	10–30	6.08–12.26	274.25–291.35	24
Anne Tamaskovics et al.	2026	CO ₂	MeOH, DEG, MEG	2.9–19.5	1.5–8.0	263.0–284.0	46
Lafond et al.	2014	CH ₄	MeOH + NaCl	2.8–20	7.0–20.0	271.99–285.99	8
Liang & Zhao	2024	Natural gas	NaCl + EG	10	1.95–4.11	279.35–285.35	9
Qazi Nasir et al.	2014	Natural gas	NaCl+CaCl ₂ +MeOH+MEG	10–20	2.97–7.44	268.35–281.85	18
Hun Park et al.	2019	CH ₄	NH ₄ Cl+MeOH+MEG	5–20	4.36–11.64	264.71–281.13	32
Fu et al.	2024	CH ₄	KCl+NaCl+MeOH	3.55–60	2.95–13.18	234.03–286.01	52
Yan Li et al.	2021	CO ₂	NaCl,CaCl ₂ ,MeOH,PrOH,EG,MEG	3.6–20	0.58–4.05	258.15–279.65	82

This dataset contains a wide variety of gas compositions and inhibitor combinations, including both single-component and multi-component gases. The gases included in this dataset are CH₄, C₂H₆, C₃H₈, C₄-hydrocarbons, N₂, CO₂. The dataset covers both single-stage inhibition (sI) and two-stage inhibition (sII). There are 11 types of inhibitors represented in this dataset: NaCl, KCl, MgCl₂, CaCl₂, NH₄Cl, MeOH, PrOH, EG, MEG, DEG and DMSO. The pressure range of all data sources is 0.23–73.28 MPa, and the temperature range is 234.03–299.35 K. The vast majority of gas processing systems operate under pressures and temperatures that fall outside of this range. Therefore, this dataset represents one of the primary advantages of the predictive framework described in this dissertation over many previously proposed ML-based hydrate predictions.

2.4 Thermodynamic Consistency of Hydrate Phase Equilibrium Data

2.4.1. *The Need for Consistency Assessment*

As stated earlier in this chapter, another significant issue that needs to be addressed before developing any predictive model for gas hydrate phase equilibria is the quality and reliability of the experimental data that is used to develop the model. Experimental measurements of gas hydrate phase equilibrium contain inherent uncertainty associated with the precision and accuracy of pressure and temperature measurements, equilibration times, sample preparation methods and protocols, mixing efficiencies and differences in experimental equipment/apparatus and methodologies used by different researchers (Sa et al., 2018). Although there is a large collection of gas hydrate phase equilibrium data available in scientific publications to date, no formal methodology has been presented to evaluate the thermodynamic consistency of reported data, making it very important to do so (Sa et al., 2018).

Because machine learning algorithms do not incorporate any physical constraints into their representation of relationships between variables like thermodynamic models do, if a machine learning model is trained using thermodynamically inconsistent data, then the inconsistencies will be incorporated into the learned feature-output mapping and manifest as physically unreasonable prediction behaviour once the model is trained. As such, the quality of the data used to train the model determines its reliability and it is essential to accurately evaluate all reported phase equilibrium data to improve existing thermodynamic models and improve predictions of phase behavior [Sa et al., 2018]. An example illustrating why evaluating data quality is important can be seen in the results from Rasoolzadeh et al. (2025), who found that removing 14 inconsistent data points from a database consisting of 136 records reduced the average absolute deviation in temperature for C₂H₆ hydrates from 0.61 K to 0.20 K, a reduction in deviations of nearly a factor of three due solely to screening data quality.

2.4.2 *The Sa et al. (2018) Consistency Methodology*

Sa, Hu, and Sum (2018) describe a rigorous qualitative three-criteria framework for determining the thermodynamic consistency of reported inhibited hydrate phase equilibrium data. Unlike other frameworks or methodologies, this approach uses fundamental thermodynamic equations to evaluate data consistency, requires no adjustment parameters and evaluates each dataset individually (each combination of gas composition, inhibitor type/concentration).

Criteria 1 - Linearity according to the Clausius-Clapeyron relation: Criterion 1 assesses whether or not reported hydrate phase equilibrium data conform to the Clausius-Clapeyron equation, which describes a linear relationship between $\ln(P)$ and $1/T$ for a small temperature interval (approximately 10–20K), assuming that neither changes in the hydrate dissociation enthalpy ΔH_d nor the gas compressibility factor Z occur within that temperature range:

$$\frac{d \ln(P)}{d(1/T)} = - \frac{\Delta H_d}{zR}$$

To determine whether or not each dataset conforms to a linear relationship, linear regressions are performed between $\ln(P)$ vs. $1/T$ for each dataset. The quality of fit between observed values and calculated values is evaluated using the coefficient of determination R^2 for each dataset. Based on R^2 values, each dataset is classified as follows: pass ($1-R^2 < 2.5\%$), acceptable ($2.5\%-5\%$) or fail ($>5\%$). This criterion will identify datasets whose experimental scatter or systematic errors result in large deviations from expected linear behaviour (Sa et al., 2018).

Criteria 2 - Consistency of Dissociation Enthalpies: Criteria 2 examines whether or not addition of a thermally induced hydrate inhibitor (THI) alters the hydrate dissociation enthalpy ΔH_d . Because THIs interact with water in solution but do not affect crystalline hydrate structures, ΔH_d should remain constant for a given hydrate former at specified conditions regardless of THI type or concentration. For each inhibited dataset, comparison is made between the slope $A = -\Delta H_d/(Z \cdot R)$, determined from linear regression analysis for that dataset's $\ln(P)$ -vs.- $1/T$ plot and slope A_{vv} from a linear regression analysis using a reference dataset describing an uninhibited system having identical hydrate formers and pressure ranges: Relative difference between slopes: $[(A_{tir}-A_{vv})/A_{vv}]$. Classification criteria: pass ($< \pm 5\%$), acceptable ($\pm 5-10\%$), or fail ($> \pm 10\%$) [Sa et al., 2018]. In this thesis, reference slope A_{vv} for single component hydrate formers is obtained from Rasoolzadeh et al. (2025), whereas vdW-P model serves as a reference for multicomponent systems.

Criteria 3 - Internal Consistency Regarding Water Activity: Criteria 3 assesses whether or not temperature depressions relative to uninhibited systems vary systematically with THI concentrations at constant pressure. Specifically, normalized temperature depressions $\Delta T/(T_o \cdot T)$ — where T_o is temperature at which an uninhibited system reaches equilibrium with gas at specified pressure T is temperature at which inhibited system reaches equilibrium with gas at specified pressure — should be directly proportional to $\ln(a_{vv})$, which decreases monotonically with increasing THI concentration. Criteria 3 thereby enables identification of

datasets whose reported temperature depressions are physically inconsistent with expected trends as THI concentrations increase (Sa et al., 2018).

Collectively, Criteria 1–3 provide a methodology combining a quantitative quality rating for each dataset examined independently with a scientifically valid rationale for retaining/rejecting datasets to be included in a modelling database. This methodology is subsequently applied to all thirty source datasets.

2.5 Machine Learning Methods Applied to Hydrate Phase Equilibrium Prediction

2.5.1 Overview of Machine Learning in Hydrate Research

Machine learning techniques for predicting gas hydrate phase equilibria have evolved greatly over the last decade. Machine learning concepts have been developed extensively over recent decades and have generated substantial bodies of work on machine learning in industrial production, cheminformatics and chemical engineering. As far as hydrate phase equilibria are concerned, these approaches have been divided into three categories: tree-based algorithms, neural networks and support vector machines (Cai et al., 2025). While high R^2 values (typically >0.97) have been reported in multiple studies cited above; there remains a common limitation among all previous work, i.e., none have implemented thermodynamic screening on training data sets or have limited themselves to studying either single-component gas systems or restricted ranges of inhibitor types.

In order to illustrate these limitations, six ML-algorithms used in this work - SVR, ANN, XGBoost, CatBoost, GBR, RF & DT - are discussed below.

2.5.2 Decision Trees

Decision Trees are a non-parametric supervised learning technique that divides feature-space into subspaces via recursive partitioning decisions to minimize a splitting-criterion — e.g., Mean Squared Error (MSE) for regression problems (Bentéjac et al., 2019). Once a tree structure is constructed (via recursive partitioning), it becomes possible to map inputs through splits represented by internal nodes; resulting branch-outputs; and terminal leaf-node predictions for outputs (Bentéjac et al., 2019). Decision Trees are transparent; require little preprocessing of data; support both continuous and categorical features; however, are prone to

overfitting as trees grow fully deep; consequently decision-trees have larger variances for prediction than ensemble-methods (González et al., 2020). In terms of hydration-prediction; Decision-Trees serve as base learners for all ensemble-methods discussed below as well as a standalone baseline-model.

2.5.3 Random Forest

Random Forest is an ensemble method derived from bagging/bootstrap-aggregation methods originally developed by Breiman (González et al., 2020). At each node split during tree construction, Random Forest samples from bootstrapped training data set with replacement; and selects randomly from m candidate features at each node split; where m typically equals $\sqrt{n}$ for n classification problems or $(1/3)n$ for n regression problems (Bentéjac et al., 2019; González et al., 2020). Final prediction is obtained by averaging predictions from all trees comprising the forest. The combination of bootstrap-sampling and random feature selection leads to greater diversity amongst trees, thus, reducing variance between trees in comparison to a single decision-tree (Bentéjac et al., 2019). Random Forest has consistently been ranked among top performers across large-scale comparative studies across disparate datasets (González et al., 2020).

2.5.4 Gradient Boosting Regression and XGBoost

Gradient boosting is an ensemble technique that involves the addition of decision trees one by one, training each to minimise the loss based on the negative gradient or pseudo-residual of the loss function in terms of the existing ensemble predictions (González et al., 2020; Bentéjac et al., 2019). In this regard, at every iteration m , a new model h_m is learned in order to approximate the residuals $r_i = -[\partial L(y_i, F(x))/\partial F(x_i)]$, and then the ensemble is updated via $F_m(x) = F_{m-1}(x) + v \cdot \rho_m \cdot h_m(x)$, where v is the learning rate and is responsible for the shrinkage of each update [Bentéjac et al., 2019]. Through this iterative process, the loss is gradually decreased, leading to extremely low-bias models, but proper regularisation is needed to avoid overfitting.

XGBoost, eXtreme Gradient Boosting, is a fast and well-regularised version of gradient boosting proposed by Chen and Guestrin (González et al., 2020). This technique includes additional regularisation terms to the gradient boosting objective function, namely, $\Omega(f) = \gamma T + \frac{1}{2} \lambda \|w\|^2$, which accounts for model complexity using the number of leaf nodes T and a leaf score vector w (González et al., 2020). Besides, the introduction of the second-order Taylor expansion of the loss function results in faster and easier optimisation as well as allows employing various

custom loss functions. XGBoost employs approximate techniques for splitting, including weighted quantile sketch, sparse-friendly functions for finding good splits, and cache-aware parallelisation of tree building processes, thus significantly outperforming gradient boosting when processing big data (González et al., 2020). It is one of the best-performing machine learning methods used in various competitions and has even proven capable of predicting hydrate equilibrium with $R^2 = 0.9956$ (Cai et al., 2025).

2.5.5 CatBoost

CatBoost (Categorical Boosting) is a gradient boosting algorithm that was developed by Yandex, and which can deal directly with categorical features and avoid their one-hot encoding as a pre-processing step; it solves the issue of prediction shift that arises with other implementations of gradient boosting (González et al., 2020). CatBoost uses a form of ordered boosting based on permutations of the training dataset where the residual for each element of the dataset is determined from a model estimated based on samples appearing earlier in the permutation, thus not suffering from bias caused by using the same data points for model fitting and residuals calculation. CatBoost makes use of symmetric (or oblivious) trees as base learners, which results in well-balanced models with a lower tendency to overfit, in addition to faster predictions (González et al., 2020). In the case of hydrate prediction, CatBoost benefits particularly from its ability to deal with categorical features like the type of inhibitor or gas composition labels, since these are often represented as categorical features, not numeric.

2.5.6 Support Vector Regression

Support Vector Regression (SVR) is an extension of Support Vector Machines to solve regression problems (Smola and Schölkopf, 2004). Specifically, ε -SVR searches for a function $f(x) = \langle w, x \rangle + b$ such that the deviation of this function from targets y_i does not exceed ε for all training samples, while at the same time the function is as flat as possible, i.e., minimises the value of $\|w\|^2$. Samples lying outside the ε -insensitive tube provide the error term, which is linearly penalised in ε -SVR. The parameter C allows us to find a balance between a maximally flat surface and tolerance of deviations exceeding ε (Smola and Schölkopf, 2004). As we have seen when solving the dual of the SVR optimisation problem, the final result will depend solely on dot products of pairs of training samples; using the kernel trick, we can use an arbitrary positive definite function instead of dot product, and most often it would be the RBF kernel function $K(x_i, x_j) = \exp(-\gamma\|x_i - x_j\|^2)$, used for non-linear regression problems.

2.5.7 Artificial Neural Networks

ANNs refer to computational models that consist of interconnected layers of artificial neurons where each neuron performs a weighted linear summation of its inputs followed by a non-linear activation function (Schmidhuber, 2015). In a traditional feedforward multilayer perceptron (MLP), data flows through the layers from the input layer to the hidden layers and then to the output layer. The network parameters are estimated using the backpropagation algorithm which estimates the partial derivative of the loss function for each parameter by the chain rule and adjusts the parameters using gradient descent algorithms and their adaptive forms (Schmidhuber, 2015). ANNs are universal function approximators that can approximate any complex non-linear input-output relationship if provided enough capacity (width and depth). ANNs have been successfully used to predict hydrate equilibrium states with high correlation coefficients ($R^2 > 0.985$) (Cai et al., 2025).

2.5.8 Model Evaluation Framework

Evaluating a model for machine learning algorithms performing regression tasks involves choosing a range of metrics describing the characteristics of the model (Varoquaux and Colliot, 2023). For the purpose of this thesis, the list of evaluation metrics is as follows:

The **coefficient of determination (R^2)** is the metric quantifying the proportion of variability of the target variable explained by the model in relation to its total variability, varying from 0 to 1 (the former corresponds to no explanatory power, while the latter means perfect model fit). The **Root Mean Squared Error (RMSE)** represents the average prediction error expressed in units of the target variable, namely kelvins (K), and has an increased sensitivity to larger errors compared to smaller ones. The **Mean Absolute Percentage Error (MAPE)** is an absolute measure of predictive accuracy as a percentage value and independent of the scale of the target variable, thus allowing comparing accuracy levels between various sets with differing temperatures (Varoquaux and Colliot, 2023). Lastly, the **Akaike Information Criterion (AIC)** and the **Bayesian Information Criterion (BIC)** are two related measures based on information theory penalizing complexity of the model; the lower values of each correspond to a better trade-off between the two components (Varoquaux and Colliot, 2023). Comparing AIC/BIC metrics in training and testing datasets allows diagnosing the problem of overfitting: a large difference between AIC/BIC scores implies overfitting and poor generalization capacity.

2.6 Research Gaps and Positioning of This Thesis

From the review of the literature, two main research gaps can be identified, which are directly addressed by the present thesis:

Gap 1 – No thermodynamic consistency screening applied in previous ML studies. As highlighted in the literature review of ML-based hydrate prediction studies, it is evident that none of the previous studies mentioned in Table 4 applied any sort of thermodynamic consistency check to the training data before building models on the basis of the datasets compiled from the literature. Models built on uncleaned literature data sets run the risk of learning thermodynamically inconsistent data points into their mappings, resulting in potential loss of physical meaning of the results produced by such models as well as making the model less generalisable. Gap 1 is being addressed in the present thesis by using the complete three-criterion methodology of Sa et al. (2018) as a data quality measure prior to building ML models, which reduces the initial compilation of 1,751 data points to 1,167 thermodynamically consistent data points and then further to a cleaned data set of 922 data points.

Gap 2 – No assessment of physical trend validation of ML models' predictions. Previously published ML studies on hydrate equilibrium prediction generally focus on assessing the performance of the developed models based only on statistical performance metrics such as R^2 , RMSE calculated based on test data. Very few papers have been found in which researchers have tested whether the developed models actually learn the correct physics of the phenomena (such as monotonic decrease of the formation temperature with an increase of inhibitor concentration under constant pressure, correct order of cation strength in salt, and multicomponent effects of gases).

3. METHODOLOGY

3.1 Overview of the Methodological Framework

The research methodology of the present thesis comprises of the following four steps that are connected to each other in sequence: (1) comprehensive acquisition and collation of available experimental data regarding gas hydrate phase equilibrium from relevant scientific literature; (2) thermodynamically consistent screening of all data gathered through the use of Sa et al.'s (2018) three criteria approach, with equilibrium temperature values for uninhibited systems computed via Rasoolzadeh et al.'s (2025) formula and the van der Waals–Platteeuw equation; (3) exploratory data analysis, processing, and feature engineering; and (4) creation, tuning, training, testing, and comparison of six different machine learning regression models. After selecting the optimal model, the same model will then be used for trend identification and physical validation, as well as supervised interpolation on training domain data. The complete methodological pipeline is presented schematically in Figure 6.

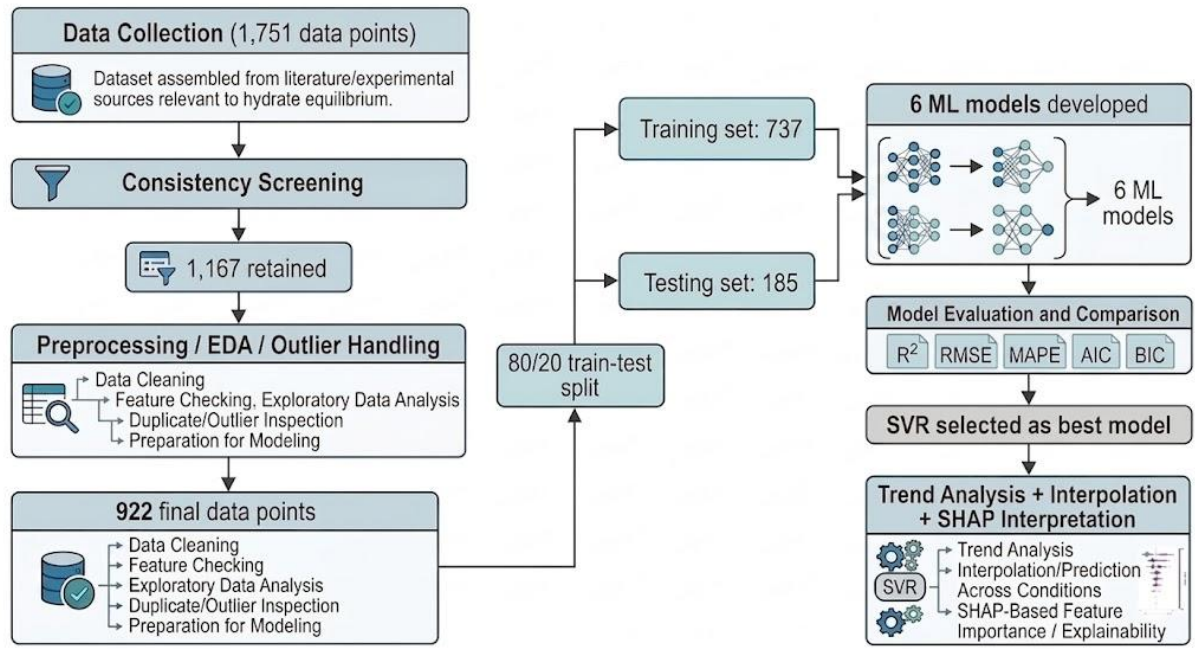


Figure 6. Full methodological pipeline flowchart

All computations were done in Python employing the Anaconda package and Spyder development environment. The main libraries used in computations included: NumPy and pandas to deal with data and perform computations; scikit-learn for modelling (SVR, RF, GBR, DT using MLPRegressor), pre-processing (StandardScaler, Pipeline), validation (RepeatedKfold), parameter tuning (RandomizedSearchCV, GridSearchCV), evaluation of

results; xgboost (with XGBRegressor library from xgboost.sklearn) for XGBoost; catboost for CatBoost; matplotlib and seaborn for visualization purposes; scipy for statistics; joblib for serializing models and shap for feature importance. The calculations of thermodynamically consistent properties of multicomponent gas mixtures were carried out using Python scripts. Single component inhibited equilibrium temperatures were calculated in Excel spreadsheet applying the formulas developed by Rasoolzadeh et al. (2025). The same seed (42) was used in all random computations to make all the calculations fully reproducible.

3.2 Experimental Data Collection

3.2.1 Data Sources and Collection Procedure

A comprehensive database of experimental phase equilibrium data for gas hydrates with thermodynamic hydrate inhibitors was meticulously gathered from 30 research articles published between 1987 and 2026. All data were manually obtained from the tabulated numerical form in the original documents without the use of any software for digitising figures at any point. Only data presented in the numerical form in the original documents were used, thereby ensuring that there would be no transcription errors caused by uncertainties associated with figure interpolations.

Thirty original research papers with six gas compounds (methane, ethane, propane, butanes, nitrogen, carbon dioxide) in single and multi-component forms, eleven inhibitor categories (NaCl, KCl, MgCl₂, CaCl₂, NH₄Cl, MeOH, propanol, EG, MEG, DEG, DMSO) and some combinations of mixed inhibitors were considered. The entire list of sources along with relevant information about the final cleaned datasets is given in Table 1. The distribution of the final dataset by inhibitor type and by gas composition is given in Figure 7 and 8. Prior to any quality assessment, the raw dataset comprised a total of 1,751 data points and had been compiled from fifty published articles.

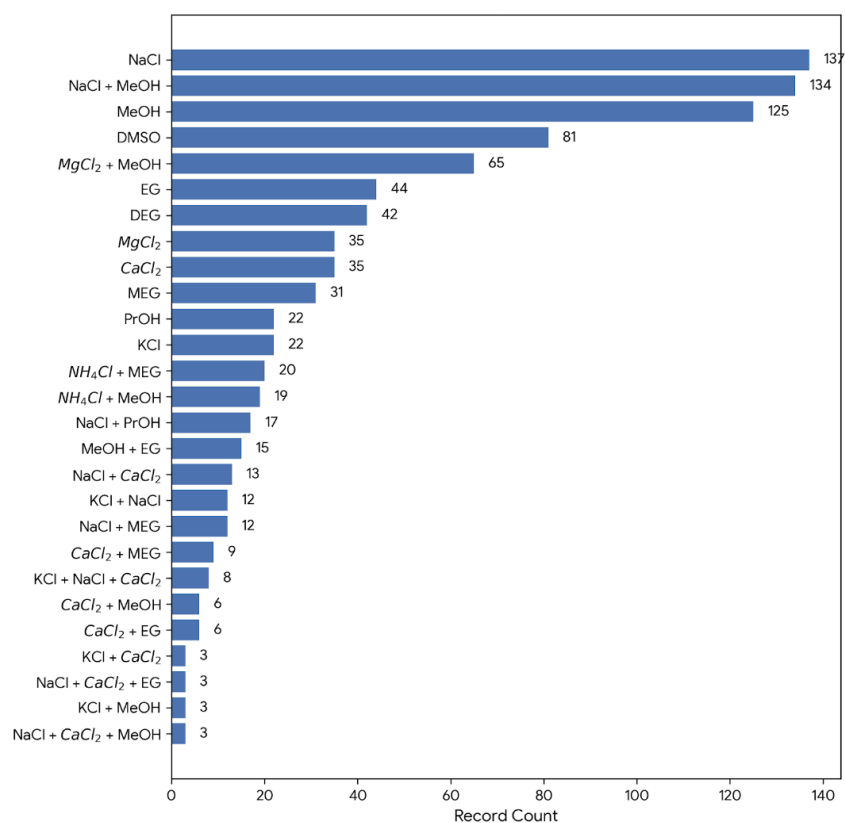


Figure 7. Distribution of the dataset by inhibitor type

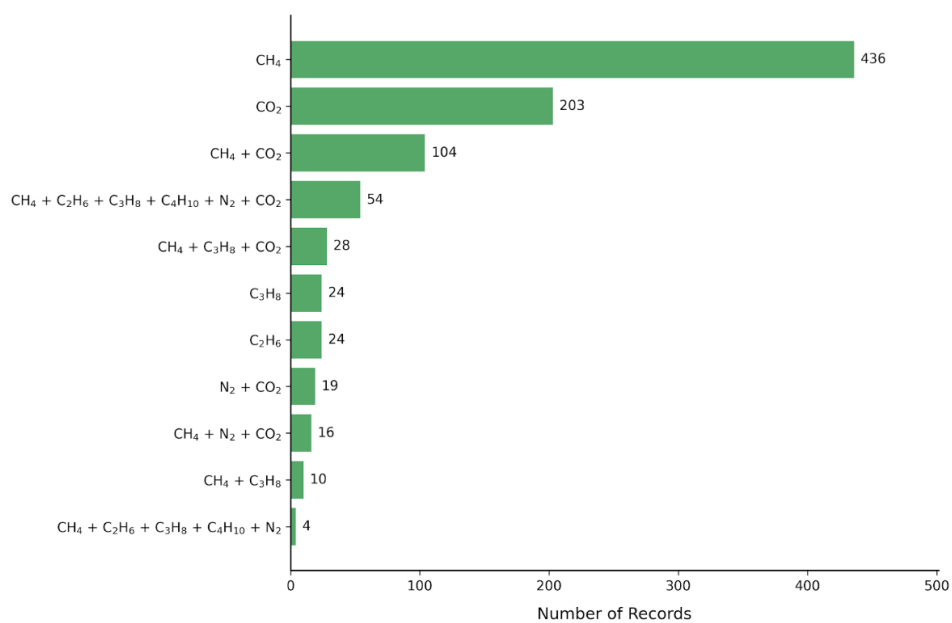


Figure 8. Distribution of the dataset by gas composition

3.3 Thermodynamic Consistency Screening

3.3.1 Rationale

Prior to any preprocessing or modelling, all 1,751 experimental data points were evaluated using a thermodynamic consistency screening procedure. This procedure was deemed necessary owing to the reason that whereas a substantial body of experimental hydrate phase equilibrium data has been documented in the literature, no formal technique for evaluating thermodynamic consistency of such data has yet been proposed, which is much needed (Sa et al., 2018). Any statistical modelling algorithm that utilizes thermodynamically inconsistent data may learn to embed errors from the experiment into its mappings, resulting in highly accurate yet physically inaccurate models. To the best of my knowledge, the current thesis represents the first effort in the machine-learning based hydrate phase prediction literature to conduct rigorous thermodynamic consistency testing on experimental data.

