

**Data-Driven Evaluation of Thermodynamic Hydrate
Promoters: Performance of water-soluble, water-insoluble,
semi-clathrate and gas-phase promoters**

by

Alikhan Ziyabek

THESIS SUPERVISOR

Shams Kalam

THESIS CO-SUPERVISOR

Masoud Riazi

Ali Rasoolzadeh (External from Behbahan Khatam Alanbia University of Technology, Iran)

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 Engineering

Nazarbayev University

10.04.2026

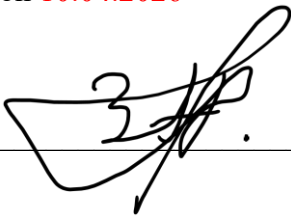
ORIGINALITY STATEMENT

I, **Alikhan Ziyabek**, 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 **10.04.2026**



A handwritten signature in black ink, consisting of stylized letters and a large loop, is written over a horizontal line.

ABSTRACT

In this work, we propose an objective comparative evaluation of the thermodynamic characteristics of hydrate promoters belonging to four different classes based on the performance of machine learning predictive models of hydrate equilibrium pressure. The dataset includes roughly 1,900 points of experimentally obtained phase equilibrium data for 21 promoter molecules across nine guest gas types. For each class of promoters, four algorithms – namely Random Forest, XGBoost, CatBoost, and Support Vector Regression (SVR) – were applied and evaluated on the following metrics – R^2 , RMSE, AIC, and BIC – taking into account promoter identity using molecular descriptors such as molecular weight, dipole moment, and LogP.

For three out of four promoter classes, Support Vector Regression (SVR) demonstrated the best performance. In the case of semi-clathrates, SVR showed $R^2 = 0.934$ and RMSE = 0.724 MPa while the second-place XGBoost reached only $R^2 = 0.71$ owing to a very high temperature-pressure correlation ($r \approx -0.98$). As for water-insoluble promoters, SVR performed equally well, achieving $R^2 = 0.84$ and RMSE = 1.12 MPa. In the case of water-soluble promoters, CatBoost outperformed SVR ($R^2 = 0.92$ and RMSE = 0.91 MPa) due to the ability of this algorithm to exploit an ordered boosting process in regions close to stoichiometry. No clear winner was found for the gas-phase category, as R^2 did not exceed 0.67 for any of the algorithms due to the insufficient number of samples and limitations in molecular descriptor space. External validation confirmed the generalizability of the developed models, with mean absolute errors (MAE) of 0.237 MPa (water-soluble), 0.147 MPa (water-insoluble), and 0.114 MPa (semi-clathrate).

A ranking of the four classes according to cross-category performance of ML models is as follows: semi-clathrates ($R^2 = 0.934$) > water-solubles ($R^2 = 0.92$) > water-insolubles ($R^2 = 0.84$) > gas-phase ($R^2 = 0.67$). Thus, it can be assumed that ML performance is a metric indicating the degree of thermodynamic coherence within a category, supporting our hypothesis.

DEDICATION

To my family and friends.

ACKNOWLEDGMENT

I would like to express my sincere gratitude to my supervisors, Dr. Shams Kalam and Dr. Masoud Riazi, for their invaluable guidance, constructive feedback, and continued support throughout the course of this research. Their expertise and encouragement were instrumental in shaping both the direction and quality of this work.

I am also grateful to the School of Mining and Geosciences at Nazarbayev University for providing the academic environment and resources that made this research possible.

TABLE OF CONTENTS

TABLE OF CONTENTS.....	VI
LIST OF FIGURES	X
LIST OF TABLES	XII
1. INTRODUCTION.....	1
1.1 Background/Problem Definition.....	1
1.2 Objectives of the Thesis	4
1.2.1 Main Objectives	4
1.2.2 Specific Objectives.....	5
1.3 Scope of Work	5
2. LITERATURE REVIEW.....	6
2.1 Overview of Thermodynamic Hydrate Promoter Categories	6
2.2 Water-Soluble Thermodynamic Promoters.....	6
2.2.1 Category Overview and Promotion Mechanism	6
2.2.2 Tetrahydrofuran - Benchmark Compound	7
2.2.3 Other Water-Soluble Promoters	8
2.3 Water-Insoluble Thermodynamic Promoters.....	9
2.3.1 Category Overview and Promotion Mechanism	9
2.3.2 Cyclopentane - Benchmark Compound	10
2.3.3 Other Water-Insoluble Promoters	11
2.4 Semi-Clathrate Hydrate Promoters.....	13
2.4.1 Category Overview and Structural Distinction	13
2.4.2 TBAB - Benchmark Compound.....	13
2.4.3 Other Semi-Clathrate Promoters	14
2.5 Gas-Phase Thermodynamic Promoters.....	15
2.6 Comparative Overview of Promoter Categories	16
2.7 Machine Learning in Gas Hydrate Phase Equilibrium Prediction	17
2.7.1 Traditional Thermodynamic Modeling Approaches and Their Limitations	17
2.7.2 Overview of ML Landscape.....	18
2.7.2 Previous Machine Learning Studies on Promoter-Assisted Hydrate Systems	19
2.7.3 Research Gap and Justification for the Present Study	22
3. METHODOLOGY.....	23
3.1 Data Collection and Compilation.....	23
3.1.1 Data Sources and Scope	23

3.1.2 Dataset Composition and Size.....	23
3.1.3 Raw Variable Structure	24
3.2 Data Preprocessing.....	25
3.2.1 Data Cleaning	25
3.2.2 Unit Standardization and Concentration Conversion	26
3.2.3 Feature Engineering	26
3.2.4 Feature Scaling	29
3.2.5 Feature Selection and Exploratory Data Analysis.....	29
3.3 Machine Learning Algorithms	34
3.4 Model Training and Validation	35
3.4.1 Train-Test Split	35
3.4.2 Cross-Validation.....	35
3.4.3 Hyperparameter Tuning	35
3.4.4 Model Evaluation Metrics.....	36
3.4.5 Internal and External Validation	37
3.5 Comparative Evaluation Framework.....	38
4. RESULTS.....	38
4.1 Overview	38
4.2 Water-Soluble Promoters	39
4.2.1 Model Performance	39
4.2.2 Best Model - CatBoost	45
4.2.3 Internal Validation – CatBoost.....	45
4.2.4 External Validation - CatBoost	47
4.3 Water-Insoluble Promoters.....	48
4.3.1 Model Performance	48
4.3.2 Best Model — SVR.....	54
4.3.3 Internal Validation – SVR.....	54
4.3.4 External Validation — SVR.....	56
4.4 Semi-Clathrate Promoters.....	58
4.4.1 Model Performance	58
4.4.2 Best Model — SVR.....	64
4.4.3 Internal Validation – SVR.....	64
4.4.4 External Validation — SVR.....	66
4.5 Gas-Phase Promoters.....	68
4.5.1 Model Performance	68

4.5.2 Best Model — SVR.....	74
4.5.3 Internal Validation – SVR.....	74
4.5.4 External Validation — Not Performed.....	76
4.6 Summary of Results Across All Categories	76
5. DISCUSSION	77
5.1 Overview of Discussion Approach	77
5.2 Water-Soluble Promoters	77
5.2.1 Algorithm Comparison and CatBoost Performance.....	77
5.2.2 Internal Validation: Water-Soluble Promoters (CatBoost)	78
5.2.3 External Validation Interpretation.....	81
5.2.4 Physical Interpretation.....	81
5.3 Water-Insoluble Promoters	81
5.3.1 Algorithm Comparison and SVR Performance.....	81
5.3.2 Internal Validation: Water-Insoluble Promoters (SVR).....	82
5.3.3 External Validation Interpretation.....	84
5.3.4 Physical Interpretation.....	85
5.4 Semi-Clathrate Promoters.....	85
5.4.1 Algorithm Comparison and Dramatic SVR Dominance	85
5.4.2 Internal Validation: Semi-Clathrate Promoters (SVR).....	86
5.4.3 External Validation Interpretation.....	89
5.4.4 Physical Interpretation.....	90
5.5 Gas-Phase Promoters	90
5.5.1 Algorithm Convergence and Performance Limitations.....	90
5.5.2 Internal Validation: Gas-Phase Promoters (SVR).....	91
5.5.3 Why External Validation Was Not Performed.....	94
5.6 Cross-Category Comparative Analysis	94
5.6.1 Performance Ranking and Thermodynamic Coherence.....	94
5.6.2 SVR Dominance Across Three Categories	95
5.6.3 Implications for Promoter Selection.....	96
5.7 Limitations of the Study	97
5.7.1 Dataset Size and Distribution Imbalances.....	97
5.7.2 Gas-Phase Category Limitations	98
5.7.3 External Validation Sample Sizes	98
5.7.4 Feature Engineering Limitations	98
5.7.5 Absence of Baseline Comparison.....	99

6. CONCLUSIONS AND RECOMMENDATIONS.....99

6.1 Conclusions 99

6.1.1 Category-Specific Conclusions 99

6.1.2 Cross-Cutting Methodological Conclusions 101

6.2 Recommendations 102

6.2.1 Data Expansion and Curation..... 102

6.2.2 Feature Engineering Enhancements 102

6.2.3 Algorithmic and Modeling Framework Improvements..... 103

6.2.4 Industrial Application Directions 103

7. REFERENCES.....104

LIST OF FIGURES

Figure 1. Three types of hydrate structures and their associated cages. (Majid et al., 2021).....	2
Figure 2. Example of hydrate equilibrium (no THI/THP, curve B) shift as a result of the addition of thermodynamic hydrate inhibitors (THIs, curve A) or thermodynamic hydrate promoters (THPs, curve C) (Majid et al., 2021).....	3
Figure 3. Pearson Correlation Heatmap for Water-soluble promoter dataset.	31
Figure 4. Pearson Correlation Heatmap for Water-insoluble promoter dataset.	32
Figure 5. Pearson Correlation Heatmap for Semi-clathrate promoter dataset.....	33
Figure 6. Pearson Correlation Heatmap for Gas promoters dataset.	34
Figure 7. RF Scatter Plot for Water-soluble training dataset	41
Figure 8. RF Scatter Plot for Water-soluble testing dataset.	41
Figure 9. XGBoost Scatter Plot for Water-soluble training dataset.	42
Figure 10. XGBoost Scatter Plot for Water-soluble testing dataset.....	42
Figure 11. CatBoost Scatter Plot for Water-soluble training dataset	43
Figure 12. CatBoost Scatter Plot for Water-soluble testing dataset	43
Figure 13. SVR Scatter Plot for Water-soluble training dataset	44
Figure 14. SVR Scatter Plot for Water-soluble testing dataset	44
Figure 15. CatBoost predicted vs actual for different promoter concentrations for WS.....	46
Figure 16. CatBoost predicted vs actual for different gas compositions for WS.	46
Figure 17. CatBoost predicted vs actual for different promoters at the same concentration for WS....	47
Figure 18. P-T Equilibrium curve for experimental and predicted values with WS CatBoost model. .	47
Figure 19. SVR Scatter Plot for Water-insoluble training dataset	50
Figure 20. SVR Scatter Plot for Water-insoluble testing dataset	50
Figure 21. XGBoost Scatter Plot for Water-insoluble training dataset.....	51
Figure 22. XGBoost Scatter Plot for Water-insoluble testing dataset.....	51
Figure 23. RF Scatter Plot for Water-insoluble training dataset	52
Figure 24. RF Scatter Plot for Water-insoluble testing dataset.....	52
Figure 25. CatBoost Scatter Plot for Water-insoluble training dataset	53
Figure 26. CatBoost Scatter Plot for Water-insoluble testing dataset.....	53
Figure 27. SVR predicted vs actual for different gas compositions for WiS.	55
Figure 28. SVR predicted vs actual for different promoter concentrations for WiS.....	55
Figure 29. SVR predicted vs actual for different promoters at the same concentration for WiS.....	56
Figure 30. P-T Equilibrium curve for experimental and predicted values with WiS SVR model.	57
Figure 31. SVR Scatter Plot for Semi-clathrate training dataset.....	59
Figure 32. SVR Scatter Plot for Semi-clathrate testing dataset	60

Figure 33. XGBoost Scatter Plot for Semi-clathrate training dataset	60
Figure 34. XGBoost Scatter Plot for Semi-clathrate testing dataset	61
Figure 35. CatBoost Scatter Plot for Semi-clathrate training dataset.....	61
Figure 36. CatBoost Scatter Plot for Semi-clathrate testing dataset	62
Figure 37. RF Scatter Plot for Semi-clathrate training dataset.....	62
Figure 38. CatBoost Scatter Plot for Semi-clathrate testing dataset	63
Figure 39. SVR predicted vs actual for different promoters at the same concentration for SC.	65
Figure 40. SVR predicted vs actual for different gas compositions for SC.	65
Figure 41. SVR predicted vs actual for different promoter concentrations for SC.	66
Figure 42. P-T Equilibrium curve for experimental and predicted values with SC SVR model.	67
Figure 43. RF Scatter Plot for Gas-phase promoter training dataset.....	70
Figure 44. RF Scatter Plot for Gas-phase promoter testing dataset	70
Figure 45. CatBoost Scatter Plot for Gas-phase promoter training dataset.....	71
Figure 46. CatBoost Scatter Plot for Gas-phase promoter testing dataset	71
Figure 47. XGBoost Scatter Plot for Gas-phase promoter training dataset	72
Figure 48. XGBoost Scatter Plot for Gas-phase promoter testing dataset	72
Figure 49. SVR Scatter Plot for Gas-phase promoter training dataset.....	73
Figure 50. SVR Scatter Plot for Gas-phase promoter testing dataset.....	73
Figure 51. SVR predicted vs actual for same promoter at different concentrations for GP.....	75
Figure 52. SVR predicted vs actual for different promoter concentrations for GP.....	75

LIST OF TABLES

Table 1. Summary of water-soluble thermodynamic promoter phase equilibrium studies.....	8
Table 2. Summary of water-insoluble thermodynamic promoter phase equilibrium studies.....	11
Table 3. Summary of semi-clathrate thermodynamic promoter phase equilibrium studies.....	14
Table 4. Cross-category comparison of thermodynamic hydrate promoter characteristics.	16
Table 5. Comparison of existing ML studies on promoter-assisted hydrate systems with the present thesis.....	21
Table 6. Summary of compiled datasets by promoter category.	24
Table 7. Input features used for each promoter category.	28
Table 8. Test set performance metrics for water-soluble promoter models.	39
Table 9. Hyperparameters summary for water-soluble category for all models	39
Table 10. External validation performance - water-soluble CatBoost model.	48
Table 11. Test set performance metrics for water-insoluble promoter models.	48
Table 12. Hyperparameters summary for water-insoluble category for all models	49
Table 13. External validation performance — water-insoluble SVR model.....	57
Table 14. Test set performance metrics for semi-clathrate promoter models.	58
Table 15. Hyperparameters summary for semi-clathrate category for all models.	58
Table 16. External validation performance - semi-clathrate SVR model.	67
Table 17. Test set performance metrics for gas-phase promoter models.	68
Table 18. Hyperparameters summary for gas-phase promoters category for all models.	68
Table 19. Summary of best model performance across all four promoter categories.	76
Table 20. Cross-category performance ranking and structural characteristics.....	94

1. INTRODUCTION

1.1 Background/Problem Definition

The freshwater availability and global demand for cleaner, more efficient energy carriers can be considered as the most crucial problems of the industrial sector. Since natural gas remains a critical fuel in the pathway towards decarbonization, the safe, efficient, and economically viable storage and transportation of gaseous energy resources have become a research priority of significant urgency. Among the techniques that are being evaluated for gas storage and transportation, clathrate hydrate-based systems have attracted a sufficient degree of interest among scientific and industrial researchers due to their unprecedented volumetric capacity of storage and thermodynamic properties of the hydrate phase.