3.3.2 Calculation of Uninhibited Reference Equilibrium Temperatures

Single-Component Systems:

The three consistency tests described by Sa et al. (2018) involve comparing inhibited phase equilibrium data to the corresponding hydrate equilibrium curve for the reference case of an uninhibited equilibrium for the gas mixture at the relevant pressures. Two approaches were applied to estimate the reference temperatures for the uninhibited case based on the type of system considered.

For datasets including a single gas forming hydrates (CH_4 , C_2H_6 , C_3H_8 , CO_2 , or N_2), the temperature of the uninhibited hydrate dissociation T for a given pressure P was estimated from the analytical solution proposed by Rasoolzadeh et al. (2025). The basis for this approach is the dependence of the enthalpy of hydrate dissociation ΔH_d on the temperature, which is formulated as follows:

$$\Delta H_d = K_1 + \frac{K_2}{T}$$

Combining this with the Clausius-Clapeyron equation for hydrate phase equilibrium:

$$\frac{d \ln P}{d(1/T)} = - \frac{\Delta H_d}{zR}$$

and integrating with separation of variables, Rasoolzadeh et al. (2025) derived the following closed-form equation for hydrate dissociation pressure as a function of temperature:

$$P = \exp \left(\frac{A}{T} + \frac{B}{2T^2} + C \right)$$

where A (K), B (K²), and C (dimensionless) are empirically fitted constants determined for each hydrate former by regression against a large consistent experimental database. Rearranging this equation algebraically yields a quadratic expression in T, solved explicitly to give the hydrate dissociation temperature as a function of pressure:

$$T = \frac{2A \mp \sqrt{4A^2 + 8B(\ln P - C)}}{4(\ln P - C)}$$

These calculations were performed in Microsoft Excel for each data point in the single-component inhibited datasets. The fitted constants A, B, and C used for each hydrate former are reproduced in Table 2, taken directly from Rasoolzadeh et al. (2025). The model achieves an average absolute deviation of 0.25 K across 1,879 consistent experimental data points for seven pure hydrate formers at pressures from 0.093 to 493 MPa, making it a highly accurate and reliable reference. It is noted that C₄ hydrocarbons present in some multicomponent datasets were not treated by this single-component equation but were instead handled within the multicomponent vdW-P framework described in the following subsection.

Table 2. Fitting parameters from Rasoolzadeh et al. (2025) used for single-component uninhibited equilibrium temperature calculations

Hydrate Former	A (K)	B (K ²)	C (dimensionless)
CH ₄	-35,267.1988	7,563,853.7909	79.3648
C ₂ H ₆	-83,024.2542	20,385,878.1017	166.6007
C ₃ H ₈	-137,298.4641	33,158,390.9325	278.6568
CO ₂	-91,168.9851	22,628,075.9882	182.3344
N ₂	-8,056.1434	89,155.4526	31.6824

Note: C₄H₁₀ hydrocarbons present in some multicomponent datasets were handled within the multicomponent vdW-P framework described below.

Multicomponent hydrate formers:

For problems related to the composition of natural gases and hydrate-forming gases for binary and ternary combinations, the calculation of the uninhibited equilibrium temperature was done by iteratively solving the van der Waals-Platteeuw (vdW-P) solid solution model equation (Rasoolzadeh et al., 2025), which is the governing equation for hydrate phase equilibrium. The

van der Waals-Platteeuw model was initially formulated by van der Waals and Platteeuw (1959), followed by various other authors such as Parrish and Prausnitz (1972), Sloan and Koh (2008), Medeiros et al. (2020), among others, who applied the principles of statistical thermodynamics of clathrate hydrates and set the chemical potentials of the water in the empty hydrate lattice equal to that of the co-existing aqueous solution. The equilibrium condition according to the formulation in this thesis work is (Rasoolzadeh et al., 2025):

$$\sum_{h=1}^{N_{cav}} v_h \ln \left(1 + \sum_{j=1}^{N_C} C_{hj}(T) f_j \right) = \frac{\Delta \mu_w^0}{RT_0} - \int_{T_0}^T \frac{\Delta h_w^{\beta-L}(T)}{RT^2} dT + \int_0^P \frac{\Delta v_w^{\beta-L}}{RT} dP - \ln(a_w)$$

For the uninhibited reference system computed in this thesis, the water activity term is set to $a_w = 1$, which is thermodynamically exact since the reference aqueous phase consists of pure liquid water with no inhibitor present. The equation is solved iteratively for T at each given pressure P . The complete implementation is described step by step below.

Step 1 - Thermodynamic reference properties and cage specifications

Constants on the right-hand side of the vdWP equation, namely, $\Delta \mu_w^0$, Δh_w^0 , and $\Delta v_w^{\beta-L}$, are related to the difference in water chemical potential, molar enthalpy, and molar volume, respectively, between an ideal hydrate framework with no hydrate (β -phase) and the liquid water phase under reference conditions ($T_0 = 273.15$ K, $P_0 = 0$). The thermodynamic properties, along with the details of cages for hydrate structures I and II, such as cavity type, cavity fraction per water molecule v_h , cavity radius R' , and coordination number z , are summarized in Table 3 below.

Table 3. Cages specifications (Rasoolzadeh et al. 2025)

Structure	I (Small)	I (Large)	II (Small)	II (Large)
Cavity description	5 ¹²	5 ¹² 6 ²	5 ¹²	5 ¹² 6 ⁴
Cavity	Small	Large	Small	Large
v_h	2/46	6/46	16/136	8/136
Average cavity radius (Å)	3.95	4.33	3.91	4.73
Coordination number	20	24	20	28
$\Delta \mu_w^0$ (J·mol ⁻¹)	1263.6	1263.6	882.8	882.8
Δh_w^0 (J·mol ⁻¹)	-4858.9	-4858.9	-5202.2	-5202.2
$\Delta v_w^{\beta-L}$ (cm ³ ·mol ⁻¹)	4.6	4.6	5.0	5.0

Step 2 - Enthalpy difference as a function of temperature

The temperature-dependent enthalpy difference $\Delta h_{vv}^{(\beta-L)}(T)$ between the empty hydrate lattice and the liquid water phase is computed using Equation 12 of Rasoolzadeh et al. (2025), based on the expression originally from Sloan and Koh (2008):

$$\Delta h_w^{\beta-L}(T) = \Delta h_w^0 + \int_{T_0}^T [-38.12 + 0.141(T - T_0)] dT$$

where Δh_{vv}^0 is the reference enthalpy difference at $T_0 = 273.15$ K (taken from Table 1), and the heat capacity difference between the empty hydrate lattice and liquid water is expressed as a linear function of temperature through the empirical constants $-38.12 \text{ J}\cdot\text{mol}^{-1}\cdot\text{K}^{-1}$ and $0.141 \text{ J}\cdot\text{mol}^{-1}\cdot\text{K}^{-2}$ (Rasoolzadeh et al., 2025; Sloan and Koh, 2008).

Step 3 - Fugacity calculation using the Peng-Robinson equation of state

The fugacity f_j of each hydrate-forming gas component j in the vapour phase was calculated using the Peng-Robinson (PR) equation of state (Peng and Robinson, 1976):

$$P = \frac{RT}{V_m - b} - \frac{a(T)}{V_m(V_m + b) + b(V_m - b)}$$

where the temperature-dependent energy parameter $a(T)$ and the co-volume parameter b are computed from:

$$a(T) = 0.45724 \frac{R^2 T_c^2}{P_c} \alpha(T), b = 0.07780 \frac{RT_c}{P_c}$$

$$\alpha(T) = \left[1 + \kappa \left(1 - \sqrt{T/T_c} \right) \right]^2, \kappa = 0.37464 + 1.54226\omega - 0.26992\omega^2$$

Fugacity coefficient ϕ_j of each individual gas j – and consequently its fugacity $f_j = \phi_j \cdot y_j \cdot P$ – was determined from the PR EOS on the basis of the relation between the fugacity and the residual Gibbs free energy. The values of critical temperature T_c , critical pressure P_c , and acentric factor ω needed for calculation purposes for each gas are tabulated in Table 4 together with other gases such as CH_4 , C_2H_6 , C_3H_8 , CO_2 , and N_2 .

Table 4. The acentric factors and the critical properties of the hydrate formers (Rasoolzadeh et al. 2025)

Hydrate Former	T_c (K)	P_c (MPa)	ω
CH_4	190.56	4.599	0.01155
C_2H_6	305.32	4.872	0.09949

Hydrate Former	Tc (K)	Pc (MPa)	ω
C ₃ H ₈	369.83	4.248	0.15229
CO ₂	304.21	7.383	0.22362
N ₂	126.20	3.400	0.03772
H ₂ S	373.53	8.963	0.09417
O ₂	154.58	5.043	0.02218

Step 4 - Langmuir constant computation via the Kihara potential

The Langmuir constant $C_{hj}(T)$ quantifies the interaction between hydrate-forming molecule j and cavity type h of the hydrate lattice, and is computed from the spherical cell integral (Rasoolzadeh et al., 2025; originally from Parrish and Prausnitz, 1972):

$$C_{hj}(T) = \frac{4\pi}{kT} \int_0^{R'-a} \exp\left(-\frac{w(r)}{kT}\right) r^2 dr$$

where k is the Boltzmann constant ($k = 1.380649 \times 10^{-23} \text{ J}\cdot\text{K}^{-1}$), R' is the mean cavity radius of the h -type cavity (Å), a is the radius of the spherical core of the Kihara potential of gas j (Å), r is the radial distance between the centres of the guest molecule and the cavity, and $w(r)$ is the spherically averaged cell potential.

The cell potential $w(r)$ is obtained by integrating over the spherical shell the interaction potentials of all the molecules of the host water within one cavity, as proposed by McKoy and Sinanoglu (1963) and formulated by Parrish and Prausnitz (1972) and Sloan and Koh (2008):

$$w(r) = 2z\varepsilon \left[\frac{\sigma^{*12}}{R'^{11}r} \left(\delta^{10} + \frac{a}{R'} \delta^{11} \right) - \frac{\sigma^{*6}}{R'^5r} \left(\delta^4 + \frac{a}{R'} \delta^5 \right) \right]$$

where the function δ^N is defined as:

$$\delta^N = \frac{1}{N} \left[\left(1 - \frac{r}{R'} - \frac{a}{R'} \right)^{-N} - \left(1 + \frac{r}{R'} - \frac{a}{R'} \right)^{-N} \right]$$

Here, z represents the coordination number of the cavity (number of water molecules in cage wall), ε represents the maximum attractive potential energy (J) for gas-water attraction, σ^* stands for the collision diameter at zero potential (Å), and a is the hard-core radius (Å). All

three parameters (a , σ^* , and ϵ/k) make up the Kihara potential parameters of gas-water interactions, presented in the Table 5 below. For calculating the integral in Langmuir constant equation, numerical integration in Python was used (via `scipy.integrate.quad`), with integration boundaries from $r = 0$ to $r = R' - a$ (values beyond $r = R' - a$ do not have physical meaning due to overlapping of cores). The values of the Langmuir constants were calculated independently for each type of cavity (small 5^{12} and large $5^{12}6^2$ for structure I; small 5^{12} and large $5^{12}6^4$ for structure II) and each gas j for each iteration temperature T .

Table 5. Kihara potential parameters for various hydrate formers (Rasoolzadeh et al. 2025)

Hydrate Former	a (Å)	σ^* (Å)	ϵ/k (K)	a (Å)	σ^* (Å)	ϵ/k (K)
CH ₄	0.3834	3.14393	155.593	0.3216	3.1275	161.614
C ₂ H ₆	0.5651	3.24693	188.181	0.6018	3.2088	176.028
C ₃ H ₈	0.6502	3.41670	192.855	0.6828	3.4771	185.399
CO ₂	0.6805	2.97638	175.405	0.6318	3.083	168.450
N ₂	0.3526	3.13512	127.426	0.3512	2.7903	141.899
H ₂ S	0.3600	3.10000	212.047	0.4508	3.2160	199.174
O ₂	0.3600	2.76730	166.370	0.3706	2.6507	166.002

Step 5 - CO₂ solubility correction and water activity

In cases where CO₂ is the hydrate-forming compound, the effect of CO₂ solubility in the aqueous phase should not be overlooked, as opposed to hydrocarbons that are not acidic and have a negligible quantity dissolved in water (Rasoolzadeh et al., 2025). The mol fraction of CO₂ dissolved in water, x_{CO_2} , was calculated by using equation below of the Krichevsky-Kasarnovsky formula provided by Nasrifar and Moshfeghian (2001):

$$x_{CO_2} = \frac{f_{CO_2}}{\eta_3} \exp \left\{ -\frac{P\eta_1 - \eta_2}{RT} \right\}$$

where f_{CO_2} is the gas-phase fugacity of CO₂ (MPa), P is the total pressure (MPa), R is the universal gas constant (8.314×10^{-3} MPa·m³·kmol⁻¹·K⁻¹), T is the temperature (K), and the empirical constants η_1 , η_2 , η_3 for CO₂ are 0.3411, 22.33, and 899,683.5 respectively (Nasrifar and Moshfeghian, 2001). The water mole fraction in the aqueous phase is then obtained from:

$$x_w = 1 - x_{CO_2}$$

and the water activity for the uninhibited CO₂-containing system is estimated using the Lewis-Randall rule (Nasrifar and Moshfeghian, 2001):

$$a_w = x_w$$

This adjustment addresses the small decrease in water activity due to CO₂ dissolution, which is not insignificant for the CO₂ hydrate systems in the consistency verification dataset. For all hydrate-forming substances that are not acids (CH₄, C₂H₆, C₃H₈, C₄ compounds, N₂), $a_{vv} = 1$ was used without modification.

Step 6 — Iterative solution for T at each pressure P

Given all component-related parameters, equation of the vdW-P equilibrium condition is solved numerically by iterations at the desired equilibrium temperature T at every given pressure P, as follows:

- 1 First, an initial estimate T^0 of temperature is calculated from the Rasoolzadeh et al. (2025) single-component equation at the corresponding gas component parameter values, or by linear interpolation along the pressure-temperature curve in the case of a multicomponent gas mixture.
- 2 Next, all temperature-dependent parameters are calculated using the current temperature value, including enthalpy of vaporization $\Delta h_{vv}^{(\beta-L)}(T)$, fugacity f_j of each gas component j using the PR equation of state, Langmuir constants $C_{hj}(T)$ using Kihara parameters and a_{vv} using the Krichevsky-Kasarnovsky equation in case of CO₂ presence.
- 3 LHS and RHS of vdWP equation are computed. In case $|LHS-RHS| < 10^{-6}$, the temperature T is considered converged and accepted as the uninhibited equilibrium reference temperature T_0 at that pressure.
- 4 Otherwise, bisection algorithm is used to refine temperature estimation at a physically justified temperature interval (normally between 240-310 K for the inhibited hydrate formation process) and Steps 2-4 are repeated until convergence is reached. This numerical routine was coded in Python and applied for every multicomponent gas dataset in the compilation.

Thus, obtained temperatures T_0 at every pressure were taken as uninhibited equilibrium reference temperatures needed for calculation of ΔT , $\Delta T/(T_0 T)$, and the slope A_{vv} for consistency criteria 1, 2, and 3 respectively.

Identification of hydrate structure

Prior to solving Eq. 11, it is necessary to find out whether hydrate crystals formed by gas mixture are of structure I or II, because only those specific cage parameter values v_h , R' , z and thermodynamic references $\Delta\mu^\circ_{vv}$, Δh°_{vv} , $\Delta v^\wedge(\beta-L)_{vv}$ associated with a particular structure will apply. Structure assignment was done according to the standard rule based on sizes of guest molecules (Sloan and Koh, 2008): methane and ethane hydrates are of structure I; propane, isobutane and natural gas mixtures which contain significant fractions of these hydrocarbons hydrates are of structure II; CO₂ hydrates are of structure I; N₂ hydrates are of structure II. For mixed gas compositions, the structure was determined based on dominant heavy gas component and then checked using published experimental results.

3.3.3 The Three Consistency Criteria of Sa et al. (2018)

Consistency evaluation has been performed separately for each individual dataset, defined as unique combination of gas composition, inhibitor type and inhibitor concentration according to the algorithm described by Sa et al. (2018).

Criterion 1 - linearity of the Clausius-Clapeyron plot. As predicted by the Clausius-Clapeyron equation, the dependence $\ln P(T)$ at the point of hydrate equilibrium should be linear within temperature interval of 10–20 K (Sa et al., 2018):

$$\frac{d \ln P}{d(1/T)} = - \frac{\Delta H_d}{zR}$$

where ΔH_d represents the heat of dissociation of hydrate and z is the compressibility factor of gas. Linearization is possible if the values of ΔH_d and z remain nearly constant within the small temperature interval covered by the dataset. Linear regression analysis was performed for each dataset using the expression $\ln P = (A/T) + B$, where the correlation coefficient (R^2) was used to evaluate the goodness of fit. Classification of consistency is:

Table 6. Classification of consistency criterion 1

Classification	Criterion
Pass	$(1 - R^2) \times 100 < 2.5\%$
Acceptable	$2.5\% \leq (1 - R^2) \times 100 < 5\%$
Fail	$(1 - R^2) \times 100 \geq 5\%$

Criterion 2 - Heat of dissociation. Since simple THIs do not change the hydrate crystal lattice, they affect hydrates solely by decreasing the activity of water in the aqueous phase. Thus, the heat of dissociation of the hydrate ΔH_d , which is a property dependent exclusively on the crystal structure and the guest molecule type, ought to be constant for any THI regardless of its concentration (Sa et al., 2018). The coefficient A_{thi} obtained for each THI inhibited system from linear regression was then compared to the coefficient A_{vv} from the uninhibited system for the same hydrate formers:

$$\text{Slope deviation} = \frac{A_{THI} - A_W}{A_W} \times 100\%$$

Table 7. Classification of consistency criterion 2

Classification	Criterion
Pass	$ \text{slope deviation} < \pm 5\%$
Acceptable	$\pm 5\% \leq \text{slope deviation} < \pm 10\%$
Fail	$ \text{slope deviation} \geq \pm 10\%$

Criterion 3 – Consistency of water activity. The normalized temperature difference $\Delta T/(T_o T)$ is determined from each data point, where T_o represents the hydrate equilibrium temperature without inhibition at a certain pressure level while T denotes the inhibited temperature. Within the thermodynamic theory of Sa et al. (2018), it is expected that this parameter is equal to $\ln(a_w)$ and is independent of the temperature:

$$\frac{\Delta T}{T_o T} = \beta \ln a_w(x_i) \approx \text{const.}$$

where $\beta = -nR/\Delta H_d$ is a constant depending on hydrate structure, guest species, and hydration number. The relative standard deviation (RSD) of $\Delta T/(T_o T)$ across all data points in each dataset quantifies consistency:

Table 8. Classification of consistency criterion 3

Classification	Criterion
Pass	$\text{RSD of } \Delta T/(T_o T) < 5\%$
Acceptable	$5\% \leq \text{RSD} < 10\%$
Fail	$\text{RSD} \geq 10\%$

Pass + acceptable acceptance criteria were used, based on the criteria recommended by Sa et al. (2018), which recognizes the presence of some level of uncertainty associated with all hydrate phase equilibrium experiments. The datasets that were categorized as “fail” on any one

of the criteria were omitted from the modelling dataset. It was important to choose the Pass + Acceptable acceptance criteria as it ensured a representative distribution of data sets of the particular inhibitors in the training set without omitting a majority due to the Pass criteria.

3.3.4 Consistency Screening Results

After the application of all three criteria to the complete set of 1,751 data points, 1,167 data points were retained, thereby achieving a data filtration ratio of 33.4%. These 584 eliminated points were spread out over many different papers and systems of inhibitors, with the primary types of eliminated points in the Heat of Dissociation (Criterion 2 eliminations) and thermodynamically non-realistic trends in the temperature depression as a function of concentration (Criterion 3 eliminations). The importance of such a data filtration step can be appreciated by its exact parallelism to the results obtained by Rasoolzadeh et al. (2025) for uninhibited systems.

3.4 Exploratory Data Analysis and Preprocessing

3.4.1 Exploratory Data Analysis

After performing consistency checks on the dataset, exploratory data analysis (EDA) was conducted systematically before fitting any models. This was done using matplotlib and seaborn, and involved the following steps:

Correlation heatmap analysis. Figure 33 that could be found in Appendix section shows the Pearson correlation heatmap that was created for all 37 predictor variables and the response variable (temperature) for the original dataset without consistency check and preprocessing. This helped in determining highly correlated pairs of variables, especially among the inhibitor concentration columns and the gas composition columns. Multicollinearity could be an issue when some modeling techniques are used. Patterns in correlations between individual inhibitors and the response variable were also observed to verify that there is a negative correlation between higher concentrations of inhibitors and lower temperatures for hydrate formation. Moreover, inputs with minimal correlation were identified in order to eliminate them from the dataset. Heatmap of the final dataset inputs is given in the Figure 9 below.

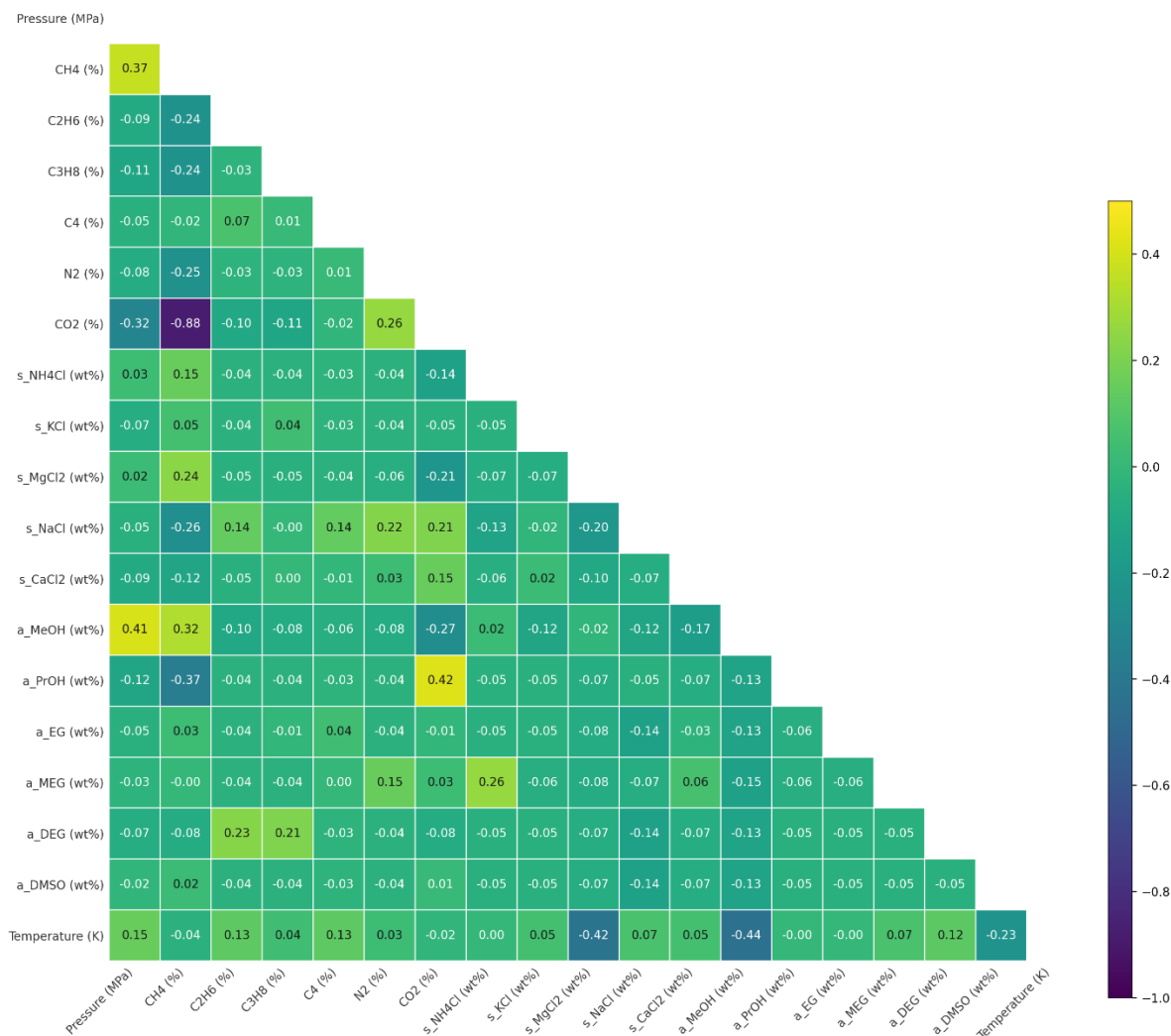


Figure 9. EDA outputs: correlation heatmap after preprocessing

Distribution Analysis

The statistical distributions of each feature and the target variable were investigated by means of histograms and density plots. Skewness and kurtosis were calculated using `scipy.stats`. The distribution of the target variable, temperature (K), was inspected in terms of its normality and whether there is any clustering in small temperature intervals. The distribution of each inhibitor concentration feature was explored for zero-inflation, as it is anticipated due to the high frequency of zero concentrations for each type of inhibitor, resulting from the multi-column encoding approach that was used. The distribution analysis plot could be found in Appendix Section as Figure 34.

Box Plot Analysis

The Box plots were developed for every individual feature as well as the temperature feature considered as the target feature. Box plots help to identify the data points considered to be outliers since they lie beyond 1.5 IQR away from the quartiles. Box Plots also enabled the evaluation of the measure of central tendency and dispersion for each feature under consideration. The box plots for each input parameter groups are given below in Figures 10-15.

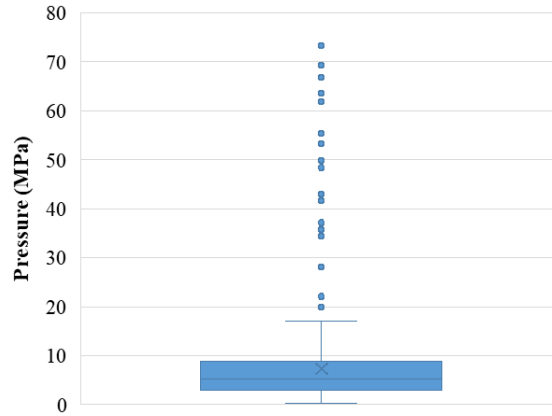


Figure 10. EDA outputs: box plots for Pressure input.