Clathrate hydrate, commonly known as the gas hydrates, are non-stoichiometric crystalline solid compounds where guest gas molecules, such as methane, carbon dioxide, hydrogen, nitrogen, and other types of hydrocarbons, are trapped within hydrogen-bonded water structures, at high pressure and low temperature (Majid et al., 2021). There are three main crystallographic structures that are recognized in the literature: structure I (sI), structure II (sII), and structure H (sH), each having unique cage geometries and size compatibility with their guest molecules (Majid et al., 2021; Wu et al., 2023). The hydrate structures are represented in Figure 1. Gas hydrates have an impressive energy density; one cubic meter of hydrated methane under standard temperature and pressure has the capacity to hold around 164 volumes of gas (Majid et al., 2021), making the hydrate phase an attractive storage medium for solid natural gas (SNG) technology.

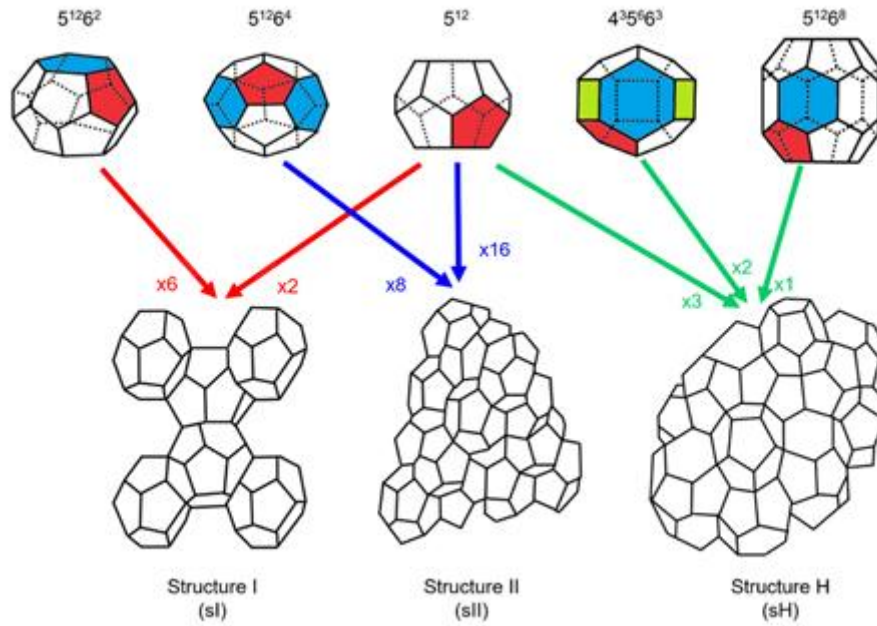


Figure 1. Three types of hydrate structures and their associated cages. (Majid et al., 2021).

Apart from the gas storage, clathrate hydrates have been investigated for a wide variety of engineering applications, such as carbon dioxide capture and sequestration, gas separation, hydrogen storage, and seawater desalination (Ho-Van et al., 2019; Majid et al., 2021). In the desalination setting, cyclopentane hydrates have been of particular interest as they can grow at ambient pressure and at a near-ambient temperature, while excluding salt ions from the crystal structure during growth (Ho-Van et al., 2019). The enclathration of CO₂ as compared to other gases in carbon capture applications offers a thermodynamic foundation to hydrate-based separation processes (Majid et al., 2021). These diverse applications highlight the technological applicability of gas hydrates as functional materials in a sustainable energy platform.

Nevertheless, the practical implementation of hydrate-based systems at an industrial scale is significantly challenged by the harsh temperature and pressure conditions required to stabilize the hydrate phase for common gas formers such as methane and carbon dioxide in pure water systems. Equilibrium conditions of pure methane hydrate demand several megapascals of pressure and temperatures slightly above the freezing point of water - conditions that impose considerable capital and operating expenses on any hydrate-based operations (Majid et al., 2021; Wu et al., 2023). This thermodynamic obstacle is the main issue that the research of hydrate promoters is aimed at.

Thermodynamic hydrate promoters (THPs) are chemical compounds that, when added to the hydrate-forming system, cause the phase equilibrium curve to shift to milder conditions, that is, lower pressures and/or higher temperatures compared to the pure water baseline (Majid

et al., 2021), as shown in Figure 2. This promotional effect is occurring since the THP molecules directly take part in the hydrate structure, usually occupying the large cages of the sII or sH lattice, stabilizing the crystal at conditions that would be thermodynamically unfavourable by the gas former alone (Wu et al., 2023). The size and character of this change are critically dependent on the physicochemical characteristics of the promoter, such as its molecular size, polarity, its capacity to dissolve in water, its cage occupancy, and its concentration in the aqueous phase.

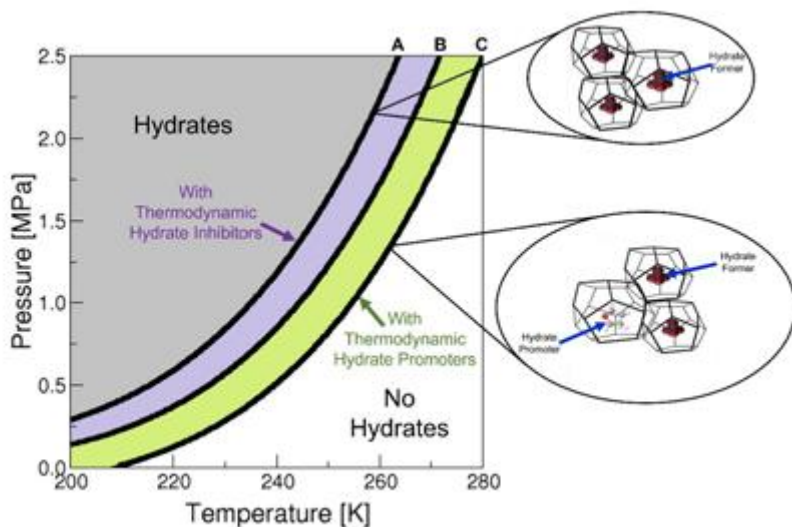


Figure 2. Example of hydrate equilibrium (no THI/THP, curve B) shift as a result of the addition of thermodynamic hydrate inhibitors (THIs, curve A) or thermodynamic hydrate promoters (THPs, curve C) (Majid et al., 2021).

The existing literature has classified some different categories of thermodynamic hydrate promoters depending upon their interactions with water and the structural contribution to the hydrate phase. Water-soluble promoters, including tetrahydrofuran (THF), propylene oxide, 1,4-dioxane, acetone, 1,3-dioxolane, and tetrahydropyran, are insoluble in water and are usually able to fill the large cages of sII hydrates, allowing hydrate formation to occur at much lower pressures (Majid et al., 2021; Wu et al., 2023). The water-insoluble promoters, such as cyclopentane, cyclohexane, methylcyclohexane, dimethylcyclohexane (DMCH), methylcyclopentane, neopentane, and similar cyclic compounds, are dissolved in water and act as promoters at the water-hydrocarbon interface, also forming sII hydrates with the guest gas in the small cages (Wu et al., 2023; Ho-Van et al., 2019). The structurally distinct group of semi-clathrate formers, including tetrabutylammonium and tetrabutylphosphonium salts, such as TBAB, TBAC, TBANO₃, TBAF, TBPB and isopropylamine, is fundamentally different from classical THPs: their ionic components partially replace the water molecules in the cage

lattice itself, acting simultaneously as a host and a guest (Majid et al., 2021). A fourth type is gas-phase promoters, where heavier hydrocarbon gases, for instance ethane in binary methane-ethane systems, or propane in ternary methane-ethane-propane systems, act as thermodynamic promoters of methane hydrate formation by preferentially occupying the large sII cages.

Although there has been a significant amount of experimental literature describing the behavior of individual promoters or small subsets thereof, a comparative assessment of the thermodynamic performance of all four categories based on a unified computational framework has previously not been reported. The literature contains a significant amount of phase equilibrium measurements of individual promoter-guest gas mixtures under limited concentration and pressure-temperature conditions, and uses either classical thermodynamic models or empirical correlations to analyze the data (Atta et al., 2025). Such methods can provide useful insights about single systems, but they are ill-suited to cross-category comparison and do not utilize the predictive benefits that modern machine learning methods can provide.

The development of machine learning as a phase behavior modeling tool for gas hydrate is a methodological advancement in the area. Algorithms based on artificial neural networks, support vector regression, random forest, and gradient boosting methods have been reviewed and found to exhibit high predictive accuracy for hydrate formation temperature and pressure in a variety of systems (Atta et al., 2025). These numerically-based methods are especially well-adapted to the modelling of nonlinear, multivariate, and complex relationships between inputs to the system - gas composition, promoter concentration, and thermophysical promoter properties - and the thermodynamic output of equilibrium pressure. Furthermore, the availability of large, curated datasets compiled from the published experimental literature provides a robust foundation for training and validating such models across diverse promoter categories.

1.2 Objectives of the Thesis

1.2.1 Main Objectives

The main objective of this thesis is to conduct a data-driven comparative evaluation of the thermodynamic performance of water-soluble, water-insoluble, and semi-clathrate hydrate promoters through the construction, training, and assessment of machine learning models

capable of predicting the hydrate equilibrium pressure for multicomponent gas-promoter systems.

1.2.2 Specific Objectives

In pursuit of the main objective, the following specific objectives were established:

1. To systematically collect, compile, and preprocess published experimental phase equilibrium data for hydrate systems containing water-soluble, water-insoluble, semi-clathrate, and gas-phase thermodynamic promoters, spanning a range of guest gases, promoter concentrations, and pressure-temperature conditions.
2. To engineer physically and chemically meaningful input features describing the thermodynamic and structural properties of the promoter compounds, enabling the models to generalize across chemically diverse promoter species within each category.
3. To develop and train machine learning regression models - including Random Forest (RF), CatBoost, XGBoost, and Support Vector Regression (SVR) - for each promoter category, targeting the prediction of hydrate equilibrium pressure (MPa) as the output variable.
4. To evaluate and compare model performance across promoter categories using standard regression metrics, and to validate model generalizability by assessing predictive accuracy on phase equilibrium data not included in the training dataset.
5. To interpret the comparative model performance across categories as a quantitative indicator of each category's thermodynamic coherence and predictability, thereby providing a data-informed basis for ranking and discussing the relative effectiveness of each promoter class.

1.3 Scope of Work

This thesis is exclusively limited to thermodynamic hydrate promoters; kinetic hydrate promoters and thermodynamic inhibitors are outside its scope. The analysis encompasses four promoter categories: water-soluble cyclic ethers and related polar compounds (THF, propylene oxide, 1,4-dioxane, acetone, 1,3-dioxolane, and tetrahydropyran), water-insoluble cyclic hydrocarbons and their derivatives (cyclopentane, cyclohexane, methylcyclohexane, dimethylcyclohexane, methylcyclopentane, neopentane, and cyclopentane derivatives), semi-clathrate formers based on quaternary ammonium and phosphonium salts (TBAB, TBAC,

TBANO₃, TBAF, TBPB, and isopropylamine), and gas-phase promoters comprising binary and ternary hydrocarbon gas mixtures in which ethane or propane serve as thermodynamic co-promoters for methane hydrate formation.

The guest gases considered across these systems include methane, carbon dioxide, hydrogen, nitrogen, oxygen, ethane, propane, krypton, and monofluoromethane (CH₃F), as present in the compiled experimental datasets. All data were sourced exclusively from the peer-reviewed experimental literature; no independent experiments were conducted in this work. The machine learning models developed in this work are trained and evaluated as predictive tools for equilibrium pressure as a function of temperature, gas composition, promoter identity, and promoter concentration. The scope does not extend to kinetic modeling, molecular simulation, or process-scale techno-economic assessment.

2. LITERATURE REVIEW

2.1 Overview of Thermodynamic Hydrate Promoter Categories

The basic characteristics of gas hydrates, such as crystal structure, phase equilibria, thermodynamic stability, and applications in industry, were discussed in Chapter 1. This chapter is dedicated specifically to an overview of both experimental and theoretical literature about thermodynamic hydrate promoters, divided into four classes of interest for this thesis: water-soluble, water-insoluble, semi-clathrates, and gas promoters. For each class, their mechanism of action, relevant benchmark molecules, the scope of guest gases, and the key results from quantification of phase equilibria are discussed. The chapter finishes with a brief discussion of the role played by machine learning in gas hydrate research and the gap addressed by this thesis.

2.2 Water-Soluble Thermodynamic Promoters

2.2.1 Category Overview and Promotion Mechanism

In terms of thermodynamic promoters for hydrates, water-soluble promoters refer to organic solutes that dissolve well in water across the range of concentrations relevant to hydrate formation and have a direct role in hydrate lattice formation through occupation of the large (5¹²6⁴) cage volume. In terms of their thermodynamic effects, the presence of these promoters results in a rightward shift in the hydrate equilibrium curve, towards lower pressures and higher temperatures relative to the gas-water system, with the extent depending on the concentration

of the promoter used, its molecular geometry, and the identity of the guest gas (de Deugd et al., 2001; Seo et al., 2001). In terms of thermodynamic promotion, the theoretical optimum corresponds to the stoichiometric concentration where the promoter saturates all the large cage cavities, with any excess not providing further promotion (Zhong et al., 2014).

In terms of the molecular factors affecting the efficacy of thermodynamic promotion, the most important are ring size, number of oxygen atoms in the molecule, and overall molecular symmetry. In terms of ring size and oxygen atoms, five-ring compounds containing a single oxygen atom, especially THF, produce the strongest promotion, because of their superior geometry with the large cages and high dipole interaction with the cage's walls. Meanwhile, six-ring compounds and those with more than one oxygen atom produce weaker thermodynamic effects, due to lattice strain caused by their higher polarities (Wu et al., 2023; Seo et al., 2001).

2.2.2 Tetrahydrofuran - Benchmark Compound

Tetrahydrofuran (THF, C_4H_8O) has received the largest number of investigations concerning a thermodynamic promoter of hydrate formation in water and acts as a category representative (Wu et al., 2023). THF leads to the formation of the sII hydrate at nearly ambient pressure levels, with its stoichiometry being close to the 1:17 THF/water molar ratio, equal to 5.56 mol%. The presence of THF alters the conditions of methane hydrate equilibrium, shifting the temperature by 4-20% at a constant pressure level, with higher alterations occurring at a lower pressure (Majid et al., 2021). In the case of CO_2 hydrates, Delahaye et al. (2005) reported shifts of 2-20% at a constant pressure between 0.2 and 3.5 MPa. Lee et al. (2012) provided extensive datasets for three different THF systems, including THF + CH_4 , THF + CO_2 , and THF + CO_2/CH_4 mixtures. The presence of sII structure in the $CO_2/CH_4/THF$ system was confirmed using the Clausius-Clapeyron method ($\Delta H_d = 57.2 \pm 5.72$ kJ/mol). Moreover, it was revealed that the promotion effect saturates above $x_{THF} = 0.03$, meaning that higher concentrations do not provide additional benefits. Further investigations have shown that THF could promote hydrate formation in a saline solution (Kumar et al., 2018; Mech & Sangwai, 2014), in hydrogen storage (Hashimoto et al., 2007; Strobel et al., 2009), and even in mixed gases concerning pre-combustion CO_2 capture processes (Lee et al., 2010; Kang et al., 2001; Park et al., 2013). However, the use of THF is associated with significant problems due to its classification as a suspected carcinogen and high volatility (Wu et al., 2023).