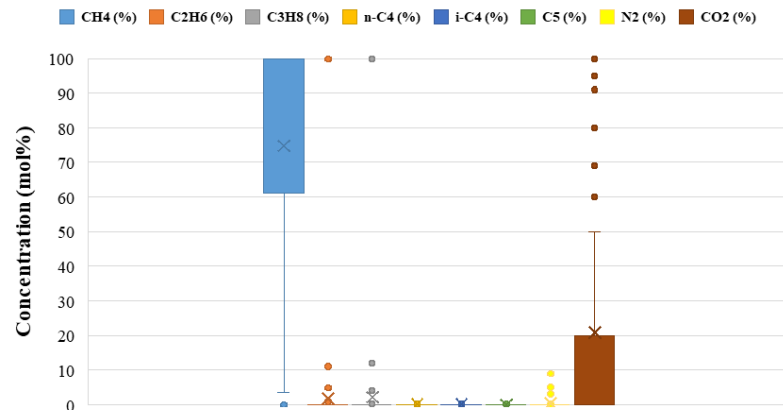


Figure 11. EDA outputs: box plots for Gas composition input

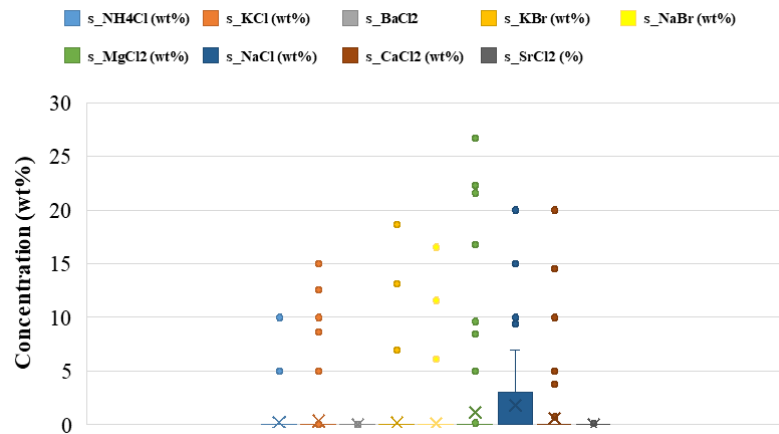


Figure 12. EDA outputs: box plots for Salts Inhibitors inputs

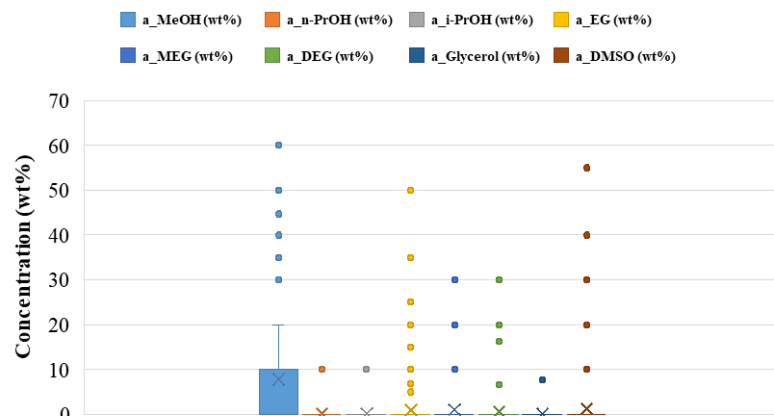


Figure 13. EDA outputs: box plots for Alcohols Inhibitors inputs

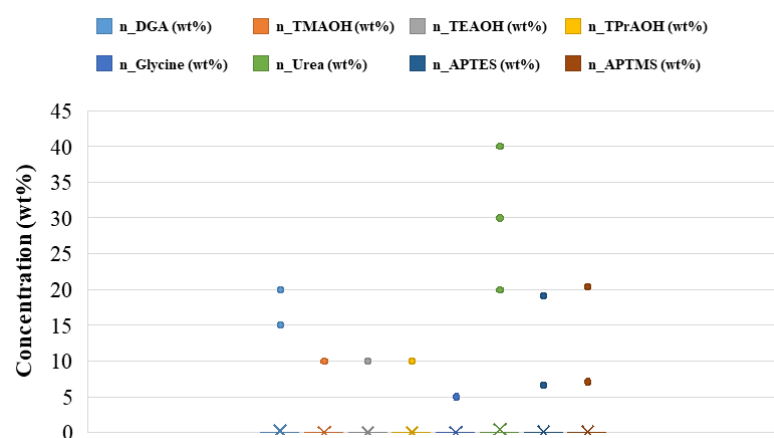


Figure 14. EDA outputs: box plots for Nitrogen-based Inhibitors inputs

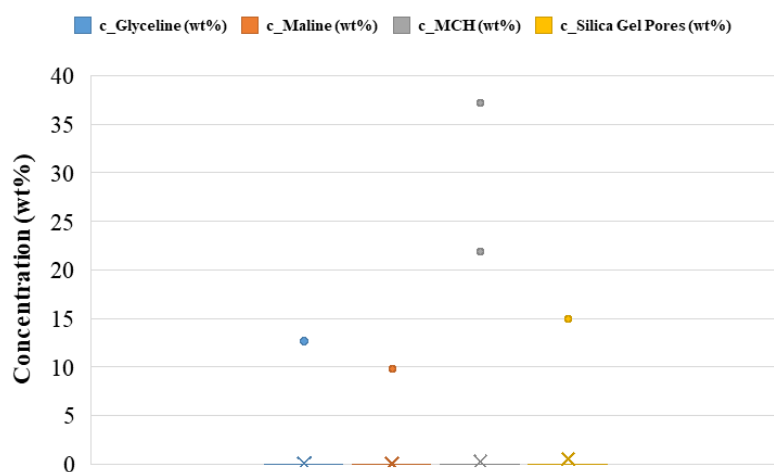


Figure 15. EDA outputs: box plots for Complex Inhibitors inputs

3.4.2 Outlier Removal

Outliers were manually detected after EDA based on a combination of statistical analysis (using box plots) and physical plausibility analysis. Outlier points were analysed one-by-one within

their feature vectors and corresponding papers to see whether the outlier point was a physically valid result of an experiment conducted under extreme conditions or just erroneous data entry or measurement. Only those data points that did not fit a physically plausible range or were statistically far from other points without any thermodynamic reason for it were excluded. This exclusion together with removal of duplicates resulted in exclusion of 245 data points from the initial sample of 1,167 points.

3.4.3 Feature Matrix Construction and Encoding

The final cleaned dataset of 922 data points was imported into Python using pandas from a structured Microsoft Excel workbook. The input feature matrix X and the target vector y were defined as:

```
X = Total_Data[['Pressure (MPa)', 'CH4 (%)', 'C2H6 (%)', 'C3H8 (%)', 'C4 (%)', 'N2 (%)', 'CO2 (%)',
```

```
's_NH4Cl (wt%)', 's_KCl (wt%)', 's_MgCl2 (wt%)', 's_NaCl (wt%)', 's_CaCl2 (wt%)',
```

```
'a_MeOH (wt%)', 'a_PrOH (wt%)', 'a_EG (wt%)', 'a_MEG (wt%)', 'a_DEG (wt%)', 'a_DMSO (wt%)']
```

```
y = Total_Data['Temperature (K)']
```

There are 18 input features in the feature matrix, while the output variable has only one continuous target – temperature for hydrate formation in Kelvin. All of them are divided into three groups: one continuous thermodynamic property (pressure); six gas composition properties (moles in %); five salt inhibitors concentrations (wt%); six organic inhibitors concentrations (wt%). The salt inhibitors will have the prefix 's_', while the prefix 'a_' will be used to denote alcohol/glycol/solvent inhibitors.

It allows us to keep all meanings of inhibitors and gas components intact without encoding them either categorically or using one hot encoding. If an inhibitor type is not present for some particular examples, then the appropriate column has zero values; likewise, in case there is only one component in gas, then all columns except the one with the existing compound will have zeros. Thus, the model will get the additive effect and potential synergy of inhibitors if several of them are combined in one particular example: e.g., for example when NaCl and MeOH inhibitors are combined, the 's_NaCl' and 'a_MeOH' columns will both contain the values of their concentrations.

3.4.4 Duplicate Removal

After the creation of the feature matrix, all exact duplicate entries, which are those that have the same values in all the 18 feature inputs and the temperature output, were eliminated. Duplicate entries might occur where the same experimental data have been published in different studies, when the author provides rounded average numbers together with raw data, or when data from several sources are combined. Keeping duplicates increases the size of the training set and gives an inaccurate representation of certain experimental conditions.

3.4.5 Feature Scaling

No feature scaling was used in tree-based models (DT, RF, GBR, XGBoost, CatBoost) since tree-based algorithms are insensitive to any monotonic feature scaling. The feature scaling does not affect their results as the only information the algorithms use when making splits is the order of the values (González et al., 2020).

Standardisation of features should be performed for Support Vector Regression because SVR makes calculations of distances in feature space, where such distances are highly dependent on features' scales. For this purpose, a standard Scaler from scikit-learn was included into the Scikit-learn Pipeline:

```
pipeline = Pipeline([  
  
    ('scaler', StandardScaler()),  
  
    ('svr', SVR())  
  
])
```

It transforms each feature by adjusting its mean value to zero and scaling its standard deviation to one. The transformation is done according to the train dataset's statistics.

3.4.6 Train-Test Split

The final dataset after cleaning that contained 922 rows of data was then divided into train and test datasets using an 80/20 split of the data as follows:

```
X_train, X_test, y_train, y_test = train_test_split(  
  
    X, y, test_size=0.2, random_state=42)
```

It produced a train dataset containing 737 rows of data and a test dataset containing 185 rows of data. The random state value was set to 42 throughout the process. The test data was completely held back from the models' training and hyperparameter tuning processes but was used exclusively for performance evaluation in the final stage.

3.5 Machine Learning Model Development

3.5.1 Models Evaluated

A total of six machine learning regression models were created, trained, and compared: Decision Tree (DT), Random Forest (RF), Gradient Boosting Regression (GBR), XGBoost, CatBoost, Support Vector Regression (SVR). Collectively, the algorithms used represent all four of the algorithmic categories present in the hydrate phase equilibrium ML literature: decision tree algorithms, ensemble bagging algorithms, ensemble gradient boosting algorithms, kernel algorithms, and neural network algorithms. The range of the models allows for a comprehensive comparison between the various families in terms of predictive power, generalisation capabilities, and physical consistency.

3.5.2 Hyperparameter Tuning Procedure

The optimisation of the hyperparameters of all six algorithms was done in a two-step procedure. In the first step, a hybrid approach involving both the RandomizedSearchCV and the GridSearchCV function from the scikit-learn library was employed. A 5-fold cross-validation was used, with R^2 as the scoring criterion and a random state of 42 fixed for reproducibility. In the second step, fine-tuning of the hyperparameters with the greatest impact on the model was performed manually by analysing the disparity between training and testing data results. This strategy was selected due to the efficiency of RandomizedSearchCV when exploring large areas of hyperparameter space (González et al., 2020), and the possibility to refine specific hyperparameters found in the first step. All cross-validation was performed exclusively on the training dataset, and no information was available regarding the withheld testing dataset.

3.5.3 Decision Tree Regressor

Decision Tree (DT) regressor is a non-parametric supervised learning method developed in its current form by Breiman et al. (1984) using the Classification and Regression Tree (CART) approach. The CART technique uses a splitting criterion for creating a binary decision tree based on predictor variables, thus recursively segmenting the predictor space into binary

segments in order to maximize the homogeneity of samples in each resulting node. In the case of a regression tree, CART relies on the principle of residual reduction to choose the next best split: at every iteration, all possible binary splits are considered and the one which results in the minimal weighted sum of mean squared error in the two resulting nodes is used. The procedure continues until certain stopping criterion is fulfilled (e.g. minimum number of samples per leaf node, minimum impurity decrease, maximum tree depth).

Averaged target values of the training samples reaching the respective terminal leaf node serve as predictions for all samples arriving at that particular leaf. Decision trees provide great interpretability, do not require feature scaling, are able to work with categorical predictors without any modification and do not suffer from linearity constraints. Nevertheless, CART can build unstable trees as a small change in the sample data can result in large differences in tree structure; furthermore, individual decision trees are susceptible to overfitting when growing to their maximum depth. Hence, maximum tree depth is the only regularization method employed for the standalone DT model in this analysis.

The following hyperparameters were optimized: `max_depth` (the maximum number of splits between root and leaf), `min_samples_split` (the minimum number of samples necessary to perform an internal node split), and `min_samples_leaf` (the minimum number of samples at each leaf node). The optimized hyperparameter values are listed in Table 9.

Table 9. Hyperparameter search space — Decision Tree (DT)

Hyperparameter	Search Space	Notes
Tree Depth & Size		
max_depth	[12, 15, 18, 20, 25, 30, 35]	<i>Controls tree complexity; higher = more overfit risk</i>
min_samples_split	[2, 3, 4, 5, 6]	<i>Min samples required to split a node</i>
min_samples_leaf	[1, 2, 3, 4]	<i>Min samples required at a leaf node</i>
Feature Selection		
max_features	[0.6, 0.7, 0.8, 0.9, 'sqrt']	<i>Fraction or method for feature subset at each split</i>
Pruning		
ccp_alpha	[0.0, 1e-5, 5e-5, 1e-4]	<i>Minimal cost-complexity pruning parameter; 0 = no pruning</i>

Search: RandomizedSearchCV, n_iter=120, cv=5, scoring=R², random_state=42.

3.5.4 Random Forest

The Random Forest (RF) was proposed by Breiman (2001) as a bagging-based ensemble learning technique, which aims to reduce the instability and overfitting properties of a single decision tree by adding randomness to the process of splitting nodes. In the RF framework, a number of trees T is first grown based on bagging with feature subset selection for splitting nodes. Bagging provides an estimate of the generalisation error for the resulting ensemble as well as individual trees' strength and correlation measures through out-of-bag (OOB) samples. Briefly, the algorithm works as follows: A number of trees T are grown; each tree is fitted on a bootstrap sample of the training set, leaving one-third of the data unrepresented in that sample (OOB samples). At each node in the tree-growing process, only a random subset of features m is selected to serve as potential split variables. For regression tasks, $m = \sqrt{p}$ (number of features p in case of classification, $1/3 p$ in case of regression), while for classification tasks, $m = \log_2 p$. For regression problems, the output of the RF is the average of the outputs of the decision trees forming the RF.

Bootstrap sampling coupled with random subset feature selection in decision tree growing helps achieve diversity between the trees and significantly reduces prediction variance as compared to the single decision tree without causing an additional increase in bias. The convergence theorem for RF error rate has proven that the error rate always converges when increasing the number of trees, which makes it clear that using more than the minimum required number of trees in the ensemble does not lead to degradation of generalisation performance - a major advantage over Boosting algorithms which tend to overfit when the number of rounds increases too much. Additionally, bagging allows calculating an on-going estimate of the generalisation error of the combined trees as well as measures of their strength and correlation using the OOB error rate.

In the RF framework, the variable importance measure (mean decrease in accuracy, MDA) is calculated by measuring how much the accuracy decreases after permuting the values of particular features in the out-of-bag (OOB) samples (Breiman, 2001). The features for which this measure yields high values are considered more important. The tuned hyperparameters in RF were `n_estimators`, `max_depth`, `min_samples_split`, `min_samples_leaf`, `max_features`. Values of hyperparameters are presented in Table 10.

Table 10. Hyperparameter search space — Random Forest (RF)

Hyperparameter	Search Space	Notes
Ensemble Size		
n_estimators	[200, 300, 400, 500, 600, 800]	<i>Number of trees in the forest</i>
Tree Structure		
max_depth	[20, 30, 40, 50, 60, None]	<i>Max depth per tree; None = fully grown</i>
min_samples_split	[2, 4, 5, 8, 10]	<i>Min samples to split an internal node</i>
min_samples_leaf	[1, 2, 4, 6]	<i>Min samples at each leaf</i>
Feature Selection		
max_features	['sqrt', 'log2', 0.3, 0.4, 0.5, 0.6]	<i>Feature subset strategy at each split</i>
Bootstrap		
bootstrap	[True, False]	<i>Whether to use bootstrap samples when building trees</i>

Search: RandomizedSearchCV, n_iter=180, cv=5, scoring=R², random_state=42.

3.5.5 Gradient Boosting Regression

Gradient Boosting Regression (GBR) constructs an additive ensemble of decision trees iteratively in such a way that every subsequent tree is trained on the negative gradient of the loss function – the pseudo-residuals of the model – rather than a bootstrap sample as is done in the case of bagging (González et al., 2020; Bentéjac et al., 2019). The additive ensemble is constructed iteratively in the following manner:

$$F_m(x) = F_{m-1}(x) + v \cdot \rho_m \cdot h_m(x)$$

where $h_m(x)$ is the m th regression tree fitted on the pseudo-residuals $r_i = -[\partial L(y_i, F(x))/\partial F(x_i)]_{F=F_{m-1}}$, ρ_m is the step size chosen by a line search, and $v \in (0, 1]$ is the learning rate (also known as shrinkage parameter) (Bentéjac et al., 2019). The learning rate determines the degree to which the output of every subsequent tree influences the overall result. Lower learning rates result in better generalisation of the algorithm at the cost of requiring more trees to fit the model (González et al., 2020). GBR can theoretically be optimised with respect to any differentiable loss function. In the current study, we will be minimising the squared error loss by default in order to do regression.

While the random forest grows multiple trees independently and in parallel, the GBR grows a single decision tree in each iteration and uses it to correct the mistakes made by the previous ensemble model. Consequently, GBR achieves very low bias but is highly sensitive to the

number of iterations and learning rate and prone to overfitting if the latter are not set properly (Bentéjac et al., 2019). Regularisation techniques include shrinkage, limiting the depth of trees and possibly using a stochastic approach by fitting a decision tree to a random subset of the training data in every iteration. The most relevant hyperparameters tuned in this experiment were the `n_estimators`, `learning_rate`, `max_depth`, `min_samples_split`, `min_samples_leaf`, and subsample parameters. Values of hyperparameters are presented in Table 11.

Table 11. Hyperparameter search space — Gradient Boosting Regressor (GBR)

Hyperparameter	Search Space	Notes
Boosting Configuration		
n_estimators	[200, 300, 400, 500, 700, 900]	<i>Number of boosting stages</i>
learning_rate	[0.01, 0.03, 0.05, 0.08, 0.1]	<i>Shrinkage rate applied to each tree's contribution</i>
max_depth	[2, 3, 4, 5]	<i>Max depth of individual estimators</i>
Tree Structure		
min_samples_split	[2, 4, 6, 8, 10, 15, 20]	<i>Min samples required to split a node</i>
min_samples_leaf	[1, 2, 3, 4, 6, 8]	<i>Min samples at a leaf node</i>
Subsampling & Features		
subsample	[0.6, 0.7, 0.8, 0.9]	<i>Fraction of samples used per stage (stochastic GBR)</i>
max_features	['sqrt', 'log2', 0.4, 0.6, 0.8, None]	<i>Feature subset strategy per split</i>
Loss Function		
loss	['squared_error', 'huber']	<i>Objective function; huber is robust to outliers</i>

Search: RandomizedSearchCV, n_iter=150, cv=5, scoring=R².

3.5.6 XGBoost

The XGBoost (eXtreme Gradient Boosting) algorithm represents a highly scalable and regularised gradient boosting algorithm invented by Chen and Guestrin and extensively described in Bentéjac et al. (2019) and González et al. (2020). In addition to the traditional gradient boosting objective function, XGBoost adds an explicit regularisation term controlling the complexity of the model:

$$\text{obj} = \sum_i L(y_i, F(x_i)) + \sum_t \Omega(f_t), \quad \Omega(f) = \gamma T + \frac{1}{2} \lambda |w|^2$$

where T denotes the total number of leaf nodes, w stands for the vector of leaf scores, γ is the minimum gain necessary to generate a split (pre-pruning), and λ denotes the L2 (or Ridge) regularisation coefficient on the leaves' weight (González et al., 2020). This is how adding the explicit regularisation term to the objective function allows XGBoost to learn more generalisable models compared to standard gradient boosting approaches.

XGBoost uses the second-order Taylor approximation of the loss function to obtain analytical solutions for the optimal values of leaf scores and gain of the split, allowing one to compute the optimal solutions much faster and efficiently compared to gradient approaches used in standard GBR (González et al., 2020). Given the formula above, the optimal value w_j^* of leaf j , as well as its gain, can be computed as follows:

$$w_j^* = -\frac{\sum_{i \in I_j} g_i}{\sum_{i \in I_j} h_i + \lambda}$$

$$\text{Gain} = \frac{1}{2} \left[\frac{(\sum_{i \in I_L} g_i)^2}{\sum_{i \in I_L} h_i + \lambda} + \frac{(\sum_{i \in I_R} g_i)^2}{\sum_{i \in I_R} h_i + \lambda} - \frac{(\sum_{i \in I} g_i)^2}{\sum_{i \in I} h_i + \lambda} \right] - \gamma$$

where g_i and h_i represent the first and second-order statistics of the loss function gradient. XGBoost is an efficient and scalable ensemble learning approach based on gradient boosting algorithm showing excellent performance across many studies. For instance, this algorithm ranked among the best in machine learning competitions and hydrate equilibrium prediction studies with reported R^2 values ranging from 0.8274 to 0.9956 (Cai et al., 2025). XGBoost was implemented using the xgboost Python package (XGBRegressor interface). Mainly tuning of such hyperparameters as `n_estimators`, `learning_rate`, `max_depth`, `gamma` (γ), `reg_lambda` (λ), `subsample`, and `colsample_bytree` (proportion of feature columns sampled per tree) was conducted. Final hyperparameter values are presented in Table 12.

Table 12. Hyperparameter search space — XGBoost (XGB)

Hyperparameter	Search Space	Notes
Boosting Configuration		
n_estimators	[400, 500, 600]	<i>Number of boosting rounds</i>
learning_rate	[0.15, 0.20, 0.25]	<i>Step size shrinkage to prevent overfitting</i>

Hyperparameter	Search Space	Notes
max_depth	[2, 3, 4]	<i>Max depth of each tree; kept shallow to reduce overfitting</i>
min_child_weight	[1, 2, 3]	<i>Min sum of instance weight in a child; higher = more conservative</i>
Subsampling		
subsample	[0.6, 0.7, 0.8]	<i>Fraction of training samples per tree</i>
colsample_bytree	[0.6, 0.7, 0.8]	<i>Fraction of features sampled per tree</i>
gamma	[0.1, 0.2, 0.3]	<i>Min loss reduction required for a split</i>
Regularization		
reg_alpha	[0, 0.001, 0.01]	<i>L1 regularisation on leaf weights</i>
reg_lambda	[1, 2, 3]	<i>L2 regularisation on leaf weights</i>

Search: RandomizedSearchCV, n_iter=250, cv=5, scoring=R², random_state=42.

3.5.7 CatBoost

CatBoost is a gradient boosting implementation created by Yandex with two novel contributions to the standard algorithm and XGBoost: the direct use of categorical predictors without prior transformation and ordered boosting that avoids prediction shift (González et al., 2020). Prediction shift refers to the problem that occurs in the gradient boosting method where the residual of each training sample is calculated with respect to a model whose training was partially completed based on the same sample, resulting in overfit gradients. The former issue is addressed by CatBoost through generating random permutations of the training set and calculating the residual of each sample based only on a model estimated by preceding samples in the current permutation - an approach similar in its essence to the use of cross-validation within the boosting framework. Categorical variables are handled differently by replacing each category in a predictor with a statistic of the target based on the preceding samples in the same permutation to avoid target leakage. The comparison with simple trees-based ML model (RF) and gradient boosting model (Gradient Boosting) is shown in the Figure 16.

Figure: Random Forest vs Gradient Boosting

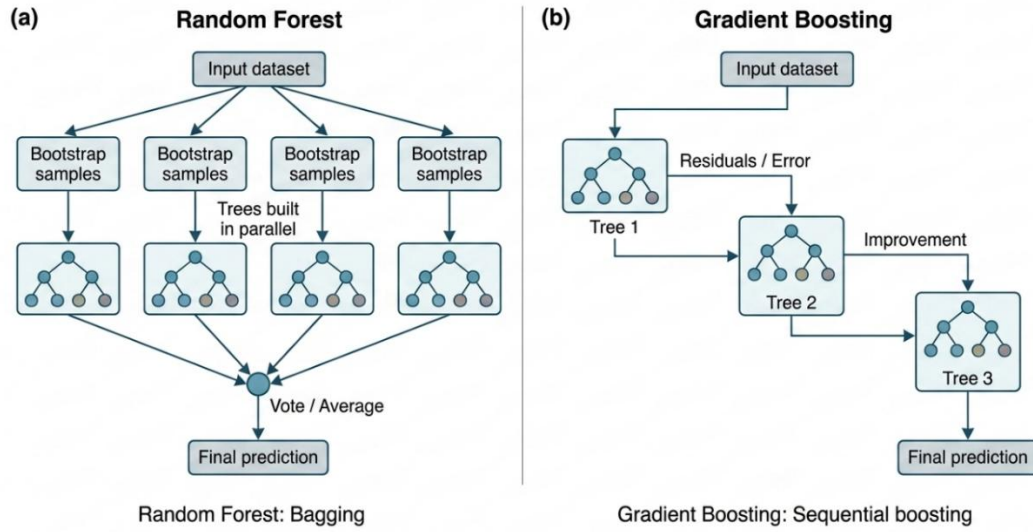


Figure 16. Random Forest and gradient boosting schematic comparison

As the base learners, CatBoost applies symmetric decision trees with each tree having nodes applying the same feature and the same splitting criteria at all levels. Such strict structure constraints may reduce expressiveness in individual trees but lead to better generalisation and faster predictions (González et al., 2020). In terms of this particular study, the properties mentioned above of CatBoost provide a great fit for the problem of predicting inhibition efficiency in the feature space of inhibitors and gases. Main tuning parameters were iterations, learning_rate, depth, l2_leaf_reg, and bagging_temperature. The final parameter settings are presented in Table 13.