The THF-promotion dependency on concentration warrants special consideration when building machine learning models based on experimental data. The available phase equilibrium data covers the THF mole fraction range from as little as 0.3 mol% to up to 10 mol%, which means that sub-stoichiometric as well as super-stoichiometric concentrations are observed. The maximum promotion effect takes place at concentrations close to the stoichiometric value of around 5.56 mol%. In terms of equilibrium pressure depression, it can be said that within the given range, the relationship with THF concentration is mostly monotonic, with higher THF concentrations being accompanied by lower equilibrium pressures until the stoichiometric threshold, after which the system stabilizes. It should be noted that the monotonicity is violated at higher concentrations due to the presence of excess THF in the liquid phase, which leads to non-ideal conditions and an associated pressure plateau (de Deugd et al., 2001; Zhong et al., 2014). Moreover, introducing THF into the system results in a structural transition from sI to sII, which disrupts the otherwise monotonic relationship between the equilibrium pressure and temperature (Lee et al., 2012). All considered together, these factors lead to a rich but nonlinear thermodynamic model of the system, making it an interesting object for machine learning regressions.

2.2.3 Other Water-Soluble Promoters

Table 1 summarizes the key phase equilibrium characteristics of the remaining water-soluble promoters studied in this thesis alongside THF for comparison. These compounds - propylene oxide, 1,4-dioxane, acetone, tetrahydropyran (THP), and 1,3-dioxolane - have all been confirmed as thermodynamic promoters for methane and/or nitrogen hydrates, though with generally weaker promotion effects than THF, consistent with structural principles discussed above.

Table 1. Summary of water-soluble thermodynamic promoter phase equilibrium studies.

Promoter	Formula	Guest gases	Conc. range (mol%)	P range (MPa)	T range (K)	Hydrate Structure Formed	Key reference(s)
THF	C ₄ H ₈ O	CH ₄ , CO ₂ , N ₂ , H ₂ , H ₂ S, CO ₂ /CH ₄ ,	0.3–10	0.1–14	275–309	sII	Lee et al. (2012); Delahaye et al. (2005);

		CO ₂ /N ₂ , CO ₂ /H ₂					Zhong et al. (2014)
Propylene oxide	C ₃ H ₆ O	CH ₄ , N ₂ , CO ₂	3	0.9– 88	275– 294	sII	Seo et al. (2001); Seo et al. (2008)
1,4- Dioxane	C ₄ H ₈ O ₂	CH ₄ , N ₂ , CO ₂	3	17– 118	275– 293	sII	Seo et al. (2001); Seo et al. (2008)
Acetone	C ₃ H ₆ O	CH ₄ , N ₂	3	19– 114	274– 289	sII	Seo et al. (2001)
THP	C ₅ H ₁₀ O	CH ₄	1	2.1– 14.1	288– 302	sII	de Deugd et al. (2001)
1,3- Dioxolane	C ₃ H ₆ O ₂	CH ₄ , CO ₂	1–20	0.1– 14	275– 299	sII	de Deugd et al. (2001); Torré et al. (2015)

The most systematic multi-compound comparison of the CH₄ and N₂ systems was made by Seo et al. (2001), which revealed a clear performance hierarchy following molecular size and polarity principles. The same ranking was observed by Seo et al. (2008) when comparing different systems of CO₂ hydrates in THF, propylene oxide, and 1,4-dioxane with regard to carbon capture applications. According to Torr   et al. (2015), 1,3-dioxolane proved itself to be able to promote CO₂ hydrates simultaneously as a thermodynamic and kinetic promoter; thus, this compound can be considered as a promising one for a combined approach despite the relatively weak thermodynamic effect compared with THF.

2.3 Water-Insoluble Thermodynamic Promoters

2.3.1 Category Overview and Promotion Mechanism

Hydrophobic organic compounds, primarily cyclic hydrocarbons, which do not dissolve in water and act as thermodynamic promoters via occupying the large cage cavities at the

hydrocarbon-aqueous liquid-liquid interface, constitute the group of water-insoluble thermodynamic promoters (Mooijer-van den Heuvel et al., 2000; Ho-Van et al., 2019). While the fact that these systems exhibit liquid-liquid interfacial structure makes their kinetics different from water-soluble systems, their fundamental mechanism of promoting hydrate formation stays the same. The majority of water-insoluble thermodynamic promoters have been investigated under excess conditions in experiments, meaning that there is an additional bulk hydrocarbon phase besides the aqueous and hydrate phases that ensures large-cage full occupation and, consequently, maximal promoting effect (Lv et al., 2013).

Formation of a certain hydrate structure depends on the molecule's size. The five-membered ring structures like cyclopentane form hydrates with the large ($5^{12}6^4$) cage of sII structure, while larger compounds such as methylcyclohexane and dimethylcyclohexane, that cannot fit the sII large cage, form hydrates in which the larger ($5^{12}6^8$) cage is stabilized by adding the water-insoluble promoters together with a small gas component, such as methane (Nakamura et al., 2003). Therefore, the structural heterogeneity leads to a broader spectrum of promotion levels compared to those in water-soluble systems, since nearly all the latter form sII hydrates. It must be noted that one of the key advantages of using water-insoluble thermodynamic promoters in hydrate formation lies in their relatively easy physical separation from the aqueous phase after the dissociation of hydrates (Ho-Van et al., 2019).

2.3.2 Cyclopentane - Benchmark Compound

Cyclopentane (CP, C_5H_{10}) is the most extensively studied compound among water-insoluble thermodynamic promoters and serves as its benchmark analogous to THF in the case of water-soluble ones (Wu et al., 2023; Ho-Van et al., 2019). CP reacts with water to form the sII structure of CP hydrate at near-ambient pressure with an equilibrium temperature equal to about 279.8–280.2 K at standard pressure and is essentially immiscible with water (its solubility is 0.156 g/L at 25°C (NIST Database)), thus allowing easy phase separation (Ho-Van et al., 2019). The initial phase equilibrium database of CP in pure water was created by Tohidi et al. (1997), while Sun et al. (2002) provided additional data about temperature shift caused by the presence of CP for the case of pure methane with 6–27% equilibrium temperature shifts in the pressure range of 0.2–7.3 MPa. Further work on CP was done by Zhang and Lee (2008), who considered the binary system consisting of CO_2 and CP and found equilibrium temperature shifts of 4–7% for CO_2 and additionally provided experimental data related to hydrogen storage. Chen et al. (2010) explored the case of CP in addition to methane in aqueous solution of NaCl,

thus proving the stability of sII structure in brine with rising pressure. Finally, Zheng et al. (2017) and Li et al. (2010) examined CP in carbon capture and desalination applications, respectively, whereas Herslund et al. (2014) showed the synergic effect of combining CP and THF for forming CO₂.

As compared with the THF dataset, the range of guest gases involved in the phase equilibrium dataset in the presence of CP is significantly larger. The guest gases analyzed in combination with CP were methane, carbon dioxide, nitrogen, hydrogen, oxygen, krypton, CH₃F (monofluoromethane), and CH₂F₂ (difluoromethane). All the data reported are obtained mostly using the condition of CP excess (Takeya & Ohmura, 2006; Takeya et al., 2007; Imai et al., 2005; Zhang & Lee, 2008). This diversity in terms of the type of guest gas leads to the fact that the range of absolute equilibrium pressures in the CP dataset is quite wide, starting from values below 0.5 MPa in CO₂ + CP systems at temperatures close to 280 K to values higher than 10 MPa in N₂ + CP systems at similar temperatures. This is caused by different properties of the investigated gas formers with regard to hydrate formation. In turn, from a machine learning standpoint, this results in heteroscedasticity that needs to be accounted for when preparing the data and evaluating model performance metrics. Another feature of the CP dataset is that all the measurements were performed under conditions of the excess of CP, implying that, in contrast to water-soluble promoters, concentration is not a free variable.

2.3.3 Other Water-Insoluble Promoters

Table 2 summarizes the phase equilibrium characteristics of the remaining water-insoluble promoters studied in this thesis.

Table 2. Summary of water-insoluble thermodynamic promoter phase equilibrium studies.

Summary of water-insoluble thermodynamic promoter phase equilibrium studies.

Promoter	Formula	Guest gases	Conditions	Structure	Key reference(s)
Cyclopentane	C ₅ H ₁₀	CH ₄ , CO ₂ , N ₂ , H ₂ , O ₂ , Kr,	Excess, 0.1– 16 MPa, 274–310 K	sII	Tohidi et al. (1997); Sun et al. (2002); Zhang & Lee (2008)

		CH ₃ F, C ₂ H ₆			
Cyclohexane	C ₆ H ₁₂	CH ₄ , N ₂ , C ₂ H ₆ , CO ₂	Excess, 0.7– 25 MPa, 274–292 K	sII/sH	Sun et al. (2002); Mohammadi & Richon (2011)
Methylcyclohexane	C ₇ H ₁₄	CH ₄ , N ₂ , C ₂ H ₆	Excess, 0.5– 11 MPa, 274–293 K	sH	Sun et al. (2002); Nakamura et al. (2003)
cis-DMCH	C ₈ H ₁₆	CH ₄	Excess, 0.5–6 MPa, 276– 294 K	sH	Nakamura et al. (2003)
Methylcyclopentane	C ₆ H ₁₂	CH ₄ , CO ₂	Excess, 0.4– 9.5 MPa, 274–292 K	sII	Lv et al. (2013); Mohammadi & Richon (2009)
Neopentane	C ₅ H ₁₂	CH ₄	Excess	sII	Tohidi et al. (1997)
CP derivatives	Various	CO ₂ , CH ₃ F, Kr, CH ₂ F ₂	Excess, various	sII	Matsumoto et al. (2014); Takeya et al. (2007)

Systematic multi-compound comparison for water-insoluble promoters was carried out by Lv et al. (2013), who established the hierarchy of water-insoluble promoters using methane with respect to phase equilibria caused by their effects. Differences in the thermodynamic effects were explained as the combined effect of the ring size, molecule polarity, and branching. Additional comparisons involving water-insoluble promoters in CO₂, N₂, and ethane systems were done by Mohammadi and Richon (2009, 2011). Conceptual understanding of the behavior of water-insoluble promoters in the context of sII and sH structures was formulated by Mooijer-van den Heuvel et al. (2000).

2.4 Semi-Clathrate Hydrate Promoters

2.4.1 Category Overview and Structural Distinction

Semi-clathrates feature a fundamental distinction from traditional clathrates, in which case the role of promoter ions is confined to occupying the cavities of already formed gas hydrate host lattice, while the ion molecules themselves become part of the host lattice structure with their ionic cores replacing water molecules in certain lattice sites (Majid et al., 2021; Muromachi et al., 2025). This results in the creation of novel crystal structures unrelated to typical sI, sII, and sH clathrates, with additional vacant lattice sites capable of hosting gas molecules under far less stringent thermodynamic conditions. In particular, TBAB shows equilibrium temperatures shift from pure gas+water composition of 1.5 to 60 K at fixed pressure, positioning semi-clathrate promoters among the most efficient known thermodynamic promoters (Majid et al., 2021).

An essential property of the current class is the fact that some semi-clathrate hydrate formers can produce two (or more) separate crystal phases dependent on stoichiometric composition, such as Type A (water-to-promoter molar ratio of ~26:1) and Type B (~38:1) in the TBAB system (Zhou et al., 2018). Phase multiplicity, coupled with a wide variety of ionic compounds used, leads to the higher inherent heterogeneity of the dataset as compared to the other two promoter classes.

2.4.2 TBAB - Benchmark Compound

Tetrabutylammonium bromide (TBAB, $[(C_4H_9)_4N]Br$) is the most thoroughly investigated compound of the category and acts as the benchmark member. The first comprehensive multi-gas experimental equilibrium dataset was published by Arjmandi et al. (2007), covering H_2 , CH_4 , N_2 , CO_2 , and natural gas in TBAB semi-clathrate hydrates at TBAB concentrations ranging from 0.3 to 4.4 mol% and pressures from 1.2 to 41.4 MPa. Li et al. (2007) and Sun and Sun (2010) analyzed the methane+TBAB combination at various concentrations. Lin et al. (2008) and Zhou et al. (2018) studied the CO_2 +TBAB system with dissociation enthalpy calculations and crystallography involved. Zang and Liang (2017) conducted research related to equilibrium conditions for CO_2/CH_4 binary mixtures in TBAB solutions applicable for natural gas purifications. Yang et al. (2015) considered the synergistic effect for TBAB+THF combinations in semi-clathrate formation. Hassan et al. (2021) performed synergistic mixture investigations specifically for methane storage. Yu et al. (2018) examined kinetic and morphology characteristics of CO_2/H_2 /TBAB.

2.4.3 Other Semi-Clathrate Promoters

Table 3 summarizes the phase equilibrium characteristics of the remaining semi-clathrate promoters studied in this thesis.

Table 3. Summary of semi-clathrate thermodynamic promoter phase equilibrium studies.

Promoter	Type	Guest gases	Conc. range (mol%)	P range (MPa)	T range (K)	Structure	Key reference(s)
TBAB	Quaternary ammonium bromide	H ₂ , CH ₄ , N ₂ , CO ₂ , natural gas, CO ₂ /C H ₄	0.3–4.4	1.2–41.4	279–298	Semi-clathrate (Type A & B)	Arjmandi et al. (2007); Sun & Sun (2010); Zhou et al. (2018)
TBAC	Quaternary ammonium chloride	CO ₂ , CH ₄ , CO ₂ /C H ₄	0.1–4.0	1.4–8.2	279–292	Semi-clathrate	Kim et al. (2015); Fan et al. (2013)
TBAF	Quaternary ammonium fluoride	CO ₂ , CH ₄ , CO ₂ /C H ₄ , flue gas	0.1–3.5	0.5–6.5	278–294	Semi-clathrate	Fan et al. (2009); Shi & Liang (2019); Xu et al. (2025)
TBANO ₃	Quaternary ammonium nitrate	CO ₂ , H ₂	0.3–3.0	1.5–23	279–293	Semi-clathrate	Babu et al. (2014)

TBPB	Quaternary phosphonium bromide	CO ₂ , CH ₄ , CO ₂ /CH ₄	0.1–3.0	0.8–8.9	278–293	Semi-clathrate	Ye & Zhang (2014); Li et al. (2017)
Isopropylamine	Primary amine	CH ₄	5.6–16.7	1.0–10.0	274–285	Semi-clathrate (structural transformation)	Lee et al. (2012)

Systematic comparisons involving halides were done by Fan et al. (2013), who reported on phase equilibrium data for TBAB, TBAC, and TBAF in the CO₂/CH₄ binary mixture with respect to thermodynamic effects. Significant differences related to the effect of anion size and charge density effects were demonstrated. High selectivity toward CO₂ vs CH₄ makes TBPB very efficient in CO₂ separation from shale gas, as demonstrated by Li et al. (2017, 2019). TBANO₃ was identified as a potential pre-combustion carbon capture promoter by Babu et al. (2014). Structural transformation behavior for isopropylamine semi-clathrate hydrates with methane co-guests was characterized by Lee et al. (2012). Most recent developments concerning semi-clathrates include papers by Mardomi et al. (2025) and Muromachi et al. (2025), which introduce new asymmetric quaternary ammonium-based semi-clathrate promoters.