Table 13. Hyperparameter search space — CatBoost

Hyperparameter	Search Space	Notes
Boosting Configuration		
learning_rate	[0.22, 0.25, 0.28, 0.30, 0.32]	Step size shrinkage
depth	[2, 3, 4]	Depth of symmetric trees
iterations	[1000, 1200, 1400, 1600]	Number of boosting iterations
Regularisation		
l2_leaf_reg	[8, 10, 12]	L2 regularisation coefficient on leaf values
random_strength	[3, 5, 7, 10]	Randomness for scoring splits; reduces overfitting
bagging_temperature	[0.3, 0.5, 0.7, 1.0]	Controls Bayesian bootstrap intensity
Tree Building		
border_count	[32, 64, 128, 254]	Number of splits for numerical features

Hyperparameter	Search Space	Notes
Fixed Setting		
loss_function	RMSE	<i>Fixed; not tuned during search</i>

Search: RandomizedSearchCV, n_iter=40, cv=5, scoring=R², random_state=42.

3.5.8 Support Vector Regression

Support Vector Regression (SVR), pioneered by Vapnik et al. and thoroughly described by Smola & Schölkopf (2004), aims to find a function $f(x) = \langle w, x \rangle + b$ that differs from real targets by less than ε while maximising model generalisation by minimising its complexity measured by $\|w\|^2$. According to the definition of the ε -insensitive loss function, it defines a "tube" around the regression curve where there is no loss (Smola and Schölkopf, 2004). The optimisation problem in the primal formulation can be stated as follows:

$$\min_{w, b, \xi, \xi^*} \frac{1}{2} \|w\|^2 + C \sum_{i=1}^N (\xi_i + \xi_i^*)$$

subject to $y_i - \langle w, x_i \rangle - b \leq \varepsilon + \xi_i$ and $\langle w, x_i \rangle + b - y_i \leq \varepsilon + \xi_i^*$, where $\xi_i, \xi_i^* \geq 0$. Regularisation parameter $C > 0$ determines the balance between model flatness and penalty of errors exceeding ε (Smola and Schölkopf, 2004). In dual formulation of this problem, it becomes clear that the optimal solution depends only on inner products of training data pairs, which allows one to use the kernel trick to conduct non-linear regression.

Radial Basis Function (RBF) kernel was chosen for this study:

$$K(x_i, x_j) = \exp\left(-\gamma \|x_i - x_j\|^2\right)$$

RBF kernel projects the input space of 18 features into an infinitely large reproducing kernel Hilbert space (RKHS), thereby allowing SVR to build highly non-linear decision surfaces in input space while solving a convex quadratic programming problem in RKHS (Smola and Schölkopf, 2004). Furthermore, unlike other kernels used in regression models (polynomial and sigmoidal), the RBF kernel does not imply any prior knowledge of functional dependence between temperature and features and has proven to work well in engineering applications with high-dimensional input spaces.

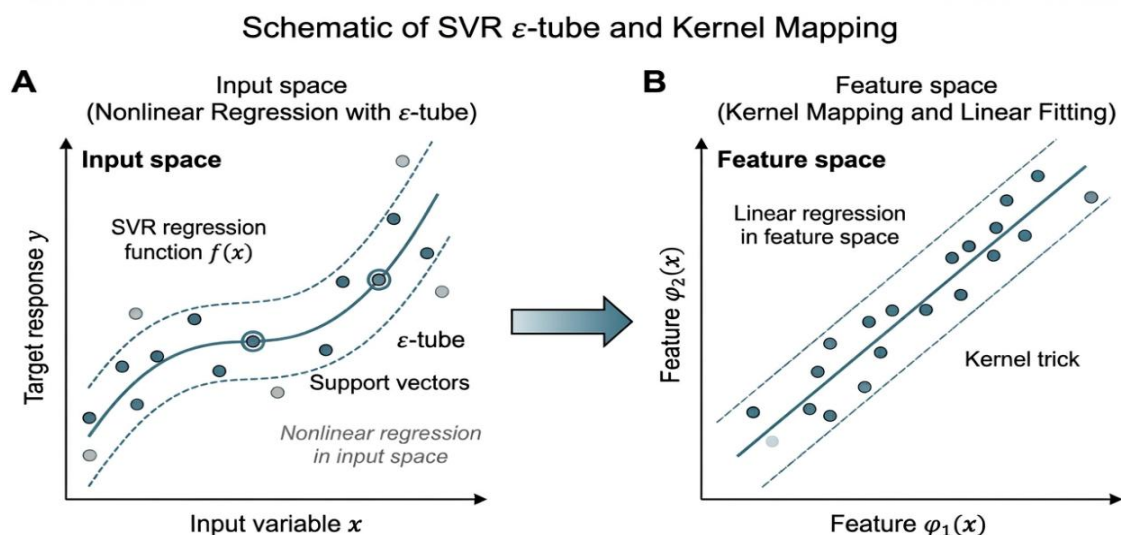


Figure 17. Schematic of SVR ϵ -tube (A) and kernel mapping (B)

Feature standardisation was applied before training within a Pipeline object as described in Section 3.4.5. The RandomizedSearchCV function ($n_iter = 72$, 5-fold CV, R^2 scoring) was used to tune hyperparameters with the following parameters and ranges defined: $kernel = 'rbf'$, $C \in \{1000, 2000, 3000, 5000\}$, $\epsilon \in \{0.1, 0.15, 0.2\}$, $\gamma \in \{0.04, 0.05, 0.06, 0.07, 0.08, 0.1\}$. Following that, hyperparameter fine-tuning was done manually with focus on minimising train-test performance difference. Final hyperparameters are listed in Table 14.

Table 14. Hyperparameter search space — Support Vector Regressor (SVR)

Hyperparameter	Search Space	Notes
Kernel & Margin		
kernel	rbf	Fixed radial basis function kernel
C	[1000, 2000, 3000, 5000]	Regularisation; higher C = smaller margin, less tolerance
epsilon	[0.10, 0.15, 0.20]	Width of the epsilon-insensitive tube; stable near 0.15
gamma	[0.04, 0.05, 0.06, 0.07, 0.08, 0.10]	RBF kernel coefficient; explored around confirmed 0.06
Preprocessing (Pipeline)		
scaler	StandardScaler	Fixed; z-score normalisation applied before SVR

Search: RandomizedSearchCV, $n_iter=72$, $cv=5$, $scoring=R^2$, $random_state=42$.

3.6 Model Evaluation Framework

Each trained model was evaluated on both the training and held-out test sets using five complementary performance metrics. Using both training and test metrics is essential for overfitting diagnosis: a large discrepancy between training and test performance indicates that the model has memorised the training data and will generalise poorly (Varoquaux and Colliot, 2023).

Coefficient of determination (R^2): Measures the proportion of variance in the target temperature explained by the model:

$$R^2 = 1 - \frac{\sum_{i=1}^N (y_i - \hat{y}_i)^2}{\sum_{i=1}^N (y_i - \bar{y})^2}$$

R^2 ranges from 0 (no explanatory power) to 1 (perfect prediction). For regression models in engineering applications, $R^2 > 0.98$ is typically considered very high accuracy.

Root Mean Squared Error (RMSE): Expresses the average prediction error in Kelvin, penalising large errors more than small ones:

$$RMSE = \sqrt{\frac{1}{N} \sum_{i=1}^N (y_i - \hat{y}_i)^2}$$

Mean Absolute Percentage Error (MAPE): A scale-independent accuracy measure in percentage:

$$MAPE = \frac{100}{N} \sum_{i=1}^N \left| \frac{y_i - \hat{y}_i}{y_i} \right|$$

MAPE is particularly useful for this study because the target variable (temperature in K) spans a range of ~ 57 K, and MAPE normalises errors relative to the absolute temperature value, enabling meaningful comparison across different temperature regions.

Akaike Information Criterion (AIC): A model selection criterion based on the theory of information that penalizes goodness of fit with model complexity, defined based on the idea of Kullback-Leibler information loss. The AIC formula is as follows: $AIC_k = n \ln(SSE) - n \ln(n)$

$+ 2(k+1)$, where n denotes the size of the training sample and k is the number of predictor terms, with smaller numbers being preferable. In the context of this research, k is treated as the number of parameters of the model estimated from the training dataset, and SSE stands for sum of squared errors of prediction. Calculation of AIC on both training and test samples allows for detection of overfitting: if the model is fitting well on the training set, producing extremely small residuals, but produces much larger residuals on the test set, its training AIC will be considerably more negative than test AIC.

Bayesian Information Criterion (BIC): Similar to the AIC but featuring a stronger penalty term for model complexity that scales with sample size logarithmically, $BIC_k = n \ln(SSE) - n \ln(n) + (k+1) \ln(n)$. The BIC penalty on the number of parameters in the model is significantly higher than that applied by the AIC after a certain sample size (which equals 7); hence, models with simpler structure are favoured. Smaller BIC values are better. This criterion may prove especially helpful in cases when one needs to compare models with different effective complexities, as was done in this study, where six models with effective complexities ranging from simple decision tree to complex ensemble classifier (XGBoost) were compared. Both AIC and BIC criteria have been calculated as part of Python machine learning evaluation code using the provided formulae with SSE evaluated based on predictions made on appropriate data subsets. All five metrics were computed on both training and test datasets for all six models and presented in Chapter 4.

3.7 Trend Analysis and Physical Validation

3.7.1 Predicted versus Actual Temperature Plots

Scatter plots of hydrate formation temperatures (predicted using SVR) were made against the experimental values for both the training and test datasets. The deviation from the 45° diagonal line representing perfect predictions was observed and analyzed quantitatively by metric distributions. This validation step reveals if there is any bias for certain temperature ranges, for instance, whether there is any systematic error in forming predictions at very low temperatures (< 250 K corresponding to high inhibitor concentrations) where the training dataset is sparse.

3.7.2 Pressure-Temperature Equilibrium Curve Comparison

The predictions from SVR will be checked physically to determine if the predictions agree with phase behavior knowledge. The SVR model will be provided with experimental inputs (such as

pressure and inhibitor concentration for a particular hydrate-forming system) and the hydrate equilibrium temperatures were plotted along with experimentally determined points from selected publications. The following three aspects will be validated through this process: (i) The increase in the hydrate formation temperature with an increase in pressure (for fixed inhibitor concentration), which implies that the P-T equilibrium curve must be inclined positively; (ii) The reduction in temperature with an increase in inhibitor concentration (for fixed pressure); and (iii) Correct ranking of inhibitors according to inhibition efficacy.

3.7.3 Temperature Depression versus Mixed Inhibitor Concentration

A specific trend analysis was done based on the dataset from Semenov et al. (2024) involving $\text{CH}_4 + \text{MgCl}_2 + \text{MeOH}$ in order to verify if the SVR model learns an accurate representation of the additivity property of the temperature depression due to inhibitors. ΔT (temperature depression, which refers to the difference between the temperature of the uninhibited equilibrium state and that of the inhibited state at the same pressure) against the concentration of MeOH in wt% at different concentrations of MgCl_2 (wt%) was illustrated in Image 2. Physical expectation dictates that ΔT increases as MeOH concentration increases for a fixed MgCl_2 value and that ΔT decreases when the MgCl_2 concentration increases for the case of zero MeOH. The analysis verifies if the SVR model successfully learns the correct physics of the system even when there are no physical constraints incorporated in the training process.

3.7.4 Supervised Interpolation within the Training Domain

An analysis for interpolation within the training domain was conducted using the Semenov et al. (2024) $\text{CH}_4 + \text{MgCl}_2 + \text{MeOH}$ dataset, where the original experimental MgCl_2 concentrations are 5.12, 8.43, 16.8, 22.28, and 26.68 wt%. New inputs were generated based on the training set by replacing the MgCl_2 concentration values with new values (specifically 12.615, 19.54, and 24.48 wt%), while all other features (pressure, composition of gases, and MeOH concentration) remain unchanged. Moreover, an additional data point at 6.775 wt% MeOH was included in order to verify if the model interpolates along the inhibitor concentration axis. Formation temperatures for each MgCl_2 concentration were plotted against the prediction points for the original experimental MgCl_2 concentration values in P versus T space.

3.8 SHAP Feature Importance Analysis

Interpretation of predictions made by the SVR model as well as quantification of the relative influence of each of the 18 input features in the predictions of the hydrate formation temperature was done using the SHapley Additive exPlanations (SHAP) approach (Lundberg and Lee, 2017). SHAP provides an approach to interpreting predictions, whereby each feature is assigned an important measure, called SHAP value, for each individual prediction. This concept has been formulated using cooperative game theory principles and each feature is considered as a player whose importance is evaluated through its marginal contributions to the prediction of the model when it is included into all possible coalitions, as follows:

$$\phi_i = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|! (|N| - |S| - 1)!}{|N|!} [f_{S \cup \{i\}}(x_{S \cup \{i\}}) - f_S(x_S)]$$

where N denotes the set of all features, S is any coalition of features which does not contain feature i , and $f_S(x_S)$ is the model prediction obtained using features in S . This method employs the Shapley values which are the game-theoretically optimal explanations of cooperative games and allows one to develop a unified framework for interpreting model predictions, which was suggested by Lundberg and Lee (2017).

The SHAP analysis was performed for the predictions made by the SVR model for the entire training + test datasets ($N = 922$) to provide a complete picture of how this model operates at all operating conditions. For each of the 922 predictions, a SHAP value was calculated for each of the 18 features, obtaining a matrix of SHAP values which could be used for determining the overall importance of features and explaining individual predictions. Such global analysis allows revealing physically relevant information about feature contributions, such as whether pressure or specific inhibitor concentrations dominates temperature prediction and whether inhibitors are correctly prioritized.

4. RESULTS

4.1 Overview

The findings from the process of constructing and evaluating the machine learning models that were discussed in Chapter 3 are presented in this chapter. The results obtained in this study are divided into four major sections: comparison of the performance of all six trained ML models according to five statistical performance metrics for both training and test sets; analysis of the most successful model (Support Vector Regression) in terms of predicted temperatures versus observed temperatures graphs; evaluation of predictions provided by SVR with respect to experimental equilibrium temperatures in the PT-plane considering different methanol concentrations; and, finally, results from the supervised interpolation analysis and the ΔT trends analysis.

4.2 Comparative Model Performance

4.2.1 Metrics Table

Table 4.1 below provides the full table of the five metrics for all six trained ML models: R^2 , RMSE, MAPE, AIC and BIC. All the values are rounded up to the decimal place allowing distinguishing between close values of the same metrics for various models.

Table 15. Performance metrics for all six ML models: final tuned run (training and test sets)

Model	R^2 Train	R^2 Test	RMSE Train (K)	RMSE Test (K)	MAPE Train (%)	MAPE Test (%)	AIC Train	AIC Test	BIC Train	BIC Test
DT	0.997	0.902	0.516	2.859	0.104	0.792	-853.0	393.1	-767.3	452.5
RF	0.999	0.967	0.020	1.651	0.001	0.432	-5257.9	207.4	-5172.2	266.8
XGBoost	0.998	0.987	0.394	0.990	0.098	0.249	-1216.5	25.1	-1130.8	84.6
CatBoost	0.998	0.989	0.423	0.931	0.071	0.153	-1230.6	11.5	-1143.1	72.6
GBR	0.998	0.989	0.447	0.986	0.105	0.250	-1044.9	29.2	-959.1	88.7
SVR	0.995	0.989	0.471	0.963	0.111	0.193	-549.1	60.0	-461.7	93.2

The dataset comprised 737 training points and 185 test points. All metrics are reported from the final tuned run. A comparison of the initial and final run metrics is discussed in Section 4.2.4.

4.2.2 Model Ranking by Test Set Performance

Examining the test set results, which constitute the primary and unbiased basis for model comparison, the seven models rank as follows in terms of RMSE (lower values are preferable):

1. **CatBoost** - RMSE = 0.931 K, $R^2 = 0.989$, MAPE = 0.153%
2. **SVR** - RMSE = 0.986 K, $R^2 = 0.989$, MAPE = 0.193%
3. **GBR** - RMSE = 0.986 K, $R^2 = 0.989$, MAPE = 0.250%
4. **XGBoost** - RMSE = 0.990 K, $R^2 = 0.987$, MAPE = 0.249%
5. **RF** - RMSE = 1.651 K, $R^2 = 0.967$, MAPE = 0.432%
6. **DT** - RMSE = 2.86 K, $R^2 = 0.902$, MAPE = 0.792%

According to the pure statistical metrics applied to the test set, CatBoost demonstrates the best performance, being characterized by the minimal RMSE (0.931 K), maximal R^2 (0.989), and minimal MAPE (0.153%). The tightest cluster is composed of CatBoost, SVR, and GBR, having equal $R^2 = 0.989$; their RMSE varies narrowly between 0.931 K and 0.986 K. However, it should be mentioned that SVR and GBR have nearly the same RMSE (0.986 K) and R^2 but different MAPE (0.193% vs. 0.250%). XGBoost is ranked fourth according to both RMSE (0.990 K) and MAPE (0.249%) and fifth regarding the slightly lower R^2 (0.987). RF is ranked fifth concerning RMSE (1.651 K) and R^2 (0.967) and sixth by MAPE (0.432%). Finally, DT is ranked sixth in RMSE (2.86 K), R^2 (0.902), and MAPE (0.792%), which can be explained by the well-known limitations of a single decision tree in modeling non-linear dependencies. The comparative graph for all models in terms of R^2 , RMSE, and MAPE is provided in Figure 18, while The AIC and BIC comparison is presented in Figure 19.

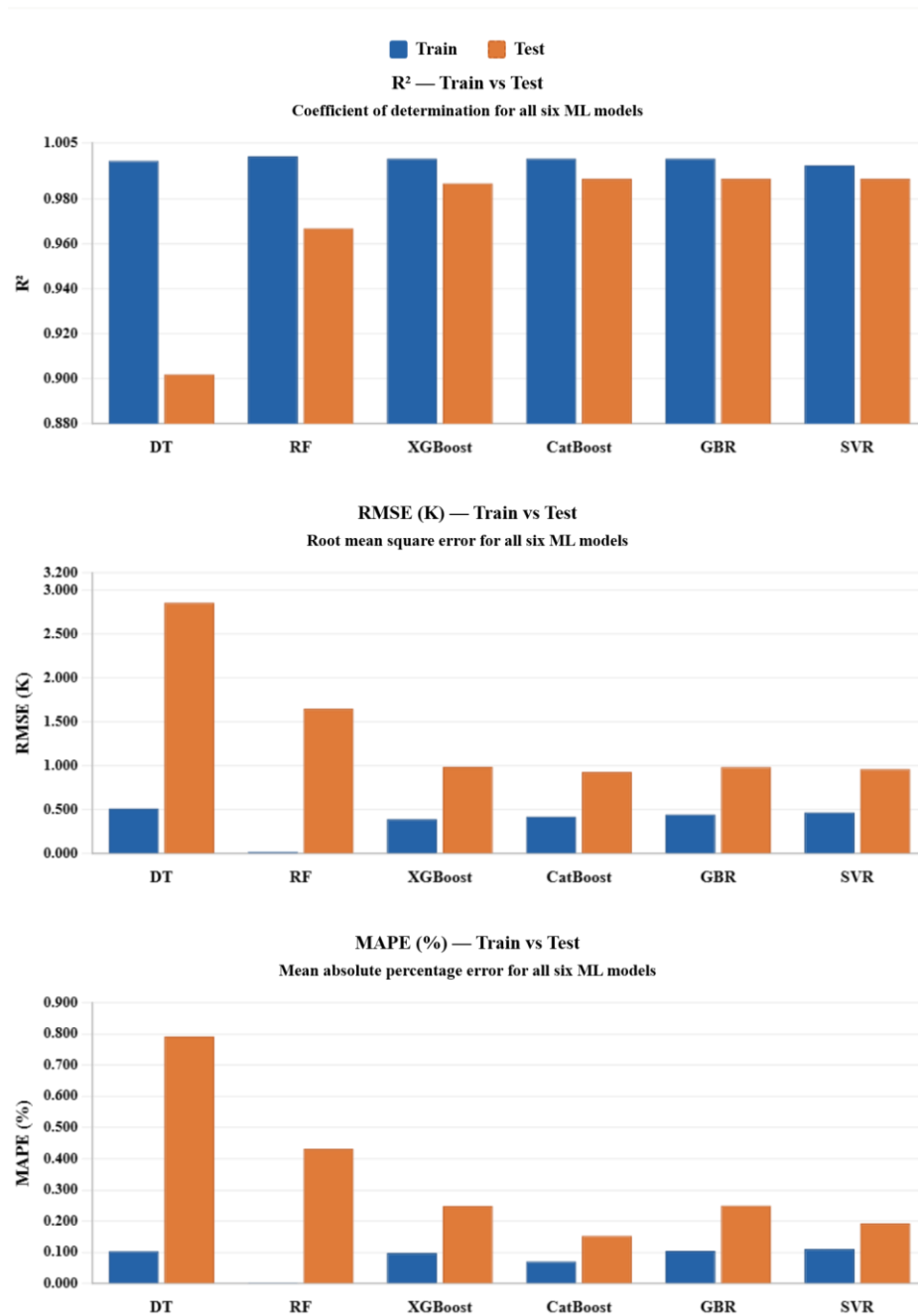


Figure 18. Model comparison bar charts: R^2 , RMSE, MAPE (train vs test) for all 6 models

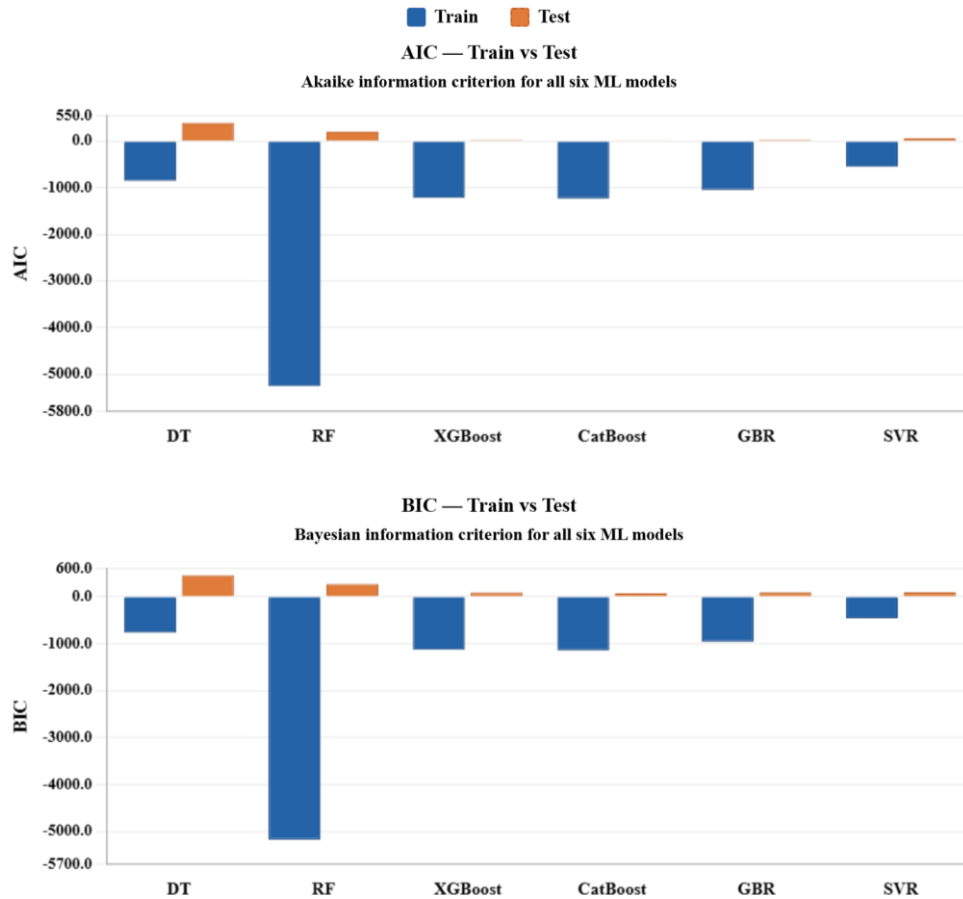


Figure 19. AIC and BIC comparison bar charts (train vs test) for all 6 models

4.2.3 Information Criteria Analysis and Overfitting Diagnosis

From another point of view, the AIC and BIC criteria offer additional information regarding the models' quality, as they include penalties for complexity and allow diagnosing overfitting based on the difference between training and test metrics. As expected, the AIC criterion is more negative in the training case for all the models, which demonstrates better adaptation to the data set that was used for learning. However, the difference between training and testing AIC is highly variable depending on the type of model and is indicative of its level of overfitting.

RF displays the highest difference between the training and testing cases according to the AIC criterion. Its training AIC equals -5257.9 , whereas the testing AIC equals 207.4 . It means that the model has memorized the training data almost completely, which is supported by its nearly perfect training RMSE of 0.020 K. At the same time, the extremely high difference between training and testing performance (training RMSE 0.020 K, testing RMSE 1.651 K; training AIC

–5257.9, testing AIC 207.4) and the relatively high test R^2 (0.967) clearly indicate the presence of overfitting. DT demonstrates different overfitting patterns since its training R^2 equals 0.997, but the testing R^2 drops to 0.902, which is the worst result among all the models. Thus, DT shows a well-known overfitting behavior because a single tree without pruning usually performs poorly when making predictions on previously unseen data. In the group of models based on gradient boosting, the training-test AIC gap is lower than for DT. Among the models, CatBoost shows the best results in terms of the AIC criterion. Specifically, its training AIC equals –1230.6, and its testing AIC equals 11.5, resulting in an approximate AIC gap of 1242. XGBoost produces similar results with an AIC gap of 1242, where the training AIC equals –1216.5, and the testing AIC equals 25.1. GBR exhibits somewhat higher AIC difference with a gap of approximately 1074 (training AIC –1044.9, testing AIC 29.2). Finally, SVR demonstrates a training AIC of –549.1 and a testing AIC of 60.0, resulting in the smallest AIC gap of 609.

In terms of BIC values, there appears to be consistency, whereby CatBoost and XGBoost have the best BIC test results of 72.6 and 84.6, respectively, implying that they have the optimal information-theoretic fit and complexity trade-off for unseen data. SVR has a slightly higher test BIC value of 93.2 due to the expense involved in the complexity approximation using a kernel, although its training BIC of –461.7 implies good regularization.

4.2.4 Comparison of Initial and Final Run Performance

The two-step method for fine-tuning was used for all models: starting with an automated step, followed by manual tuning. Evaluating the results of the two steps based on the values presented in Table 15, improvements were obtained for all models with tuning. However, the largest improvements were obtained for SVR (with the improvement of RMSE test from 2.070 K to 0.963 K). This improvement is quite large and is related both to the implementation of the standardization of features in the pipeline and the transition from the initial model that had no parameters tuned to the optimized model. Significant improvements in test RMSE were also obtained for CatBoost (from 1.026 K to 0.931 K), GBR (from 1.172 K to 0.986 K), and XGBoost (from 1.160 K to 0.990 K). No improvements in test RMSE were noticed for RF (1.695 K to 1.651 K) and DT (2.861 K to 2.859 K).

4.2.5 Model Selection: SVR as Best Overall Model

While CatBoost exhibits slightly better raw test set performance (RMSE = 0.931 K versus SVR's 0.963 K, both with $R^2 = 0.989$), SVR was chosen as the optimal model for this thesis for reasons other than raw test set performance, including:

First, the most important criterion for any engineering predictive model is physical generalization – producing physically realistic predictions for new cases where the physical processes occur in regions not observed in the training data. As seen from the supervised interpolation analysis provided in Section 4.5, the predictions generated by SVR provide physically sensible interpolations for intermediate inhibitor concentration values not covered by the training data. In contrast to this, while performing somewhat better numerically than SVR in test RMSE, CatBoost produces physically implausible interpolation results. Predictive ML models that fail to interpolate consistently across their training domains are practically useless engineering tools despite their high accuracy on test data.

Second, SVR possesses the smallest training-test AIC difference (609), showing a balance in the complexity vs. fit trade-off and implying uniform generalizability throughout the entire input domain, rather than overfitting to achieve high average prediction accuracy on a specific domain as exhibited by some other models.

Third, SVR exhibits the second smallest MAPE value of 0.193%, after CatBoost's 0.153%, which implies low, uniformly distributed relative errors even in hard-to-predict low-temperature, high-inhibitor-concentration region below 250 K

4.3 SVR Predicted versus Actual Temperature Analysis

4.3.1 Test Set Performance

The SVR-predicted hydrate formation temperature values are plotted against the corresponding experimental temperatures for the test set of 185 data points in Figure 20, with temperature varying between about 234 and 287 K. All the data points cluster very close to the 45° line, with no systematic curve, banded behavior, or regional bias discernible over the entire temperature span. This clustering around the perfect prediction line can be attributed to the test set $R^2 = 0.989$ and RMSE = 0.963 K. A few outlier points are noted, especially one point in the temperature range of about 270 to 275 K with the predicted temperature value around 276 K

compared to the experimental value of about 270 K, and another around the 280 K temperature mark. These outliers could be attributed to the few instances of the inhibitor/gas pairs, which were either inadequately covered in the training dataset or difficult from a thermodynamic point of view.

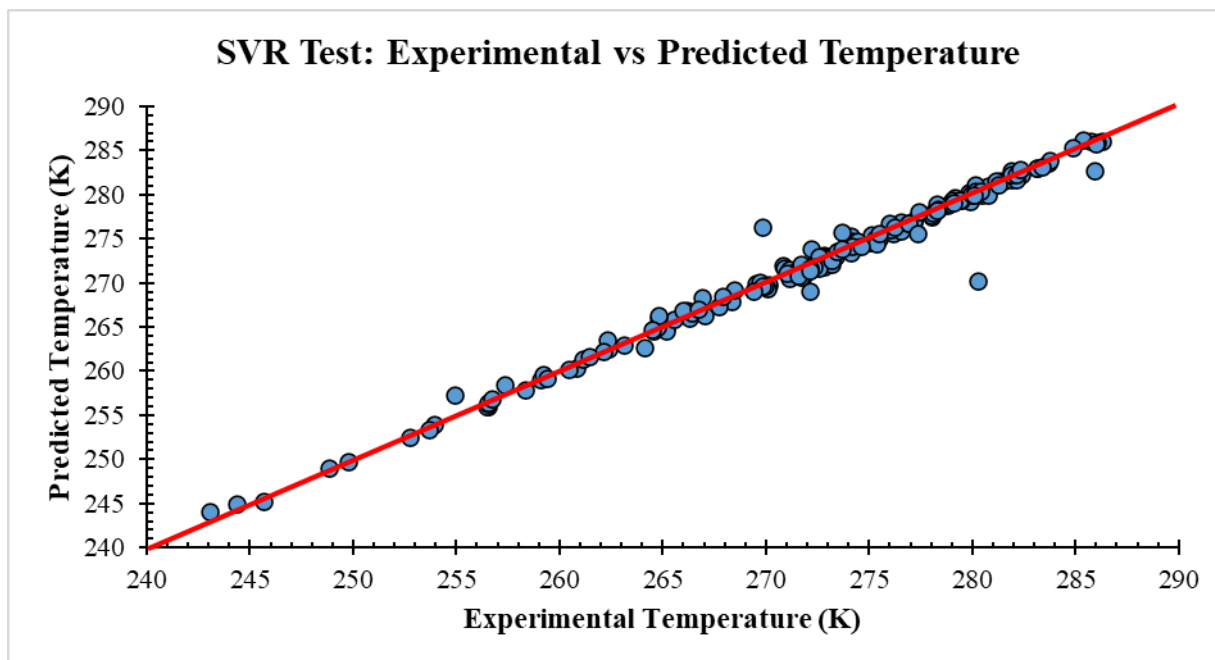


Figure 20. SVR Test: Experimental vs Predicted Temperature (K) scatter plot

4.3.2 Training Set Performance

Figure 21 below depicts the respective plot of the 737 training set data points over an extended temperature range from 235 to 291 K. In the training plot, it can be seen that all the data points follow the perfect prediction line with precision, as evident from the high training R^2 of 0.995 and RMSE of 0.471 K. Training results indicate slightly better performance than the test set, as can be seen in Figure 4.3 above, but the difference is quite small: training RMSE = 0.471 K and test RMSE = 0.963 K, which confirms that there has been no memorisation of the training data but rather learning of a generalisable representation by the SVR model. Only a few scattered points with errors in the region of 2 to 4 K can be seen in both plots, mostly in the temperature range of 265 to 275 K, representing non-linear regions of complex multicomponent gases or mixed inhibitors.

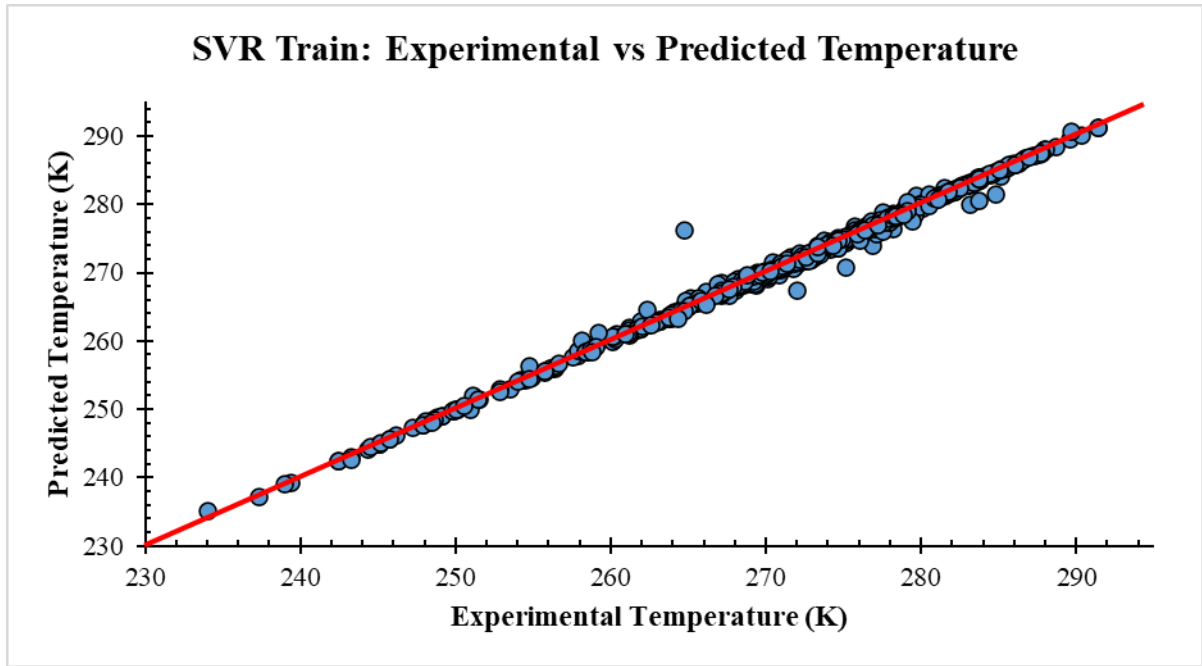


Figure 21. SVR Train: Experimental vs Predicted Temperature (K) scatter plot

The comparison of Figures 20 and 21 shows that the model under consideration is balanced and properly tuned because the distribution of residuals in both cases looks identical, which means that the generalization capability of our model can be considered realistic.

4.4 Pressure-Temperature Equilibrium Curve Comparison

4.4.1 Overview of *P-T* Analysis

In order to test if the SVR model is able to describe properly the thermodynamic behavior of the phase equilibrium curve of hydrates rather than simply the statistical distribution of temperature values in the database, a number of comparisons of phase equilibrium curves were made on the Semenov et al. (2024) system of methane hydrate formation in the presence of both MgCl_2 and methanol solutions. Comparisons were performed for four different concentrations of methanol: 0 wt%, 5 wt%, 10 wt%, and 20 wt% MeOH. Comparisons at different concentrations of MgCl_2 were done simultaneously for every concentration of methanol.

4.4.2 *P-T* Comparison at 0 wt% MeOH

Figure 22 shows the *P-T* equilibrium curves for five MgCl_2 concentrations (5.12, 8.43, 16.8, 22.28, and 26.68 wt%) in the absence of methanol. Several physically significant observations can be made from this figure.

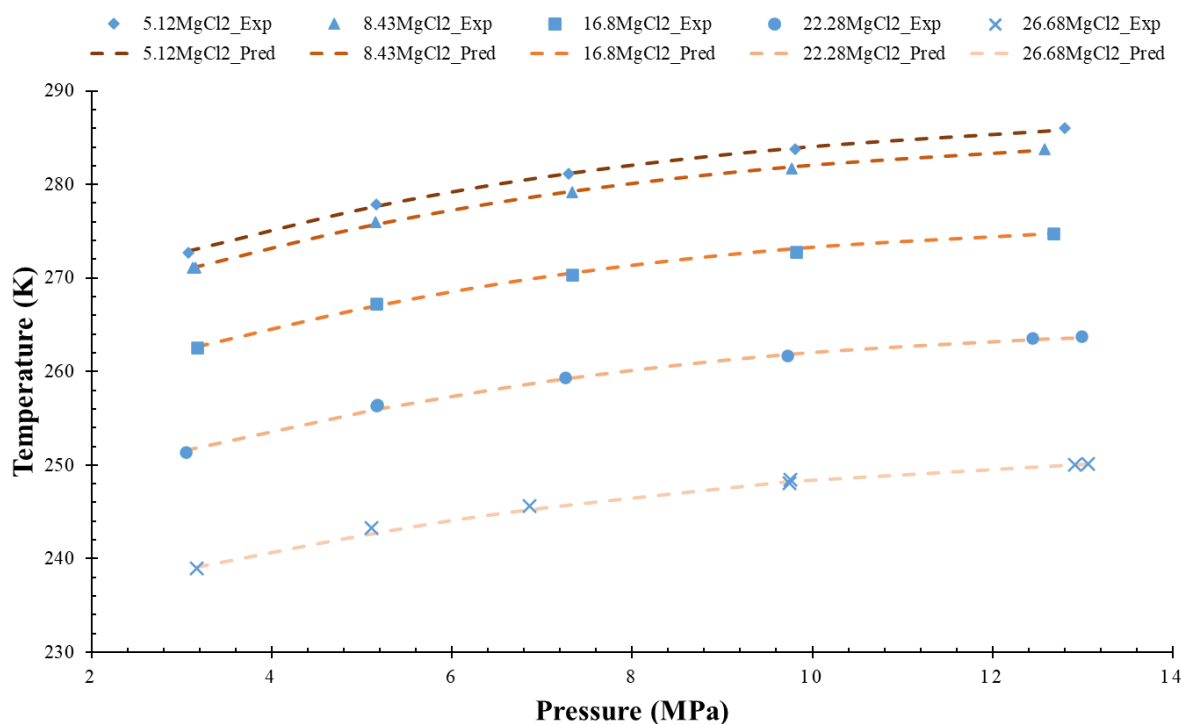


Figure 22. SVR P-T Equilibrium Curve Comparison plots at 0 wt% MeOH

First, the predictions produced by the SVR model exhibit correct positive slope of all equilibrium lines - the hydrate formation temperature rises with increasing pressure at all inhibitor concentrations according to physical principles set forth by the Clausius-Clapeyron relationship. Second, the predicted temperatures of the equilibrium are consistently decreasing with increasing MgCl_2 concentrations at all fixed pressures, correctly modeling the effect of inhibition caused by this chemical. When considering 3 MPa pressure point, the predicted temperatures range approximately from 272-273 K (when inhibitor concentration equals 5.12 wt%) down to 239 K (26.68 wt%) - approximately by 33 K at maximal concentration of inhibitor, corresponding to significant thermodynamic effect of Mg^{2+} ions on hydrate stability. Third, there is very high agreement between predicted and experimentally measured points across five different concentrations of inhibitor; the curves intersecting data points pass through or are very close to the data points. The biggest deviations occur when considering MgCl_2 concentration equal to 26.68 wt%, where equilibrium temperature becomes less than 250 K (this is the border area of data used for training the SVR model), and at 16.8 wt% MgCl_2 at some intermediate pressures.

4.4.3 P-T Comparison at 5 wt% MeOH

Figure 23 displays the P-T equilibrium curves for MgCl_2 concentrations of 5.13, 9.63, 16.82, and 21.6 wt% in the presence of 5 wt% methanol. Inclusion of 5 wt% methanol results in a decrease in the equilibrium curves relative to the 0 wt% MeOH curve – the curve corresponding to the lowest concentration of MgCl_2 (5.13 wt%) in the presence of 5 wt% MeOH achieves an equilibrium temperature maximum of about 283 K at a pressure of 13 MPa, while in the case of 5.13 wt% MgCl_2 and 0 wt% MeOH, the temperature maximum is reached at about 285 K, showing a decrease in temperature of 2 K due to the presence of MeOH. This additional temperature reduction resulting from methanol inclusion is accurately described by the proposed SVR model. The predicted curves agree well with the experimental data for each combination of MgCl_2 concentrations and MeOH. The range of equilibrium temperatures among various MgCl_2 concentrations remains unchanged, with the correct ranking among the curves. This shows that the SVR model can accurately describe hydrate equilibrium curves for mixed systems of inhibitors, which act on the system independently, reducing the hydrate equilibrium temperature.

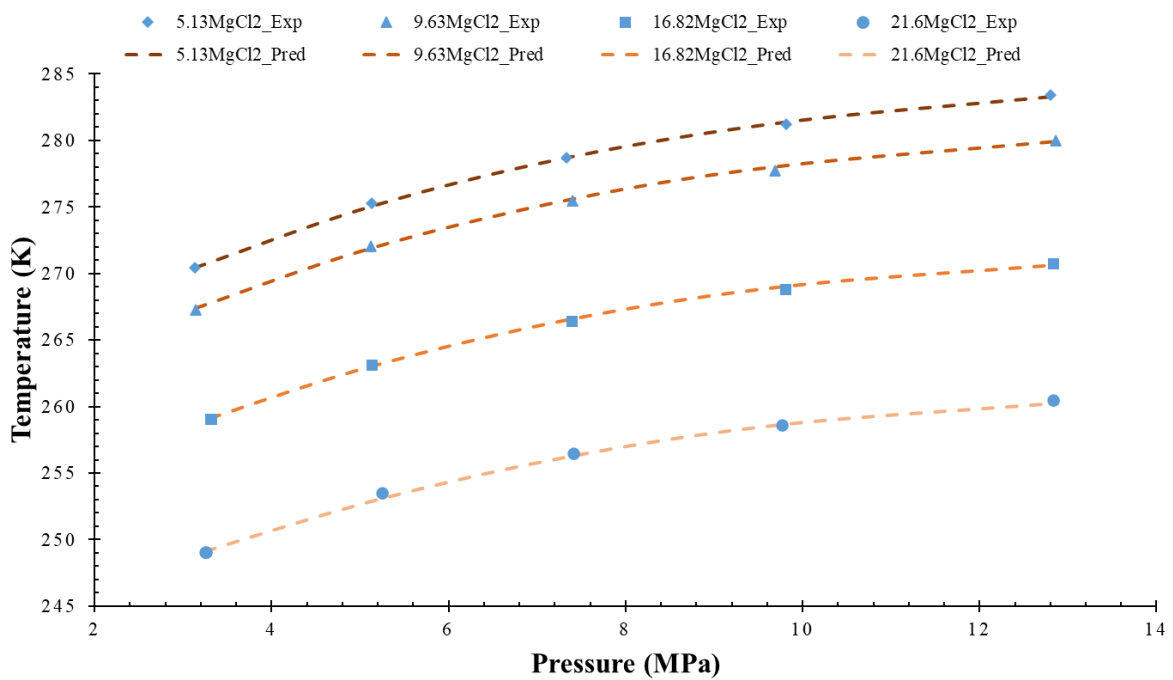


Figure 23. SVR P-T Equilibrium Curve Comparison plots at 5 wt% MeOH

4.4.4 P-T Comparison at 10 wt% MeOH

Figure 24 shows the results obtained for 4 different concentrations of MgCl_2 (5.14%, 9.62%, 16.82%, and 21.6%) at 10 wt% of methanol. The presence of additional methanol results in another decrease compared to that at 5 wt% MeOH. In the case of 10 wt% MeOH, the curve corresponding to the maximum MgCl_2 concentration (21.6%) almost touches the bottom of the graph, with equilibrium temperatures dropping to about 244 K at 3.5 MPa pressure, close to the minimum temperature value in the training dataset (234.03 K). However, the SVR model still follows the experimental points well in terms of both the magnitude of the change and even in the low-temperature part of the diagram (<250 K), despite the higher inhibitor content. It can be noted that the curve of the 21.6 wt% MgCl_2 at 10 wt% MeOH deviates more strongly from the experimental points compared to other inhibitor concentrations.

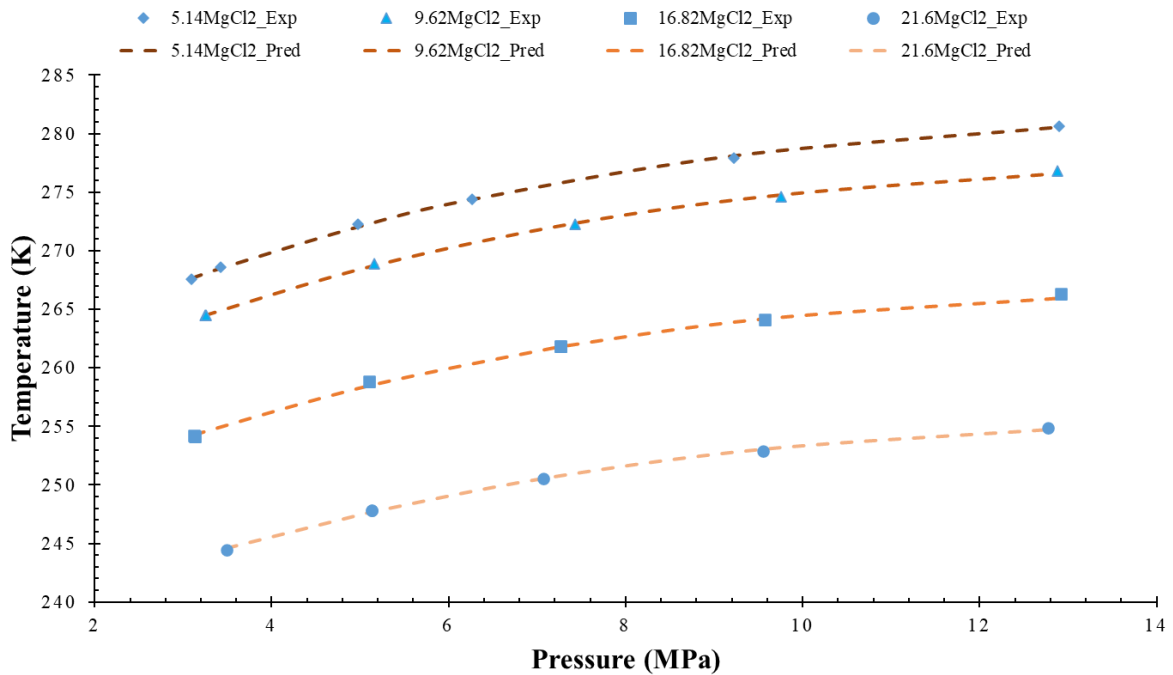


Figure 24. SVR P-T Equilibrium Curve Comparison plots at 10 wt% MeOH

4.4.5 P-T Comparison at 20 wt% MeOH

Figure 25 presents the results for MgCl_2 of concentrations 5.13, 9.62, and 16.8 wt% at 20 wt% methanol. At this maximum methanol concentration, the temperature range of equilibrium shrinks, with the minimum equilibrium temperatures reaching 244 K at 3.5 MPa in case of the 16.8 wt% MgCl_2 content. As one can observe, the SVR model successfully describes the reduction in the number of possible MgCl_2 concentrations at 20 wt% MeOH. This result

demonstrates that the model accounts for the impossibility of using higher concentrations of MgCl_2 together with 20 wt% MeOH within the boundaries of the training set. The calculated curves at 20 wt% MeOH exhibit greater dispersion around experimental points than those at lower concentrations of MeOH. In particular, for the 16.8 wt% MgCl_2 curve, the estimation error amounts to about 3–5 K in the pressure range from 3.0 to 4.0 MPa. This outcome may be explained by the difficulty of predicting extreme cases for both inhibitor concentrations. Nevertheless, the thermodynamic behavior of all calculated curves remains correct.

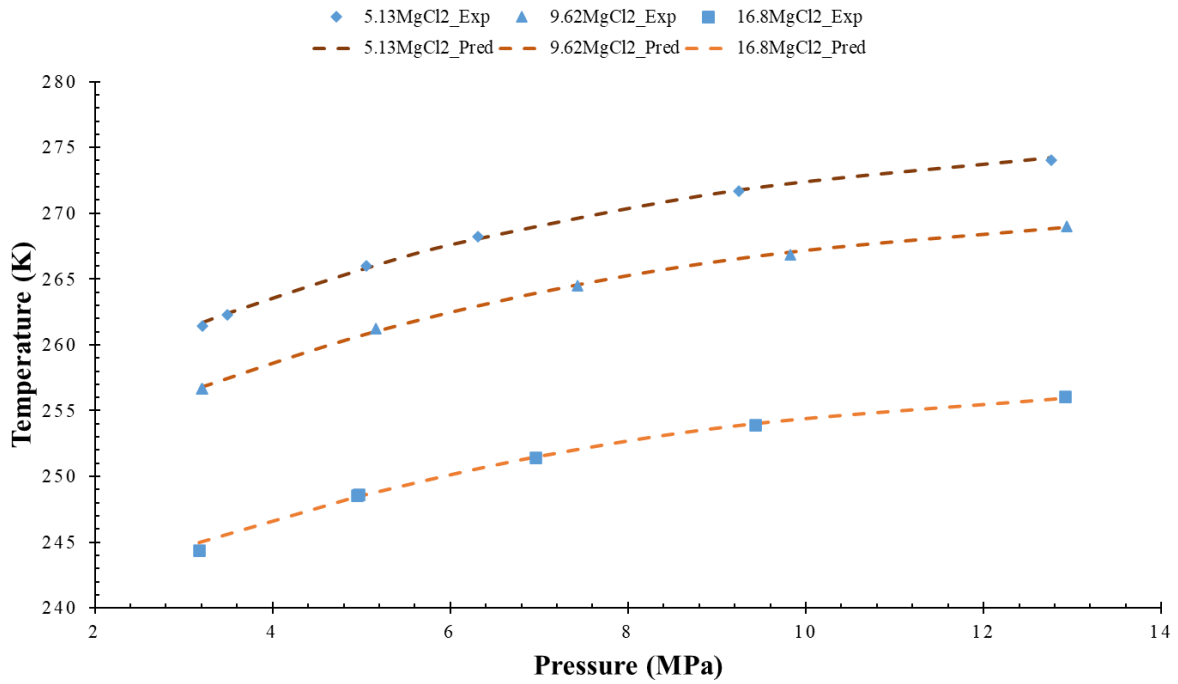


Figure 25. SVR P-T Equilibrium Curve Comparison plots at 20 wt% MeOH

In all cases where methanol concentrations were varied, from those in Figures 22 through to 25, the SVR approach is consistently capable of: (i) replicating the positive P-T slope for all equilibrium lines; (ii) ordering the different lines according to the MgCl_2 concentration; (iii) showing a correct reduction in temperature with higher methanol concentration; and (iv) depicting the additional cooling effect of both salt and organic inhibitors. In other words, the SVR algorithm has not only captured the physical meaning of inhibited hydrate thermodynamics but also trained its behavior at different temperatures.

4.5 Supervised Interpolation Analysis

4.5.1 Interpolation Approach

The supervised interpolation study was intended to determine the ability of the SVR to predict the physically realistic hydrate equilibrium conditions for inhibitor concentrations not represented in the training set, but falling within the span of the training data. The experimental data used for this study was the Semenov et al. (2024) $\text{CH}_4 + \text{MgCl}_2 + \text{MeOH}$ dataset. The training set was augmented by introducing four novel concentrations of MgCl_2 (6.775, 12.615, 19.54, and 24.48 wt%) to the existing experimental concentrations of 5.12, 8.43, 16.8, 22.28, and 26.68 wt%. All other variables remained constant for each case, matching the experimental values for pressure, gas composition, and methanol concentration. The study was conducted at four different methanol concentrations: 0, 5, 10, and 20 wt%.

The physical expectations for this test include an interpolated curve between two training curves at any of the new concentrations being able to predict an equilibrium temperature between the two training curves.