2.5 Gas-Phase Thermodynamic Promoters

Promoters in the gas-phase thermodynamic category do not involve any additional compound being added into the system; rather, the promotion effect is derived from the chemical composition of the gas phase itself. Hydrocarbons in heavier weight classes are naturally present in the natural gas streams; the presence of ethane and propane molecules in natural gas favors their preferential accommodation in the large cage (5¹²6⁴) of sII hydrate, leading to mixed sII hydrate formation at much milder conditions than when pure methane hydrate forms (Majid et al., 2021). Methane alone tends to form sI hydrate structure, whereas a small addition of propane or ethane triggers structural transition to the stable sII hydrate state, shifting the P-T equilibrium curve significantly to the low-pressure/high-temperature domain (Holder & Hand, 1982; Nixdorf & Oellrich, 1997).

In the binary systems ($\text{CH}_4/\text{C}_2\text{H}_6$), ethane plays the role of the sII large-cage occupant and methane promoter. Holder and Hand (1982) provided the fundamental dataset on multiple-phase equilibria in the binary $\text{CH}_4/\text{C}_2\text{H}_6$ system. De Menezes et al. (2019) studied the co-existence of sI and sII hydrate in the $\text{CH}_4\text{-C}_3\text{H}_8$ systems, demonstrating the structural complexity resulting from the competition between the two structures as a function of gas composition. For the ternary systems ($\text{CH}_4/\text{C}_2\text{H}_6/\text{C}_3\text{H}_8$), propane with a relatively larger diameter (approximately 6.28 Å) serves as the optimal sII large-cage occupant and is the major promoter; in addition, the presence of propane leads to much greater equilibrium shifts compared to only ethane (Adisasmito & Sloan, 1992; Nixdorf & Oellrich, 1997).

Although gas-phase promoters are important from a practical perspective and are relevant to flow assurance and natural gas processing operations, there are notable difficulties associated with data-driven modeling in this category. First, in contrast to additive promoters, there is no single, clear compound playing the role of the promoter in gas-phase systems. Therefore, molecular-level feature engineering becomes significantly harder. Second, the competition between the two structural configurations (sI and sII) results in non-linear equilibrium behavior. Regression techniques have trouble capturing such non-linearity without enough data.

2.6 Comparative Overview of Promoter Categories

The four promoter categories reviewed in this chapter differ not only in their chemical nature and structural mechanisms but also in the characteristic thermodynamic conditions they produce, the range of guest gases with which they have been studied, and the nature of the phase equilibrium datasets available for modeling. Table 4 provides a consolidated cross-category comparison that directly contextualizes the data-driven evaluation undertaken in this thesis.

Table 4. Cross-category comparison of thermodynamic hydrate promoter characteristics.

Characteristic	Water-soluble	Water-insoluble	Semi-clathrate	Gas-phase
Representative compound	THF	Cyclopentane	TBAB	Ethane/Propane

Hydrate structure formed	sII	sII (sH for larger compounds)	Semi-clathrate (unique)	sII
Typical P reduction vs. pure water	60–95%	40–80%	70–99%	20–60%
Typical T increase vs. pure water	4–20%	6–27%	10–60 K absolute	2–20%
Promoter concentration as variable	Yes	No (excess conditions)	Yes	No (gas composition)
Water miscibility	Full	Immiscible	Full (ionic solution)	N/A (gas phase)
Dataset heterogeneity	Moderate	High (multi-structure)	High (multi-phase types)	High (structural competition)
Primary guest gases in literature	CH ₄ , CO ₂ , H ₂ , N ₂	CH ₄ , CO ₂ , H ₂ , N ₂ , O ₂	CH ₄ , CO ₂ , H ₂ , N ₂	CH ₄ dominant

2.7 Machine Learning in Gas Hydrate Phase Equilibrium Prediction

2.7.1 Traditional Thermodynamic Modeling Approaches and Their Limitations

Before delving into the discussion on ML algorithms' use in this study, let us provide a brief overview of the classical thermodynamic modeling methods used historically for predicting the gas hydrate phase equilibria, since their drawbacks explain the need for the data-driven technique adopted in this research.

The most applied theoretical background for hydrate phase equilibrium modeling is van der Waals-Platteeuw (vdWP) statistical thermodynamic theory. In this theory, the formation of hydrates is represented as a gas molecule adsorption process into ideal cages, where the chemical potential of the hydrate phase depends on the degree of occupancy of each type of cage (Sloan & Koh, 2008, as cited in Majid et al., 2021). Together with an EOS for fluid phases and activity coefficient models for the aqueous phase, the vdWP model allows for physically correct calculation of three-phase equilibrium at different conditions. Commercial software packages such as CSMHyd and Multiflash are successfully applied in the oil and gas sector for estimating flow assurance conditions associated with hydrate inhibition (Majid et al., 2021).

Nevertheless, the vdWP theory is limited in applicability to promotion due to multiple reasons, as follows. Firstly, vdWP needs compound-specific Langmuir constants for each guest-cage interaction; thus, for any new promoter, there is required a fitting procedure based on experimental data. This makes the application of vdWP complicated and impractical since the objective of this research is to generalize for large amounts of diverse promoters from various classes (Atta et al., 2025). Secondly, classical vdWP does not include the structural features of semi-clathrates with a modified host lattice due to the presence of the ionic promoter. Thus, its application in the case of semi-clathrates is incorrect and requires structural modifications, which adds complications (Majid et al., 2021). Thirdly, EOS approaches suffer from their incapacity to describe mixed gases consisting of multiple components at various promoter concentrations, in which cases, nonlinearity and complexity of guest-guest and guest-host interaction led to unpredictable behavior of the system (Atta et al., 2025). The above-discussed issues generate a clear need to apply data-driven techniques for studying hydrate inhibition, which do not require the introduction of physically sound parameters but can learn from input-output relationships instead.

2.7.2 Overview of ML Landscape

Several recent studies have shown that the application of modern machine learning (ML) techniques can provide an invaluable contribution to the modeling efforts in the domain of gas hydrates thermodynamics, allowing the extraction of the underlying complex nonlinear multi-variate relationships from experimental data, without the need for prior knowledge about the nature of the physical process and the appropriate modeling structure (Talaei Khoei & Kaabouch, 2023; Atta et al., 2025). As evidenced by the systematic bibliometric review conducted by Atta et al. (2025), in their analysis of the existing literature on ML techniques

applied in the domain of hydrate thermodynamics between 2016 and 2025, the most common machine learning algorithms used in such work include artificial neural networks (ANN), support vector regression (SVR), and random forest (RF). Additionally, recent developments have seen a growing popularity of gradient boosting methods, in particular the popular XGBoost and CatBoost algorithms. In terms of the target variables, by far the most frequent one is HFT (hydration formation temperature), and the most commonly used accuracy metrics are R^2 and RMSE. According to the same review, in approximately 73.9% of cases, the datasets were literature-based experimental, with a strong emphasis on binary water-gas systems with or without a promoter or an inhibitor (Atta et al., 2025).

2.7.2 Previous Machine Learning Studies on Promoter-Assisted Hydrate Systems

Despite the continuously increasing research on gas hydrate phase equilibrium using machine learning in the literature, the investigation of thermodynamically promoted systems is still rather scarce. According to the findings presented by Atta et al. (2025), only 4.3% of the ML hydrate literature reviewed dealt with the topic of promoters, with most of the remaining articles being devoted to pure gas-water systems and inhibitor-rich systems. However, the limited number of works on the topic of ML promoters already provides valuable benchmarks for the current research thesis.