4.5.2 Interpolation Results at 0 wt% MeOH

The P-T curves for both the original five concentrations of MgCl_2 and the interpolated four concentrations of MgCl_2 are shown in Fig. 26. The orange dashed curve for 6.775 wt% MgCl_2 lies between the light blue solid curve for 5.12 wt% MgCl_2 and the grey solid curve for 8.43 wt%, predicting temperatures at 3 MPa of approximately 272–273 K between the 273 K and 272 K of the adjacent solid curves. The dark blue dashed curve for 12.615 wt% MgCl_2 lies between the grey solid curve for 8.43 wt% and the green solid curve for 16.8 wt%, which was expected since 12.615 wt% was between 8.43 wt% and 16.8 wt%. The green dashed curve for 19.54 wt% MgCl_2 lies between the green solid curve for 16.8 wt% and the yellow solid curve for 22.28 wt%, while the light blue dashed curve for 24.48 wt% MgCl_2 lies between the yellow solid curve for 22.28 wt% and the purple solid curve for 26.68 wt%. In all cases, the interpolated curves are smooth, have the correct positive P-T slope, and lie in the physically expected location within the established range of solid curves.

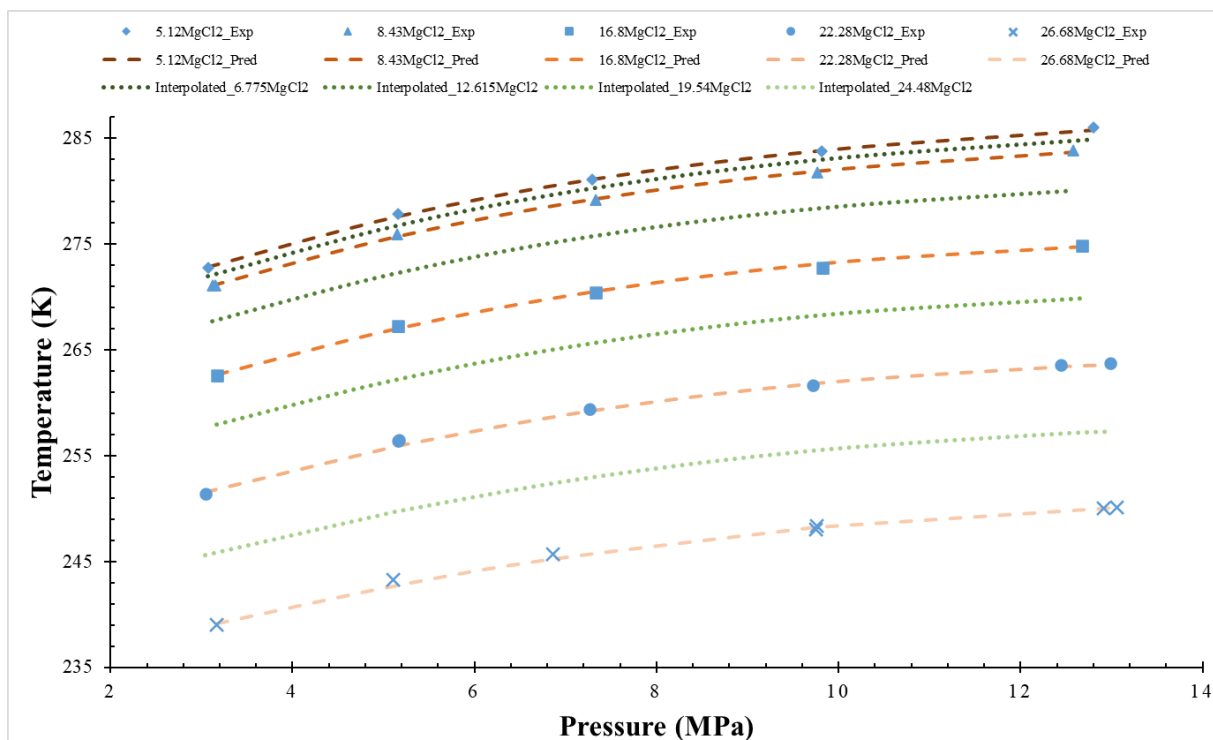


Figure 26. Interpolation P-T plots at 0 wt% MeOH

4.5.3 Interpolation Results at 5, 10, and 20 wt% MeOH

Equivalent interpolations at increased concentrations of methanol are shown in Figures 27, 28, and 29. For 5 wt% MeOH (Figure 4.10), all four interpolated concentrations of MgCl_2 are within the correct boundaries relative to the surrounding training curve over the whole range of pressures, with consistent, monotonic behavior and no crossing anomalies. Figure 4.11 shows that interpolation is correct, including the interpolated concentration of 24.48 wt% MgCl_2 (brown dashed line) which stays appropriately below the training curve of 22.28 wt% and converges towards but does not intersect the upper bound of 26.68 wt%. In this plot, minimum temperatures converge to around 236 K at 3.5 MPa, representing the lowest limit of the training range.

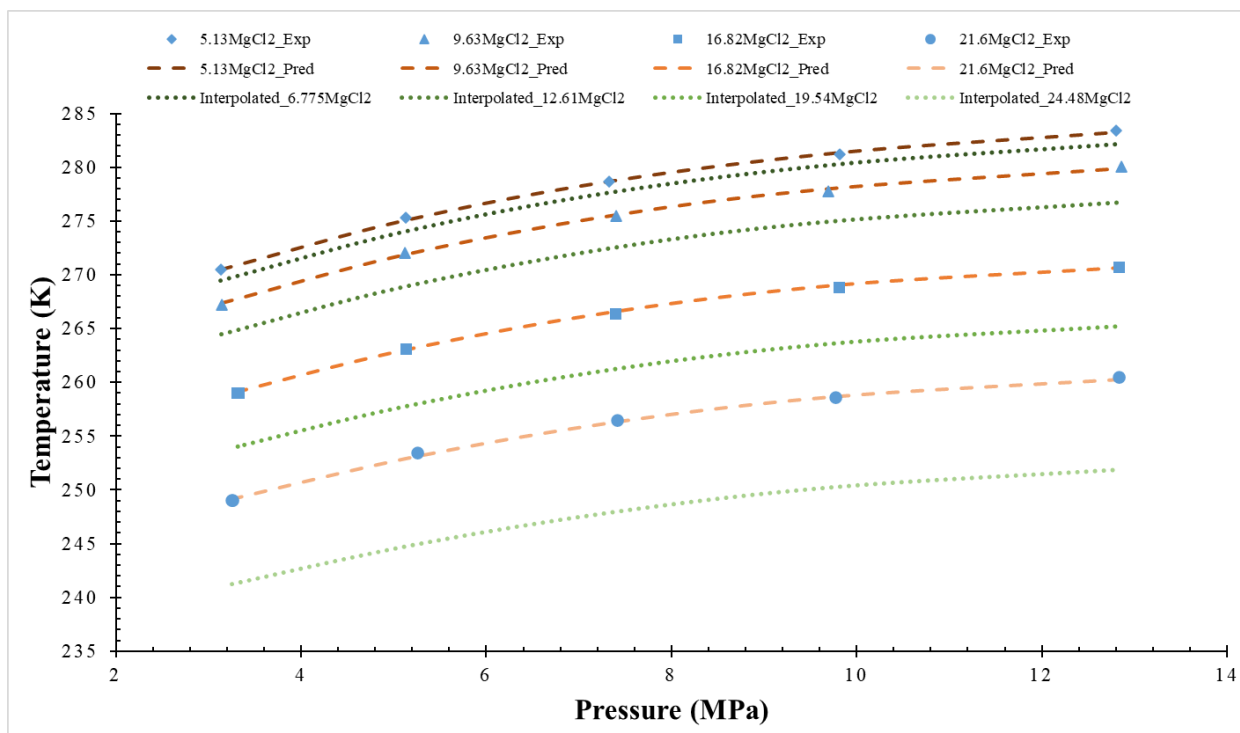


Figure 27. Interpolation P-T plots at 5 wt% MeOH

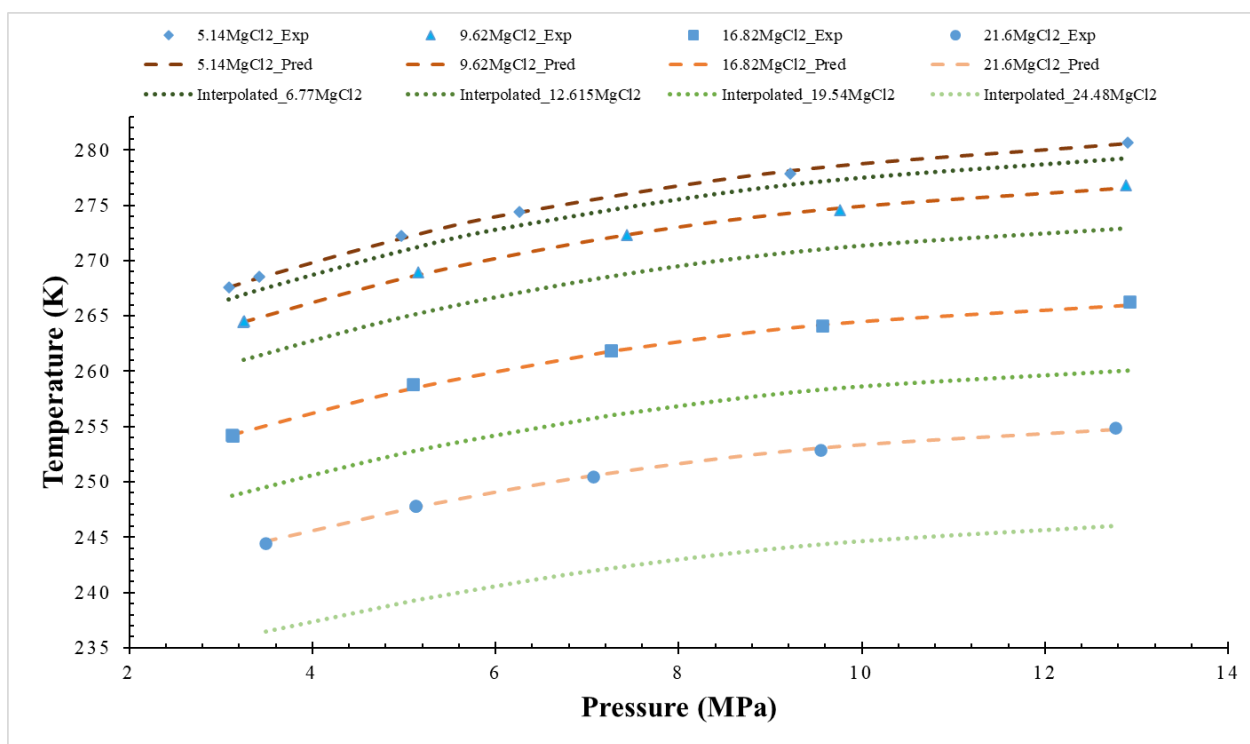


Figure 28. Interpolation P-T plots at 10 wt% MeOH

With a MeOH content of 20 wt%, the data points clearly have less concentration variation for MgCl_2 owing to coverage problems at the highest combination of inhibitive agent concentrations, but once again the interpolation functions (at 6.775, 12.615, and 19.54 wt%) MgCl_2 fit nicely within their neighboring training curves. The interpolated function at 19.54

wt% MgCl_2 is appropriately below that of the 16.8 wt% MgCl_2 training function and approaches about 239 K at 3.5 MPa.

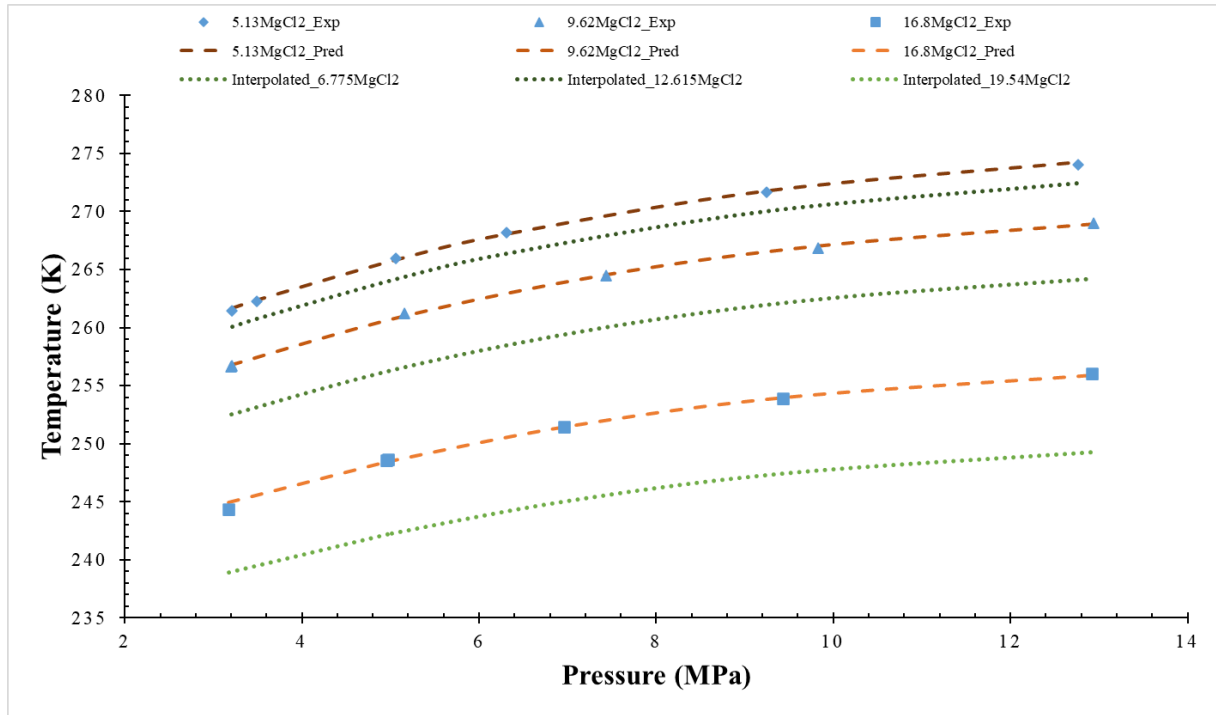


Figure 29. Interpolation P-T plots at 20 wt% MeOH

Overall, considering the interpolation of all four levels of MeOH together, it is clear that the SVR algorithm has effectively learned a smooth physical mapping function of the feature MgCl_2 on the hydrate formation temperature. Moreover, the SVR algorithm is able to interpolate not only in the inhibitor concentration feature domain but also to handle the concurrent presence of another inhibitor (i.e., MeOH). Importantly, this can be accomplished while allowing no values of new inhibitor concentrations within the training data.

4.6 Temperature Depression Trend Analysis

Figure 30 shows the effect of ΔT (difference between the temperature at which the pure methane hydrate is in equilibrium and the inhibited one at the same pressure level, calculated around 7 MPa) on methanol mole fraction (0, 5, 10, and 20 wt%) plotted against different concentrations of MgCl_2 (5.12, 9.63, 16.8, and 21.6 wt%). Experimental data for ΔT and SVR-predicted ΔT data are represented by symbols linked with straight lines and dashed lines, respectively.

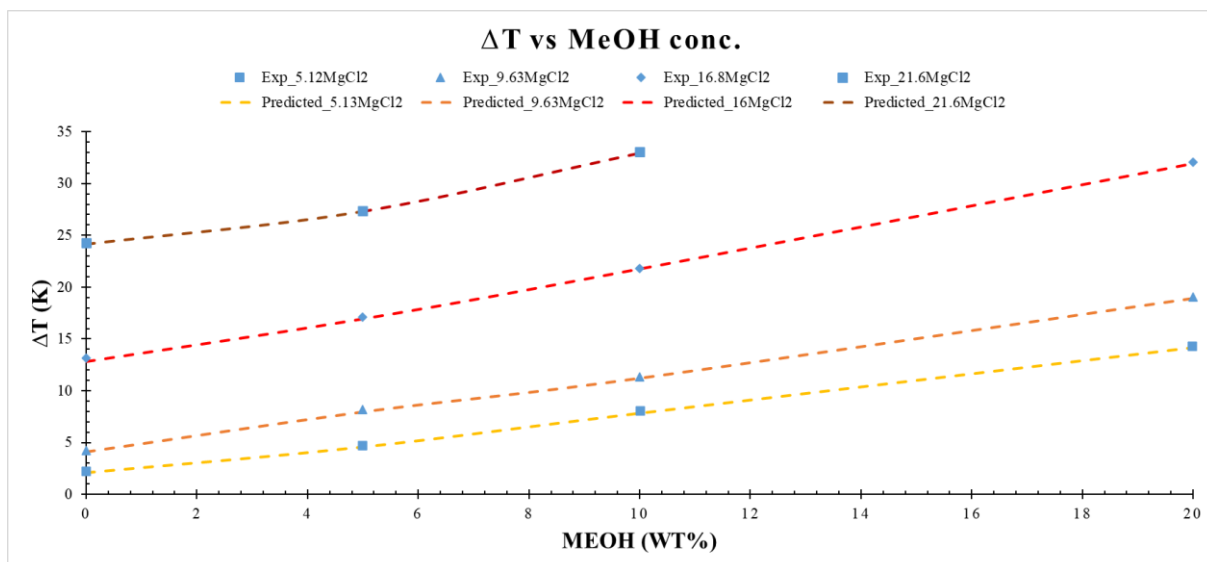


Figure 30. ΔT vs MeOH concentration trend analysis

Physically important points of this figure include the following. First, SVR is able to reproduce the monotonous dependence of ΔT with the increase of methanol concentration for all four MgCl_2 concentrations considered. In particular, an approximate linear increase of ΔT when increasing the concentration of methanol in its range from 0 to 20 wt% takes place in all four cases, starting from a specific initial value depending only on the inhibition due to MgCl_2 (when $\text{MeOH} = 0$ wt%) to a value of suppression when the concentration of methanol reaches 20 wt%. It should be noted that such behavior of ΔT is physically reasonable and corresponds to the typical behavior characteristic of organic inhibitors of the Hammerschmidt type. Second, the value of ΔT at zero methanol content increases with MgCl_2 concentration, which physically reflects the more significant inhibition provided by MgCl_2 at higher concentrations. Specifically, when there is no methanol ($\text{MeOH} = 0$ wt%), the value of ΔT changes from about 2 K for MgCl_2 equal to 5.12 wt% to 24.4 K at 21.6 wt% MgCl_2 – thus, the difference between these two values reaches more than 22 K solely because of the difference in the concentration of magnesium chloride. Finally, SVR provides a very good fit of predicted ΔT to experimental values regardless of the concentration of either MgCl_2 or methanol. Indeed, for 5.12 and 9.63 wt% MgCl_2 , predicted and experimental dependencies of ΔT are indistinguishable visually, while for 16.8 and 21.6 wt% MgCl_2 , the predictions are close to the experimental curve, even though some deviations can be observed at intermediate methanol concentrations. Moreover, in all these cases, the direction of trends was correctly reproduced by SVR. Most importantly, however, the additivity of the effect of two inhibitors on ΔT was successfully modeled without imposing any specific requirement of that kind during training.

4.7 SHAP Feature Importance Analysis

4.7.1 Global Feature Importance Ranking

In Figure 31, there is the list of global feature importances for the SVR model in terms of the mean absolute SHAP value of each of the 18 features in the input space based on the entire dataset of 922 data points. The more important features are ranked first. Physical insight can be gained from the findings about the impact of the features on the output.

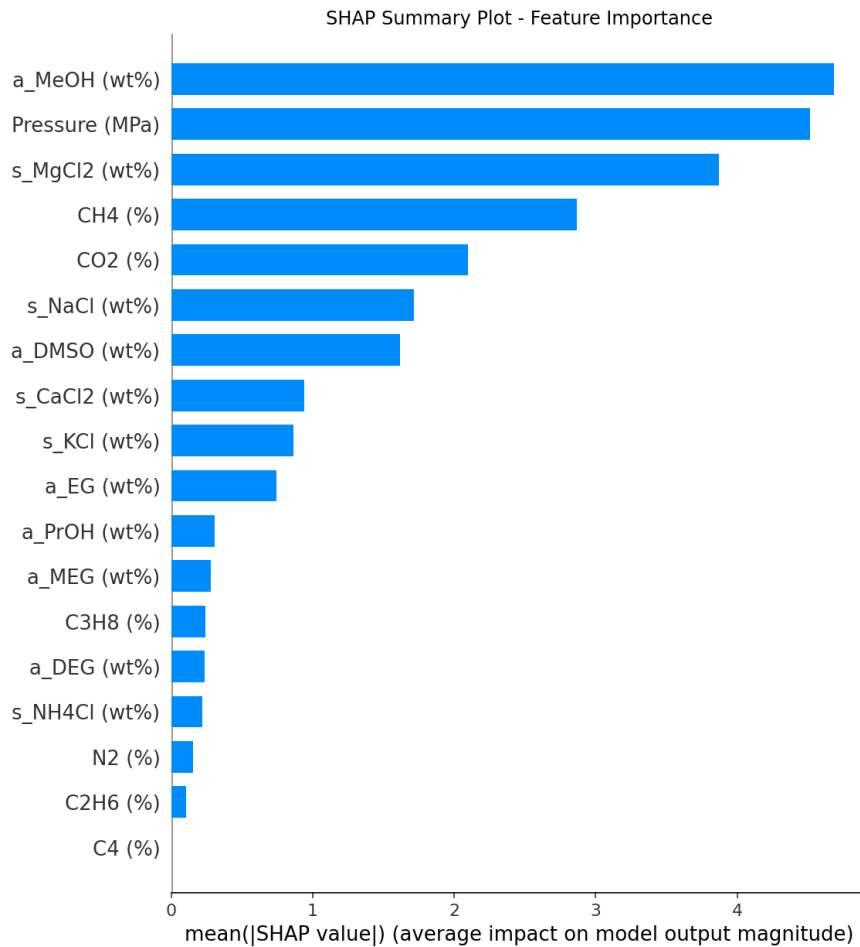


Figure 31. SHAP Summary Plot: Feature Importance (bar chart)

Concentration of methanol (a_MeOH, wt%) is the dominant variable with an average SHAP value of approximately 4.6 K. The predominance of the methanol influence on temperature is physically justified by the fact that the number of measurements with the addition of methanol in the experimental setup exceeds any other inhibitor and includes the largest possible range of its concentrations (0-50 wt%). It leads to a decrease in the temperature by tens of kelvins and thus becomes the leading component in the variance of T.

Secondly, we can mention **pressure (MPa)** that has the mean absolute SHAP value of around 4.5 K. Physically, it corresponds to the fact that the positive P-T slope of the equilibrium curve makes pressure variation in the interval from 0.23 to 73.28 MPa lead to the shift of temperature in the range of approximately 50 K for the constant value of inhibitors. Therefore, pressure is the key factor that influences the SVR results.

Thirdly, we need to consider **MgCl₂ concentration (s_MgCl₂, wt%)** with a mean SHAP value of approximately 3.9 K. Although the number of samples with the addition of MgCl₂ is much smaller than the number of samples with the addition of methanol, the contribution of this variable is large because it has a very powerful inhibition effect per one percent of weight concentration due to high charge density of Mg²⁺ ions.

Fourth, we have the **methane mole fraction (CH₄, %)**, which is the fourth in the list with a mean absolute SHAP value of approximately 2.9 K. This result is associated with the fact that the structure of hydrates depends on the composition of the gas mixture and, thus, the methane mole fraction affects the value of the equilibrium temperature.

Fifth, **CO₂ mole fraction (CO₂, %)** with the SHAP value of approximately 2.1 K corresponds to the relatively strong suppression effect of CO₂ on the hydrate formation process compared to CH₄. Therefore, even a slight addition of this gas molecule to the gas mixture causes the change of the equilibrium temperature.

Sixth is the **concentration of NaCl (s_NaCl, wt%, ≈1.7 K)**, and seventh - **DMSO (a_DMSO, wt%, ≈1.6 K)**. Note that the physical interpretation is valid here - NaCl is less effective in the prevention of hydrate formation than MgCl₂ because of the difference in charge-to-radius ratio. However, the high concentration range of DMSO used by authors of Semenov et al. papers made its effect much higher than for other glycols and salt hydrate inhibitors.

The rest of variables - CaCl₂, KCl, EG, PrOH, MEG, C₃H₈, DEG, NH₄Cl, N₂, C₂H₆, and C₄H₁₀ - possess a relatively low mean absolute SHAP values (less than 1.0 K). Their low impact on the model predictions is associated with small numbers of the corresponding samples and weak absolute effects. The zero SHAP values for the concentration of C₄H₁₀ (%) also can be easily justified by the lack of this substance in data points.

4.7.2 Feature Effect Directionality

The beeswarm plot in Figure 32 illustrates the SHAP values of each attribute, whereby the dots depict individual data, while their horizontal position denotes the SHAP value (effect on the model's prediction in K), and the color shows the attribute value (red is high; blue is low).

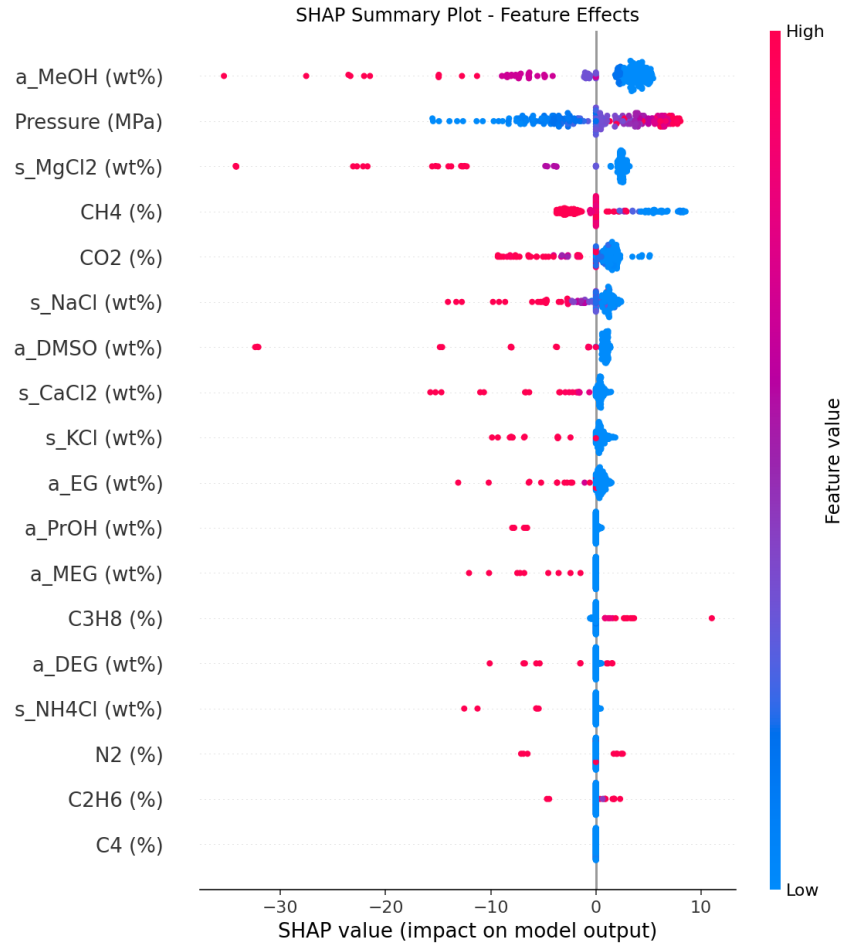


Figure 32. SHAP Summary Plot: Feature Effects (beeswarm)

4.8 Summary of Results

The SHAP values for a_MeOH (wt%) are mainly negative (with high feature values in red extending to the left towards approximately -35 K), confirming the strong inhibiting effect of high methanol concentration on the formation temperature - as expected. Conversely, low MeOH concentrations (in blue) have close to zero SHAP values, as expected. Pressure (MPa) has mainly positive SHAP values, confirming that increasing pressure raises the formation temperature (with high pressures in red reaching around +8 K). Thus, this feature

properly represents the positive slope of the P-T line for hydrate equilibrium. For s_MgCl_2 (wt%), high feature values in red yield mainly negative SHAP values (the spread reaches up to -30 K for the highest values). This confirms the inhibiting effect of $MgCl_2$ on the formation temperature, which is quite strong for high concentrations. For CH_4 (%), high methane fractions result in SHAP values that are slightly positive (with red feature values), while low methane fractions result in negative SHAP values (blue). The reason is related to the thermodynamics of the datasets: High CH_4 fractions usually imply the dominance of CH_4 hydrates (thus higher formation temperature), while low CH_4 usually implies either CO_2 or other components in the hydrate formation system. The sign and the extent of influence of each feature on the target parameter according to SHAP analysis have clear thermodynamic interpretation and consistency with known physical/chemical behavior of gas hydrate inhibition. This constitutes strong evidence for the validity of learned dependencies by SVR and their non-statistical nature.

5. DISCUSSION

5.1 Overview

The findings are interpreted and contextualised in the current chapter. These will be examined with respect to five themes, which include: (1) the performance of the machine learning (ML) pipeline compared to conventional thermodynamic methods in terms of statistics and physics; (2) the rationale behind the selection of SVR as the optimal model without having the minimum test root mean squared error (RMSE); (3) the physical interpretation of SHAP values; (4) trends and interpolations in view of their thermodynamic implications; and (5) the scope of this study and future directions.

5.2 ML Pipeline Performance in the Context of Classical Thermodynamic Approaches

5.2.1 Comparison Framework

An appropriate assessment of the accuracies of the developed machine learning models must take into account the performance of other available methods for estimating hydrate equilibrium conditions, whether traditional thermodynamic models or other ML works in the past. It should be emphasized, however, that a quantitative comparison based on the use of the same test set is impossible in the present study since none has conducted the exact same experiment using the exact same type of consistency screened data set. There are well-known benchmark accuracies of thermodynamic models which can serve as a good basis of comparison.

5.2.2 Classical Thermodynamic Model Accuracy on Inhibited Systems

The accuracy that can be attained using state-of-the-art thermodynamic models for pure gas hydrates without inhibition can be impressively high. For example, according to Rasoolzadeh et al. (2025), a state-of-the-art thermodynamic model based on the vdW-P model for gas hydrates with inhibitors has an average absolute deviation of only 0.25 K over 1,879 consistent data points over the pressure range 0.093–493 MPa for 7 pure gas hydrate-forming substances (Rasoolzadeh et al., 2025). Commercial thermodynamic software like CSMGem gives average absolute temperature deviations in the order of 0.2–0.4 K for natural gas hydrate systems containing inhibitors over 1,685 consistent data points (Koh et al., 2011).

However, the quality of prediction using thermodynamic models is compromised to a great extent in the presence of inhibitors, especially if the type of inhibitor present belongs to those represented in the dataset. The vdW-P model requires proper selection of activity coefficient models for each inhibitor type, with the use of two distinct kinds of electrolyte models necessary in the case of salt inhibition compared to organic inhibition, and very poor prediction in the case of multiple inhibitors. As highlighted in Cai et al. (2025), while vdW-P-based models give fairly accurate predictions in the broadest possible pressure and inhibitor concentration range, provided an advanced activity coefficient model is applied, their quality is significantly reduced for highly concentrated inhibitors, mixed inhibitors, and multicomponent gases. In addition, as pointed out in Kashish et al. (2024), special attention should be given prior to applying thermodynamic models with high concentrations of one or several inhibitors, because the presence of water-gas-alcohol, water-gas-glycol, or water-gas-electrolyte makes these systems non-ideal, hence specific parameterisations for each system are needed. The ML method applied in this work eliminates the need for these parameters by learning their effect on the equilibrium point of interest directly from experimental data, but in return, it cannot extrapolate outside the interpolation boundaries of the training dataset.

In this context, the SVR model developed in this thesis reaches a test set RMSE of 0.963 K and MAPE of 0.193% on 185 consistent data points involving 11 inhibitor types, 6 different gas component(s), pressures up to 73.28 MPa, and temperatures as low as 234.03 K. Three of the best performing boosting models (CatBoost, XGBoost, GBR) achieve test RMSE in the range of 0.931–0.990 K. These results are similar to or even better than those reported for classical thermodynamic models for equally complicated inhibited systems. Notably, unlike the thermodynamic models discussed above, the ML models do not require any specific parameterisations, equation of state and activity coefficient model couplings. However, it must be stressed that, while thermodynamic models could theoretically extrapolate outside the calibration range, the ML models developed in this thesis operate only in the bounds of the training data. Nonetheless, as the domain of the training data covers the majority of industrially relevant flow assurance conditions, the ML models are competitive or even superior compared to contemporary thermodynamic modelling methods.

5.3 SVR Model Selection: Physical Generalisation as the Primary Criterion

5.3.1 Statistical Metrics Versus Physical Reliability

Choosing SVR as the best model even though CatBoost had slightly better test RMSE (0.931K vs. 0.963K) calls for justification. It would be safe to state that for any prediction tool that can be used as part of flow assurance process, having accurate results on a hold-out dataset is a must-have feature, although it is not enough for choosing the best model. Another feature, which is no less important than accuracy, is being able to provide physically realistic predictions for new, unseen conditions, where there is no information to train the algorithm on.

5.3.2 Engineering Context for Interpolation Capability

During the course of field operations, the concentration of the inhibitors used in a pipeline system is usually not precisely identical to any of the specific values that have been tested through experimentation. Suppose, for instance, that an operator injects an inhibitor system composed of MgCl_2 and methanol. Then, the concentration of the mixture in the aqueous phase will fall somewhere in between the range of the training data. It follows that the machine learning model should be able to make accurate interpolations in such a way that it can provide predictions of formation temperatures at these concentrations.

5.3.3 SVR Interpolation Behavior

As shown in the supervised interpolation analysis provided in Section 4.5, SVR unequivocally interpolates smooth, monotonically ordered, and physically realistic predictions of P-T curves in conditions of MgCl_2 concentrations absent from the training data set — for all four methanol concentrations analyzed (0, 5, 10, and 20 wt%). In all cases, the predicted curves are found in the correct positions between neighboring curves corresponding to MgCl_2 concentrations found in the training set. No spurious intersections or non-monotonic behavior appears across the full pressure range of 3-13 MPa. Such behavior follows from one of the intrinsic mathematical properties of the RBF kernel employed by SVR, namely, the property of being a mapping of the input feature space to an n-dimensional feature space where the regression function is the smoothest function fitting the training data within the ϵ -tube (Smola and Schölkopf, 2004). CatBoost, however, uses symmetric decision trees as the base learners and thereby splits the

input feature space into rectangles. The resulting regression function is highly accurate for test inputs taken from the same distribution as the training set.

Nevertheless, such an approach may yield abrupt changes or even physically inconsistent results when inputs belong to those regions in the input feature space where no training inputs occur (a well-known weakness of tree ensembles for interpolation). The interpolation test results for CatBoost illustrate the phenomenon mentioned above and serve as an additional reason to choose SVR despite its slightly higher test RMSE.

One should note that SVR achieves excellent absolute engineering accuracy (RMSE of only 0.963 K or 0.395% of the temperature range in the test set) and that this result is directly comparable to average absolute deviation reported in the literature for thermodynamic models in inhibited conditions. Furthermore, after the hyperparameters tuning, SVR reaches the test R^2 value of 0.989, which means that, statistically, it performs equally to CatBoost and GBR. However, the small difference (0.032 K) in test RMSE cannot be considered significant from the engineering standpoint.

5.4 Analysis of Overfitting Patterns Across Models

5.4.1 Random Forest Overfitting

The most severe case of overfitting is seen in Random Forest; while its training RMSE is 0.020 K (which means it almost memorizes the training data down to the precision of our experiment), its test RMSE is 1.651 K, giving a ratio of around 83:1. In fact, the training AIC of -5257.9 is by far the most negative among all models, suggesting that RF is fitting a highly complex function, which matches the training data perfectly but does not generalize well. This is a well-known theoretical problem: trees in a random forest are fully grown and unpruned, and with 922 training examples and 18 input variables, the model has enough flexibility to fit such a complex decision tree that achieves zero training error. While bagging can help to reduce variance when compared to a single tree, it has been mathematically proved that the error rate of a random forest will always tend to converge with increasing numbers of trees. However, this only holds true for generalization error, and RF minimizes training error to zero without any upper bound. This, together with training R^2 of 0.999 and test R^2 of 0.967, implies that the model has learned the noise pattern in the training data set rather than the underlying physics.

5.4.2 Decision Tree Overfitting and Instability

Overfitting for the Decision Tree algorithm is shown to have a distinct character: although both its training and testing R^2 scores (0.997 and 0.902, respectively) show greater consistency compared to RF, the former's test score is the lowest among the six - RMSE = 2.859 K and MAPE = 0.792%. Decision trees are notoriously prone to generating highly irregular prediction functions, whereby small variations in the input data can dramatically change the tree's shape, and this tendency is reflected in the model's weak generalization ability on the testing set.

5.4.3 Gradient Boosting Family Balance

All three gradient boosting models - CatBoost, XGBoost, and GBR - outperform both RF and DT in terms of the balance between training and testing performance; all of them show test RMSE between 0.931–0.990 K, while the training RMSEs range between 0.394–0.447 K. The ratio between the training and testing RMSE is approximately 2–4, which is much more favourable compared to RF (~83) and DT (~5.5), suggesting very well-regularised models that learned a representative structure from training data. It is important to note that the specific regularisation strategies included in XGBoost (L2 penalty on leaf weights, γ minimum gain threshold) and CatBoost (ordered boosting, L2 leaf regularisation) help these models maintain such good balance, as compared to RF. Another machine learning model, SVR - that belongs to a completely different category, demonstrates a similar training-testing balance; it shows the training RMSE of 0.471 K and test RMSE of 0.963 K (training-testing RMSE ratio is 2.0).

5.5 Physical Interpretation of SHAP Feature Importances

5.5.1 Dominance of Methanol and Pressure

The SHAP analysis reveals that a_{MeOH} and pressure were the two key features used in predicting the hydrate formation temperature, having respective mean absolute SHAP contributions of 4.6 K and 4.5 K. These two factors have nearly equal mean absolute SHAP values and this is not surprising considering their physical significance and thermodynamic grounding since the two factors produce a large, consistent effect on the formation temperature throughout the entire range of conditions included in the database.

The fact that the methanol concentration was a more significant factor compared to all other organic inhibitors is primarily due to the wide range of concentrations available in the training

set. Specifically, the concentrations of MeOH vary widely in comparison to all other inhibitors (0–50 wt%), come from the highest number of source studies, and represent a wide range of gas compositions and pressures. Therefore, the temperature variance arising from variation in methanol concentration becomes the primary source of target variability and this is properly identified by the SVR model. It must be noted here that while the concentration of methanol is important, it does not mean that methanol is a more thermodynamically potent inhibitor than all others; indeed, the temperature depression induced by MgCl_2 exceeds that of MeOH per weight percentage addition (1.3 K/wt% MgCl_2 vs 0.5 K/wt% MeOH).

5.5.2 MgCl_2 Ranking Above NaCl: Physical Interpretation

The third feature ranked based on its mean |SHAP| value is MgCl_2 concentration (mean |SHAP| ≈ 3.9 K), which ranks considerably higher than NaCl concentration (mean |SHAP| ≈ 1.7 K). The ranking order between the two cations is physically correct since divalent cations tend to be stronger inhibitors than their monovalent counterparts. The inhibition efficiencies of different cations follow the order $\text{Mg}^{2+} > \text{Ca}^{2+} > \text{Na}^+ > \text{K}^+$ (Wang et al., 2019). This is because the mechanism behind inorganic electrolyte inhibition relies on reducing water activity through the electrostatic binding of water molecules in the solvation shell surrounding the ions. Since Mg^{2+} ions have a higher charge density (charge +2, small ionic radius), they bind more tightly and over a larger solvation shell compared to Na^+ ions, resulting in a larger reduction in water activity per unit weight concentration of added cations. Consequently, the average SHAP values of MgCl_2 concentration are higher (by a factor of 2.3) than those of NaCl concentration (3.9 K vs. 1.7 K), consistent with the greater magnitude of temperature depressions observed for MgCl_2 compared to NaCl at equal weight concentrations.

5.5.3 Low Importance of C_2H_6 , C_4 , and N_2

The near-zero SHAP values for C_2H_6 (%), C_4 (%), and N_2 (%) can be attributed to the underrepresentation of these components in the training set. The compiled database mainly includes experiments performed on methane-rich gas mixtures, with hydrocarbons such as C_2H_6 , C_3H_8 , and C_4 components present in a subset of the multicomponent gas mixture datasets. The minimal contribution of these features to the predicted temperature depressions does not indicate that C_2H_6 , C_3H_8 , and C_4 components are not important thermodynamically in affecting hydrate equilibrium; they can significantly impact hydrate structures (e.g., changing from sI to sII) and formation pressures for equivalent temperatures (Sloan, 2003; Koh et al., 2011). The

reason for their minimal contribution here is that they are currently underrepresented in the dataset, i.e., the compiled database consists of experimental data mainly obtained for methane-rich gas mixtures.

5.6 Physical Validity of Trend and Interpolation Results

5.6.1 *P-T Curve Fidelity Across MeOH Concentrations*

Comparisons between predicted and measured P-T equilibrium curves were conducted at various MeOH concentrations (0%, 5%, 10%, and 20%). The results show that the SVR model is capable of accurately capturing the fundamental properties of the thermodynamics of inhibited hydrate formation: positive P-T slopes, monotonic decrease in equilibrium temperatures with increased inhibitor concentrations, and additive temperature depression effects from co-existing salt and organic inhibitors. The observed deviations from experimental data points (mostly less than 2–3 K) are not systematic and do not display any directional trends at all MeOH levels.

However, the slightly greater deviations found at high MgCl_2 concentrations (e.g., 16.8 wt% and 22.28 wt% at 5 wt% MeOH and 26.68 wt% at 0 wt% MeOH) can be attributed to the sparseness of data points for training at extreme inhibitor concentrations. With fewer experimental data points available in the 30 source papers in the literature for the combination of high MgCl_2 concentrations and high MeOH levels, the SVR model predicts temperature depressions using interpolation within a feature space that is sparse in data. This is one limitation inherent to purely data-driven models: predictions become less accurate away from training data points, and the effect is governed by the bandwidth parameter γ of the RBF kernel. The deviations in these extreme cases do not undermine the accuracy of the SVR model for the vast majority of practical applications where the inhibitor concentrations are in the moderate range of 5–30 wt%.

5.6.2 *ΔT Trend Analysis and Additivity of Inhibition*

ΔT vs. concentration of MeOH curve is perhaps the best validation based on a fundamental consideration of the SVR model from the thermodynamic perspective. Indeed, as was stated earlier in this text, the prediction about the linear relationship of ΔT with MeOH concentration at constant MgCl_2 is quite natural and can be explained by thermodynamic theory. For moderate concentrations of the inhibitor within the limited range of temperatures used in this dataset

(~10-20 K depending on the inhibitor concentration), the relationship $\Delta T/(T_o T) \approx \beta \ln a_{vv}(x_i)$ can be considered linear. At the same time, as the log of water activity (which appears on the right-hand side) is approximately linear in the concentration of the solution, the SVR model shows excellent predictions without explicitly imposing this constraint on it. The values of ΔT at 0 wt% MeOH increase with the increasing concentration of $MgCl_2$ in agreement with the theoretical expectations. It also demonstrates that the contributions of each component of a mixture are correctly accounted for. In addition, as was mentioned above, the additive contribution of both inhibitors in a mixed solution results in the additive increase in the ΔT values; however, the somewhat larger deviations at 21.6 wt% $MgCl_2$ indicate less precise interpolation of this additive relation into the model.

5.7 Limitations of This Study

5.7.1 Dataset Size and Laboratory Data Coverage

The ultimate clean dataset of 922 points thermodynamically consistent, although thoroughly curated and much larger compared to most previous machine learning investigations in hydrates domain, is relatively small considering the 18-dimensional feature space and the complexity of inhibited hydrate system itself. Notably, all 922 data points used were derived exclusively through rigorous laboratory experiments, where single or binary gas mixtures along with single or binary inhibitors were studied under specified experimental conditions. In comparison to natural gas field compositions, the range of the composition of C_2+ gases (including C_2H_6 , C_3H_8 , $i-C_4H_{10}$, $n-C_4H_{10}$, as well as higher-order components) is considerably narrower. The extremely low SHAP values attributed to C_2H_6 , C_4H_{10} , N_2 features are explained by that fact directly: the machine learning model does not have enough training data representing the concentration of said components similar to those found in real field gas compositions. Therefore, predictions for natural gas compositions enriched with C_2+ hydrocarbons should be taken with due caution, and dataset extension into the area represents the important future direction of research.

5.7.2 Interpolation Within the Training Domain Only

It can be seen from the supervised interpolation study that the SVR model delivers valid and physically plausible predictions on inhibitor concentrations lying inside the span of the training data, which constitutes an important ability. Nevertheless, extrapolation of model predictions outside the training data domain limits was not performed, and extrapolation is not advised. The span of the pressure in the training dataset ranges between 0.23 and 73.28 MPa, whereas

for the temperature parameter it covers the interval of 234.03–291.4 K. Predictions at extreme values of the pressure and/or temperature parameters that lie outside these intervals, especially in case of extremely high pressure typical of ultra-deep-water operations or low temperatures characteristic of permafrost formations, cannot be considered valid since they lie out of the range of the provided training data. Furthermore, predictions on inhibitor concentrations outside the range of the training dataset, specifically MeOH concentration exceeding 50 wt% or NaCl concentration surpassing 26 wt%, cannot be obtained.

5.7.3 Scope: THI Systems Only, Current Gas Composition Coverage

This predictive pipeline can currently only be applied to gas hydrate formation in the presence of THIs, since KHIs and AAs are not present in the training data, and their influence on gas hydrate formation parameters is both mechanism-dependent and highly concentration-dependent. Thus, application of the current model to KHI systems requires not just the representation of inhibitors in a new feature space but possibly even a change in the output variable (for example, subcooling temperature or induction time) and a significantly larger quantity of training data. In addition, the current model only applies to gas hydrate formation without a separate hydrocarbon liquid phase; thus, multiphase systems containing liquid hydrocarbons are beyond the current scope of applicability.

As concerns the gas composition in the system, the current model can be used solely to estimate the temperature at which hydrate forms for systems with gas compositions in the six-component space (CH_4 , C_2H_6 , C_3H_8 , C_4 , N_2 , CO_2) covered by the training data set. In principle, generalization to gas compositions outside of this space (or other thermodynamic properties of the system) is feasible, provided sufficient thermodynamically consistent training data for such systems can be gathered. The consistency-screening stage, in particular, does not depend on the gas composition (since it makes use of the Clausius-Clapeyron and water activity criteria), and the multi-column feature representation makes generalization to additional components straightforward without modification of the model structure.

5.8 Significance of the Thermodynamic Consistency Screening Step

The common thread throughout the findings and analysis section of this thesis is the critical importance of the consistency screening process as a necessary first step before training ML models. The reduction of 1,751 data points down to 1,167 (or 33.4%) due to consistency screening was not simply an aesthetic choice, but an important quality control measure. The

case study provided by Rasoolzadeh et al. (2025) for unconstrained systems shows the significance of this point clearly.

None of the earlier studies that employed ML techniques for hydrate predictions, as surveyed in Section 2.5, applied a thermodynamic consistency screening process prior to training the ML models. In this thesis, with all six trained models providing test R^2 values greater than 0.90 and with the three best-performing models delivering test RMSE values less than 1.0 K, it is clear that consistency-screening yields reliable and physically plausible ML models. It is not feasible to separate the impacts of this step from the chosen ML algorithms, the particular dataset, and the feature encoding method — in order to do that, a comparative analysis would need to use identical consistent and uncleaned datasets. However, by analyzing the trends and interpolating the SVR model, it becomes clear that consistency screening enabled the learning of thermodynamically consistent feature-temperature relations.

5.9 Directions for Future Work

Based on the results and limitations identified in this thesis, the following directions are proposed for future research:

Extension of the experimental database to real natural gas compositions. The most promising and relevant near future extension of this study would be to gather phase equilibrium data on inhibited hydrate systems with realistic multicomponent gas mixtures — i.e., with ethane (C_2H_6), propane (C_3H_8), and even more complex compounds present, in realistic quantities — e.g., thermogenic gas mixtures with C_2+ content in the range of 5-20%. Such a step would provide direct validation of the problem identified as the reason behind the relatively low SHAP importance of C_2+ components.

Application of the consistency screening pipeline to additional inhibitor types. The thermodynamically consistent approach of Sa et al. (2018) applied in this study is shown in this thesis for systems of THI inhibitor. One potential future research avenue would involve extending the methodological toolset to other classes of inhibitors, such as amino acids, DESs, and ionic liquids, recently considered as a promising alternative for environmentally friendly inhibition by some authors. The screening methodology outlined here is, theoretically, applicable to any inhibitor as long as a reference uninhibited phase equilibria curve is obtainable.

Prediction performance assessment under extrapolation conditions. A study on SVR prediction performance outside the boundaries of the training domain — i.e., at pressures higher than 73 MPa and lower temperatures than 234 K — would require additional experimental data to be collected and would determine the safe extrapolation regime of the SVR model.

Integration with real-time flow assurance monitoring. The fast inference time (in microseconds) of the Support Vector Regression once the training has been completed makes it suitable for implementation in flow assurance monitoring software packages. Potential future research would involve constructing a pipeline monitoring system capable of providing hydrate risk assessments based on pressure, temperature, and inhibitor concentration measurements in the pipe.

6. CONCLUSIONS AND RECOMMENDATIONS

6.1 Summary of the Study

This thesis proposed a data-driven approach for the prediction of gas hydrate formation temperatures in thermodynamic inhibitor-containing systems via supervised machine learning. The research project has been prompted by the known limitations of classical van der Waals–Platteeuw thermodynamic approaches to modeling complex systems (e.g., mixtures of several types of inhibitors, multicomponent gases, high concentration of inhibitors) with respect to accurate prediction of hydrate phase equilibria, as well as the striking lack of data quality control for any existing machine learning-based prediction of hydrates in the literature.

This thesis aims to meet five interrelated objectives: to systematically collect phase equilibrium data for gas hydrates with different types of inhibitor additives from 30 peer-reviewed journal articles published between 1987 and 2026; to employ the data quality standard of thermodynamic consistency using the criteria developed by Sa et al. (2018); to develop and compare seven ML regression models with tuned hyperparameters; to select the best-performing model according to its physical validity; and to interpret results using SHAP importance of features. The novelty of the proposed approach in relation to any previous work employing machine learning for hydrate predictions lies in the integration of thermodynamic considerations (consistency screening, uninhibited reference, physical validation of trends) at each step of the ML workflow.

6.2 Principal Conclusions

Conclusion 1 - Thermodynamic consistency screening is a critical and quantitatively significant prerequisite for machine learning training on hydrate phase equilibrium data.

The three-criterion screening approach proposed by Sa et al. (2018) was applied to the initial dataset of 1,751 data points before model development began. This process removed 584 thermodynamically inconsistent entries (33.4%), leaving 1,167 consistent points, which were further preprocessed into a final clean training dataset of 922 points. To the author's knowledge, this type of screening has not previously been used in ML-based studies of inhibited hydrate prediction. Its importance is supported by the known sensitivity of ML models to systematic experimental errors in training data. Supporting evidence from Rasoolzadeh et al. (2025) for

uninhibited systems shows that eliminating inconsistent data can reduce temperature prediction error by up to 43% (from 0.44 K to 0.25 K) by removing points that violate Clausius–Clapeyron linearity, heat of dissociation consistency, or water activity consistency. This work extends that principle to inhibited systems, which are more data-intensive and experimentally complex, leading to higher inherent uncertainty due to inhibitor–water interactions. The fact that the trained models reproduce correct thermodynamic trends without explicit constraints provides qualitative confirmation that the screened dataset contains physically consistent information.

Conclusion 2 - All six machine learning models achieve strong predictive performance on the test set, with three gradient boosting models forming a statistically equivalent top group.

After a two-stage hyperparameter optimization process, all models achieved test R^2 values above 0.90 and RMSE values below 3.0 K. CatBoost delivered the lowest RMSE (0.931 K, $R^2 = 0.989$), followed closely by XGBoost (RMSE = 0.990 K, $R^2 = 0.989$) and Gradient Boosting Regression (RMSE = 0.986 K, $R^2 = 0.989$), forming a tightly grouped top tier. SVR also performed strongly (RMSE = 0.963 K, $R^2 = 0.989$), matching the top models in R^2 and ranking second in RMSE. Random Forest showed clear overfitting (training RMSE = 0.020 K vs test RMSE = 1.651 K), while Decision Tree had the weakest generalization (RMSE = 2.859 K). AIC and BIC analyses revealed large training–test complexity gaps for RF and DT, indicating that these models captured noise rather than underlying physical relationships. In contrast, gradient boosting models and SVR showed more balanced and efficient complexity, with AIC gaps ranging from 619 (SVR) to 1274 (XGBoost).

Conclusion 3 - SVR with an RBF kernel is the most suitable overall model when physical generalization is prioritized alongside statistical accuracy.

This conclusion is supported by two key observations. First, SVR achieves the same test R^2 (0.989) as the top-performing gradient boosting models and only slightly higher RMSE than CatBoost (0.963 K vs 0.931 K), a negligible difference of 0.032 K. Second, interpolation analysis shows that SVR produces smooth, monotonic, and physically realistic P–T equilibrium curves at MgCl_2 concentrations not included in training, across all methanol levels tested (0, 5, 10, and 20 wt%). This behavior stems from the RBF kernel, which enforces smoothness by favoring the flattest function within the ϵ -insensitive margin in reproducing kernel Hilbert space, resulting in continuous prediction surfaces. In contrast, CatBoost’s tree-based structure

partitions the feature space into discrete regions, which can lead to inconsistencies in sparsely sampled areas. SVR also has the smallest AIC gap (619), indicating the best balance between model complexity and generalization. Since real-world flow assurance scenarios often require interpolation between discrete experimental conditions, this capability is essential.

Conclusion 4 - The SVR model successfully captures the physical thermodynamics of inhibited hydrate systems without explicitly imposing thermodynamic constraints during training.

Validation tests show that the model reproduces several key behaviors: it maintains a positive P–T slope across the full pressure range (3–13 MPa), correctly orders equilibrium curves by inhibitor concentration (higher concentration leading to lower equilibrium temperature), captures the combined temperature suppression effect of MgCl_2 and methanol, and reflects the expected monotonic increase in suppression temperature (ΔT) with increasing methanol concentration at fixed MgCl_2 levels. It also correctly represents how baseline ΔT varies with MgCl_2 concentration. Together, these results demonstrate that training on thermodynamically consistent data, combined with the smooth interpolation properties of the SVR model, enables the learning of physically meaningful representations directly from experimental data.

6.3 Recommendations

Recommendation 1 - Enhance the training dataset with multicomponent compositions of real natural gas (C_2+).

The most critical near-term enhancement of the predictive framework would be the inclusion of additional data relating to the inhibited phase equilibrium of hydrates of thermogenic field gas compositions, i.e., those containing ethane, propane, and butane constituents (C_2+ concentration in the range of 5–20 mol%). The minimal SHAP influence of ethane (C_2H_6) and C_4 features in the existing model reflects their absence in the training data. Field-representative gas compositions will enable the application of the ML pipeline to actual field flow assurance analysis and is thus an essential step prior to actual field implementation.

Recommendation 2 - Apply the consistency screening/ML pipeline to novel inhibitor types.

The thermodynamic consistency methodology proposed in this study can be readily extended to the case of other inhibitors provided that the equilibrium phase diagram without inhibitors can be calculated from the Clausius-Clapeyron relation. Future work will need to incorporate inhibitors based on DESs, amino acids, and ionic liquids, active lines of research in environmentally friendly THIs (Rashnavadi et al., 2024; Cai et al., 2025). The multi-column inhibitor description adopted here will easily accommodate these new classes of THIs.

Recommendation 3 – Investigate prediction validity in the extrapolation regime and introduce uncertainty quantification.

In this work, the interpolation validity was thoroughly evaluated through the cross-validation and interpolation error analysis procedure. However, the extrapolation regime was not tested experimentally. A systematic study of equilibrium data at pressures over 73 MPa and temperatures under 243 K will determine the prediction region within which SVR provides consistent results, as extrapolation beyond this region will result in inconsistencies between predictions and known thermodynamics. The development of uncertainty quantification techniques appropriate for SVR, such as conformal prediction or Bayesian approaches, would make it possible to provide uncertainty intervals for predictions, thereby enhancing the engineering utility of the model.

Recommendation 4 – Explore the hybrid physics-based ML-thermodynamics approach.

The future direction of research may be developing the idea of a hybrid ML approach based on both consistency constraints and pattern recognition. A physics-informed SVR algorithm will be formulated, where consistency constraints, such as monotone relationship between the inhibition temperature and inhibitor concentration or consistent P-T slope according to the Clausius-Clapeyron law, can be incorporated in the form of regularization terms in the optimization process. Physics-informed ML techniques have proven successful for estimating thermophysical properties and, if employed in this context, could significantly enhance both accuracy and reliability of predictions outside the training domain.

Recommendation 5 – Implement the algorithm in a real-time pipeline monitoring system.

The efficiency of predicting hydrate temperatures with SVR lies in fast computations required to perform a prediction at runtime, which amounts to a few microseconds after training. Therefore, it would be technically feasible to implement a validated model within a flow

assurance monitoring tool, which accepts pipeline parameters such as pressure, temperature, gas composition, and inhibitor concentration as input and predicts formation risks continuously. The consistent behavior of the SVR model within the training region, proved by interpolation analysis in this study, justifies its implementation in real time applications within the same domain.

7. REFERENCES

- Afzal, W., Mohammadi, A. H., & Richon, D. (2008). Experimental measurements and predictions of dissociation conditions for methane, ethane, propane, and carbon dioxide simple hydrates in the presence of diethylene glycol aqueous solutions. *Journal of Chemical & Engineering Data*, 53(3), 663–666. <https://doi.org/10.1021/jc700457r>
- Behnammotlagh, M. A., Hashemi, R., Taheri Rizi, Z., Mohammadtaheri, M., & Mohammadi, M. (2022). Experimental study of the effect of the combined monoethylene glycol with NaCl/CaCl₂ salts on sour gas hydrate inhibition with low-concentration hydrogen sulfide. *Journal of Chemical & Engineering Data*, 67, 1250–1258. <https://doi.org/10.1021/acs.jced.1c00888>
- Bentéjac, C., Csörgő, A., & Martínez-Muñoz, G. (2019). A comparative analysis of XGBoost. *arXiv preprint arXiv:1911.01914*. <https://arxiv.org/abs/1911.01914>
- Bishnoi, P. R., & Dholabhai, P. D. (1999). Equilibrium conditions for hydrate formation for a ternary mixture of methane, propane and carbon dioxide, and a natural gas mixture in the presence of electrolytes and methanol. *Fluid Phase Equilibria*, 158–160, 821–827.
- Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5–32. <https://doi.org/10.1023/A:1010933404324>
- Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). *Classification and regression trees*. Wadsworth International Group.
- Cai, J., Tang, H., Zhang, T.-H., Xiao, P., Wu, Y.-W., Qin, H.-H., Chen, G.-J., Sun, C.-Y., & Wang, X.-H. (2025). Phase equilibria of gas hydrates: A review of experiments, modeling, and potential trends. *Renewable and Sustainable Energy Reviews*, 209, 115082. <https://doi.org/10.1016/j.rser.2024.115082>
- Dholabhai, P. D., & Bishnoi, P. R. (1994). Hydrate equilibrium conditions in aqueous electrolyte solutions: Mixtures of methane and carbon dioxide. *Journal of Chemical & Engineering Data*, 39, 191–194.
- Dholabhai, P. D., Parent, J. S., & Bishnoi, P. R. (1997). Equilibrium conditions for hydrate formation from binary mixtures of methane and carbon dioxide in the presence of electrolytes, methanol and ethylene glycol. *Fluid Phase Equilibria*, 141, 235–246.
- Du, J., Wang, X., Liu, H., Guo, P., Wang, Z., & Fan, S. (2019). Experiments and prediction of phase equilibrium conditions for methane hydrate formation in the NaCl, CaCl₂, MgCl₂ electrolyte solutions. *Fluid Phase Equilibria*, 479, 1–8. <https://doi.org/10.1016/j.fluid.2018.09.028>
- Fan, S.-S., & Guo, T.-M. (1999). Hydrate formation of CO₂-rich binary and quaternary gas mixtures in aqueous sodium chloride solutions. *Journal of Chemical & Engineering Data*, 44(4), 829–832. <https://doi.org/10.1021/jc990011b>
- Fan, S.-S., Chen, G.-J., Ma, Q.-L., & Guo, T.-M. (2000). Experimental and modeling studies on the hydrate formation of CO₂ and CO₂-rich gas mixtures. *Chemical Engineering Journal*, 78, 173–178.

- Fu, Q., Chen, M., Pang, W., Xu, Z., Liu, Z., Wen, H., & Lei, X. (2024). Experimental and modeling study on methane hydrate equilibrium conditions in the presence of inorganic salts. *Molecules*, 29, 3702. <https://doi.org/10.3390/molecules29153702>
- Gong, Y., Mendgaziev, R. I., Hu, W., Li, Y., Li, Z., Stoporev, A. S., Manakov, A. Y., Vinokurov, V. A., Li, T., & Semenov, A. P. (2022). Urea as a green thermodynamic inhibitor of sII gas hydrates. *Chemical Engineering Journal*, 429, 132386. <https://doi.org/10.1016/j.cej.2021.132386>
- González, S., García, S., Del Ser, J., Rokach, L., & Herrera, F. (2020). A practical tutorial on bagging and boosting based ensembles for machine learning: Algorithms, software tools, performance study, practical perspectives and opportunities. *Information Fusion*, 64, 205–237. <https://doi.org/10.1016/j.inffus.2020.07.007>
- Hassanpouryouzband, A., Joonaki, E., Vasheghani Farahani, M., Takeya, S., Ruppel, C., Yang, J., English, N. J., Schicks, J. M., Edlmann, K., Mehrabian, H., Aman, Z. M., & Tohidi, B. (2020). Gas hydrates in sustainable chemistry. *Chemical Society Reviews*, 49(15), 5225–5309. <https://doi.org/10.1039/C8CS00989A>
- Herri, J.-M., Bouchemoua, A., Kwaterski, M., Fezoua, A., Ouabbas, Y., & Cameirao, A. (2011). Gas hydrate equilibria for CO₂–N₂ and CO₂–CH₄ gas mixtures—Experimental studies and thermodynamic modelling. *Fluid Phase Equilibria*, 301, 171–190. <https://doi.org/10.1016/j.fluid.2010.09.041>
- Jager, M. D., Peters, C. J., & Sloan, E. D. (2002). Experimental determination of methane hydrate stability in methanol and electrolyte solutions. *Fluid Phase Equilibria*, 193(1–2), 17–28. [https://doi.org/10.1016/S0378-3812\(01\)00623-9](https://doi.org/10.1016/S0378-3812(01)00623-9)
- Kang, S.-P., Lee, H., Lee, C.-S., & Sung, W.-M. (2001). Hydrate phase equilibria of the guest mixtures containing CO₂, N₂ and tetrahydrofuran. *Fluid Phase Equilibria*, 185, 101–109.
- Kashish, Yusuf, M., Beg, M., Kamyab, H., Muzammil, K., Ahmad, M., Gupta, M., Kumar, S., Rezanian, S., & Ibrahim, H. (2024). Natural gas hydrates: A review of various inhibitors and respective mechanisms. *Journal of Molecular Liquids*, 403, 124809. <https://doi.org/10.1016/j.molliq.2024.124809>
- Keshavarz, L., Ghaani, M. R., & English, N. J. (2022). Thermodynamic evaluation of inhibitors for methane-hydrate formation. *Fuel*, 324, 124672. <https://doi.org/10.1016/j.fuel.2022.124672>
- Khan, M. S., Lal, B., Shariff, A. M., & Mukhtar, H. (2019). Ammonium hydroxide ILs as dual-functional gas hydrate inhibitors for binary mixed gas (carbon dioxide and methane) hydrates. *Journal of Molecular Liquids*, 274, 33–44. <https://doi.org/10.1016/j.molliq.2018.10.076>
- Khoo, Y. Z., Sun, D.-H., & Chen, L.-J. (2025). Potential use of deep eutectic solvents as dual-purpose green inhibitors of methane hydrate formation for flow assurance. *Journal of Molecular Liquids*, 428, 127542. <https://doi.org/10.1016/j.molliq.2025.127542>
- Koh, C. A., Sloan, E. D., Sum, A. K., & Wu, D. T. (2011). Fundamentals and applications of gas hydrates. *Annual Review of Chemical and Biomolecular Engineering*, 2, 237–257. <https://doi.org/10.1146/annurev-chembioeng-061010-114152>

- Lee, H. J., Lee, J. D., Linga, P., Englezos, P., Kim, Y. S., Lee, M. S., & Kim, Y. D. (2010). Gas hydrate formation process for pre-combustion capture of carbon dioxide. *Energy*, 35, 2729–2733.
- Li, Y., Chen, J., Gambelli, A. M., Zhao, X., Gao, Y., Rossi, F., & Mei, S. (2022). In situ experimental study on the effect of mixed inhibitors on the phase equilibrium of carbon dioxide hydrate. *Chemical Engineering Science*, 248, 117230. <https://doi.org/10.1016/j.ces.2021.117230>
- Liang, H., & Zhao, Q. (2024). Hydrate equilibrium conditions of CO₂ containing natural gas in methanol/ethylene glycol/sodium chloride aqueous solutions. *Journal of Chemical & Engineering Data*, 69, 2606–2612. <https://doi.org/10.1021/acs.jced.4c00021>
- Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, & R. Garnett (Eds.), *Advances in neural information processing systems 30* (pp. 4765–4774). Curran Associates. <https://proceedings.neurips.cc/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html>
- Maekawa, T. (2001). Equilibrium conditions for gas hydrates of methane and ethane mixtures in pure water and sodium chloride solution. *Geochemical Journal*, 35, 59–66.
- Medeiros, F. de A., Segtovich, I. S. V., Tavares, F. W., & Sum, A. K. (2020). Sixty years of the van der Waals and Platteeuw model for clathrate hydrates—A critical review from its statistical thermodynamic basis to its extensions and applications. *Chemical Reviews*, 120(24), 13349–13381. <https://doi.org/10.1021/acs.chemrev.0c00494>
- Mohammadi, A. H., & Richon, D. (2010). Phase equilibria of methane hydrates in the presence of methanol and/or ethylene glycol aqueous solutions. *Industrial & Engineering Chemistry Research*, 49(2), 925–928. <https://doi.org/10.1021/ie901357m>
- Nasir, Q., Lau, K. K., Lal, B., & Sabil, K. M. (2014). Hydrate dissociation condition measurement of CO₂-rich mixed gas in the presence of methanol/ethylene glycol and mixed methanol/ethylene glycol + electrolyte aqueous solution. *Journal of Chemical & Engineering Data*, 59, 3920–3926. <https://doi.org/10.1021/je5008313>
- Nasrifar, K., & Moshfeghian, M. (2001). A model for prediction of gas hydrate formation conditions in aqueous solutions containing electrolytes and/or alcohol. *The Journal of Chemical Thermodynamics*, 33(9), 999–1014. <https://doi.org/10.1006/jcht.2000.0811>
- Ng, H.-J., Chen, C.-J., & Saeterstad, T. (1987). Hydrate formation and inhibition in gas condensate and hydrocarbon liquid systems. *Fluid Phase Equilibria*, 36, 99–106.
- Ohgaki, K., Takano, K., Sangawa, H., Matsubara, T., & Nakano, S. (1996). Methane exploitation by carbon dioxide from gas hydrates—Phase equilibria for CO₂–CH₄ mixed hydrate system. *Journal of Chemical Engineering of Japan*, 29(3), 478–483. <https://doi.org/10.1252/jcej.29.478>
- Østergaard, K. K., Tohidi, B., Anderson, R., Todd, A. C., & Danesh, A. (2002). Can 2-propanol form clathrate hydrates? *Industrial & Engineering Chemistry Research*, 41, 2064–2068.

- Parrish, W. R., & Prausnitz, J. M. (1972). Dissociation pressures of gas hydrates formed by gas mixtures. *Industrial & Engineering Chemistry Process Design and Development*, 11(1), 26–35. <https://doi.org/10.1021/i260041a006>
- Peng, D.-Y., & Robinson, D. B. (1976). A new two-constant equation of state. *Industrial & Engineering Chemistry Fundamentals*, 15(1), 59–64. <https://doi.org/10.1021/i160057a011>
- Rasoolzadeh, A., Bakhtyari, A., Javanmardi, J., & Mohammadi, A. H. (2025). A straightforward model for determining gas hydrate phase equilibrium conditions and enthalpy of hydrate dissociation for natural gas components. *The Canadian Journal of Chemical Engineering*, 103(1), 410–424. <https://doi.org/10.1002/cjce.70011>
- Rehman, A. U., Abdulwahab, A., Kaur, A., Khan, M. S., Zaini, D. B., Shariff, A. M., & Lal, B. (2022). Experimental investigation and modelling of synergistic thermodynamic inhibition of diethylene glycol and glycine mixture on CO₂ gas hydrates. *Chemosphere*, 308, 136181. <https://doi.org/10.1016/j.chemosphere.2022.136181>
- Robinson, D. B., & Hutton, J. M. (1966). Hydrate formation in systems containing methane, hydrogen sulphide and carbon dioxide. *Journal of Canadian Petroleum Technology*.
- Sa, J.-H., Hu, Y., & Sum, A. K. (2018). Assessing thermodynamic consistency of gas hydrates phase equilibrium data for inhibited systems. *Fluid Phase Equilibria*, 473, 294–299. <https://doi.org/10.1016/j.fluid.2018.06.012>
- Saberi, A., Alamdari, A., Rasoolzadeh, A., & Mohammadi, A. H. (2021). Insights into kinetic inhibition effects of MEG, PVP, and L-tyrosine aqueous solutions on natural gas hydrate formation. *Petroleum Science*, 18, 495–508.
- Schicks, J. M. (2022). Gas hydrates in nature and in the laboratory: necessary requirements for formation and properties of the resulting hydrate phase. *ChemTexts*, 8(2), 13. <https://doi.org/10.1007/s40828-022-00164-3>
- Schmidhuber, J. (2015). Deep learning in neural networks: An overview. *Neural Networks*, 61, 85–117. <https://doi.org/10.1016/j.neunet.2014.09.003>
- Semenov, A. P., Mendgaziev, R. I., & Stoporev, A. S. (2023). Dataset for the experimental study of dimethyl sulfoxide as a thermodynamic inhibitor of methane hydrate formation. *Data in Brief*, 48, 109283. <https://doi.org/10.1016/j.dib.2023.109283>
- Semenov, A. P., Mendgaziev, R. I., Istomin, V. A., Sergeeva, D. V., Vinokurov, V. A., Gong, Y., Li, T., & Stoporev, A. S. (2024). Data on searching for synergy between alcohol and salt to produce more potent and environmentally benign gas hydrate inhibitors. *Data in Brief*, 53, 110138. <https://doi.org/10.1016/j.dib.2024.110138>
- Semenov, A. P., Mendgaziev, R. I., Stoporev, A. S., Istomin, V. A., Sergeeva, D. V., Ogienko, A. G., & Vinokurov, V. A. (2021). The pursuit of a more powerful thermodynamic hydrate inhibitor than methanol: Dimethyl sulfoxide as a case study. *Chemical Engineering Journal*, 423, 130227. <https://doi.org/10.1016/j.cej.2021.130227>
- Semenov, A. P., Mendgaziev, R. I., Stoporev, A. S., Istomin, V. A., Sergeeva, D. V., Tulegenov, T. B., & Vinokurov, V. A. (2022). Dimethyl sulfoxide as a novel thermodynamic inhibitor of

- carbon dioxide hydrate formation. *Chemical Engineering Science*, 255, 117670. <https://doi.org/10.1016/j.ces.2022.117670>
- Seo, Y.-T., Lee, H., & Yoon, J.-H. (2001). Hydrate phase equilibria of the carbon dioxide, methane, and water system. *Journal of Chemical & Engineering Data*, 46(2), 381–384. <https://doi.org/10.1021/je000237a>
- Sloan, E. D. (2003). Fundamental principles and applications of natural gas hydrates. *Nature*, 426(6964), 353–359. <https://doi.org/10.1038/nature02135>
- Sloan, E. D., & Koh, C. A. (2008). *Clathrate hydrates of natural gases* (3rd ed.). CRC Press. <https://doi.org/10.1201/9781420008494>
- Smola, A. J., & Schölkopf, B. (2004). A tutorial on support vector regression. *Statistics and Computing*, 14(3), 199–222. <https://doi.org/10.1023/B:STCO.0000035301.49549.88>
- Song, K. Y., & Kobayashi, R. (1989). Final hydrate stability conditions of a methane and propane mixture in the presence of pure water and aqueous solutions of methanol and ethylene glycol. *Fluid Phase Equilibria*, 47, 295–308.
- Taheri Rizi, Z., Aminolroayaei, M. A., Mohammad-Taheri, M., Kamran-Pirzaman, A., & Seif, M. (2026). Experimental and thermodynamic modeling of hydrate inhibitors for sour gas systems in South Pars gas fields. *SPE Journal*, 31(1), 410–??.
- Tamaskovics, A., Kummer, N.-A., & Amro, M. (2026). Thermodynamic gas hydrate inhibition for CO₂ injection in depleted gas reservoirs. *Gas Science and Engineering*, 149, 205868. <https://doi.org/10.1016/j.jgsce.2026.205868>
- Tohidi, B., Danesh, A., Todd, A. C., & Burgass, R. W. (1997). Hydrate-free zone for synthetic and real reservoir fluids in the presence of saline water. *Chemical Engineering Science*, 52(19), 3257–3263.
- Unruh, C. H., & Katz, D. L. (1949). Gas hydrates of carbon dioxide-methane mixtures. *Petroleum Transactions, AIME*, 186, 83–86.
- Valadan Zoj, A. M., Javanmardi, J., Rasoolzadeh, A., & Mohammadi, A. H. (2022). Experimental measurement and thermodynamic modeling of methane hydrate dissociation conditions in the presence of diglycolamine aqueous solution. *Industrial & Engineering Chemistry Research*, 61, 13683–13693. <https://doi.org/10.1021/acs.iecr.2c00315>
- Varoquaux, G., & Colliot, O. (2023). Evaluating machine learning models and their diagnostic value. In O. Colliot (Ed.), *Machine learning for brain disorders* (pp. 601–630). Springer. https://doi.org/10.1007/978-1-0716-3195-9_20
- Wang, Y., Fan, S., & Lang, X. (2019). Reviews of gas hydrate inhibitors in gas-dominant pipelines and application of kinetic hydrate inhibitors in China. *Chinese Journal of Chemical Engineering*, 27(9), 2118–2132. <https://doi.org/10.1016/j.cjche.2019.02.023>
- Wu, B.-J., Robinson, D. B., & Ng, H.-J. (1976). Three- and four-phase hydrate forming conditions in methane + isobutane + water. *The Journal of Chemical Thermodynamics*, 8(5), 461–469. [https://doi.org/10.1016/0021-9614\(76\)90005-8](https://doi.org/10.1016/0021-9614(76)90005-8)

8. APPENDICES

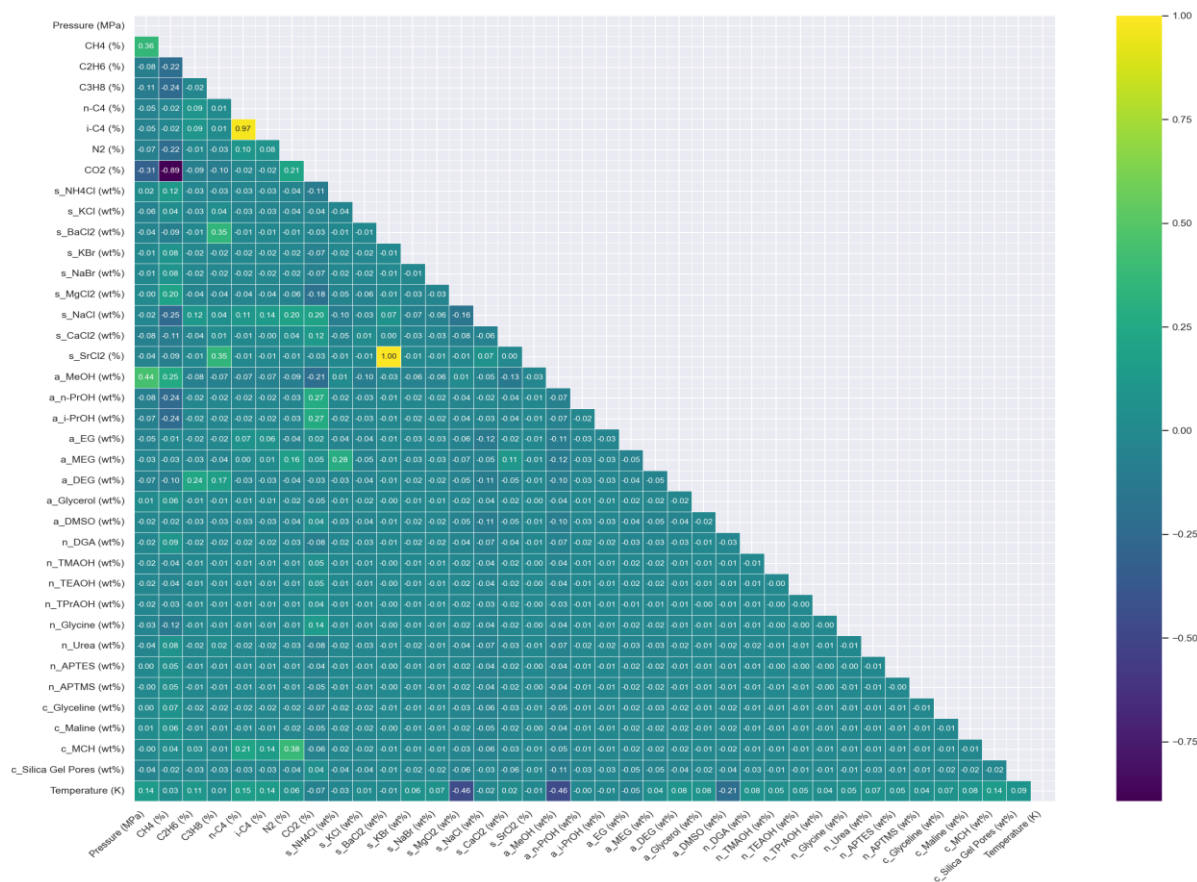


Figure 33. EDA outputs: correlation heatmap before preprocessing