One of the earliest and most relevant works on predicting hydrate dissociation is that of Eslamimanesh et al. (2011) on the development of a hybrid Group Contribution–Least Squares Support Vector Machine (GC-LSSVM) model to predict hydrate dissociation pressure for structure H methane hydrates in the presence of 21 water-insoluble hydrocarbon promoters. Eslamimanesh et al.'s (2011) work is especially interesting in terms of relevance to the current thesis due to its pioneering nature in using group contribution parameters for molecular structure description in combination with machine learning. The GC-LSSVM model developed in this study showed high accuracy with an AARD of $\sim 1.6\%$ and $R^2 \approx 0.99$ based on a dataset of approximately 298 experimental points corresponding to different sH-forming compounds. The application of group contribution parameters allowed Eslamimanesh et al. (2011) to build a universal model that generalized over multiple chemically different water-insoluble compounds as opposed to using categorical promoter labels. Despite its relevance, Eslamimanesh et al.'s (2011) work covered only sH-forming hydrates and did not consider water-soluble or semi-clathrate-forming promoters.

The most advanced and complete ML model of hydrate promoters published before the present work was described by Ahmadi et al. (2020). In their work, Ahmadi et al. considered five thermodynamic promoters classified into three classes – water-soluble (THF and acetone), water-insoluble (cyclopentane), and semi-clathrates (TBAB and TBAC) – and predicted equilibrium dissociation pressure in CO₂ hydrates for each type of promoter. For the comparison, several ML algorithms – Adaptive Neuro-Fuzzy Inference System (ANFIS), Artificial Neural Network (ANN), Hybrid Fuzzy Genetic Algorithm (HFGA), and Ordinary Kriging, along with the classical van der Waals-Platteeuw thermodynamic model, were compared. The best performance in terms of prediction accuracy was shown by ANFIS, with $R^2 \approx 0.99$, which was not only superior to the performance of other machine learning algorithms, but also surpassed the classical van der Waals-Platteeuw model. Thus, this result proves that it is indeed possible to use data-driven approaches to obtain reliable and computationally efficient predictions of promoters. However, there are certain restrictions concerning the generalizability of results – Ahmadi et al.'s (2020) study is restricted to CO₂ as the guest gas and does not include the analysis of molecular structure in promoters, limiting its generalizability. The current thesis broadens the scope of Ahmadi et al.'s (2020) work in three ways: by introducing more guest gases, applying to all four promoter types, and using molecular features in the form of dipole moment, LogP, and molecular weight.

Another relevant study for our topic is that of Mesbah et al. (2017), which investigated the ability of Multi-Layer Perceptron Artificial Neural Network (MLP-ANN) to predict equilibrium dissociation pressure of semi-clathrate hydrates formed by CO₂ + H₂, CO₂ + CH₄, and CO₂ + N₂ mixtures in aqueous solutions of TBAB. The model obtained using MLP-ANN showed high precision with $R^2 \approx 0.996$ on a large number (179) of experimental points. While it is evident that ANNs are suitable tools to predict pressures in semi-clathrates, Mesbah et al.'s (2017) model cannot be generalized to other semi-clathrates (TBAC, TBAF, TBANO₃, TBPB, and isopropylamine) due to the limited scope of their work. This is a common problem of existing ML promoters' models – being trained on the data obtained for a particular promoter or even for a specific mixture of gases, they are unable to provide a comparative evaluation of the effect of various promoters, which is the goal of the current thesis. The current work includes six different semi-clathrate promoters with around 620 data points in total.

A study conducted by Ghiasi et al. (2016) applied Least Squares Support Vector Machine (LSSVM) to predict hydrate dissociation pressure in methane, ethane, propane, isobutane, CO₂, H₂S, N₂, and natural gas hydrate mixtures in pure water and in aqueous solution

containing thermodynamic inhibitors – salts and alcohols. Although this study does not target the prediction of promoter effects on hydrates, the high accuracy obtained using LSSVM on a very large dataset (over 3,900 experimental points) shows a great potential of kernel regression models in hydrate systems. The dataset considered by Ghiasi et al. (2016) is much larger and chemically more heterogeneous in comparison with other datasets in the literature; thus, it demonstrates the possibility of using LSSVM-based SVR models for complex datasets and emphasizes the need for further research on promoter systems. As mentioned above, this is especially relevant, taking into account the lack of datasets containing both inhibitors and promoters.

Collectively, it can be said that there is a certain trend in existing ML predictor studies of hydrate promoters: most of the studies are either too narrow-scope or are built with categorical descriptions of the chemical structure of molecules in mind. Table 5 shows the comparison between the studies reviewed above and the current work.

Table 5. Comparison of existing ML studies on promoter-assisted hydrate systems with the present thesis.

Study	Algorit hm(s)	Promoter category	No. of promo ters	Guest gases	Target	Dataset size	Molecular features
Eslamimanesh et al. (2011)	GC- LSSVM	Water- insoluble (sH)	21	CH ₄	P	298	Yes (group contributio n)
Ahmadi et al. (2020)	ANFIS, ANN, HFGA, Kriging	Water- soluble, water- insoluble, semi- clathrate	5	CO ₂ only	P	Not stated	No (categoric al)

Mesbah et al. (2017)	MLP- ANN	Semi- clathrate	1 (TBA B)	CO ₂ /H ₂ , CO ₂ /CH ₄ , CO ₂ /N ₂	P	179	No
Ghiasi et al. (2016)	LSSVM	Inhibitors (not promoters)	Multiple	CH ₄ , C ₂ H ₆ , CO ₂ , N ₂ , H ₂ S	T	>3900	No

The gaps identified in Table 5 directly motivate the methodological design of the present thesis. By training separate category-specific models using RF, XGBoost, CatBoost, and SVR across all four promoter categories simultaneously, incorporating continuous molecular descriptors as promoter features, and targeting equilibrium pressure across a diverse range of guest gases, the present work addresses the three principal limitations of the existing literature identified by Atta et al. (2025): the underrepresentation of HFP as a target variable, the absence of cross-category comparative evaluation, and the lack of molecular-level feature engineering for promoter characterization.

2.7.3 Research Gap and Justification for the Present Study

The present thesis addresses the following three gaps identified in the current literature on the application of machine learning to the domain of gas hydrate thermodynamics. Firstly, there is a significantly smaller number of works focusing on the prediction of hydrate formation pressure (HFP) as opposed to hydrate formation temperature (HFT), although in many real-life cases, HFP is more relevant than HFT due to limitations imposed by the available pressure conditions. For example, as indicated by Atta et al. (2025), only approximately 6 of the 24 papers reviewed by the authors targeted HFP prediction. Secondly, no existing study has used ML in a consistent fashion to compare different categories of thermodynamic promoters, instead considering a specific promoter, a promoter category, or an inhibitor-based system. Thirdly, no feature engineering approach has been applied to incorporate the physicochemical characteristics of the promoter molecule into the model's input. The thesis addresses all three gaps at once, as discussed below in detail.

3. METHODOLOGY

3.1 Data Collection and Compilation

3.1.1 Data Sources and Scope

The data used in the current thesis were gathered through a comprehensive search of published literature regarding experimental phase equilibrium studies. Most of the papers in which data were extracted belong to the references cited in Chapter 2. The study utilized no proprietary data or data from field or laboratory measurements, which means that the work is based on literature and follows the approach for data-based assessment described in Chapter 1. Experimental data points were obtained manually from tables and figures provided by published journal papers on four classes of promoters: water-soluble promoters, water-insoluble promoters, semi-clathrates, and gas phase promoters.

Data gathering involved collecting data corresponding to specific promoter compounds defined in the thesis scope and introduced in Chapter 2. In the case of water-soluble promoters, the selected molecules included THF, propylene oxide, 1,4-dioxane, acetone, tetrahydropyran (THP), and 1,3-dioxolane for various guest gases, such as methane, carbon dioxide, hydrogen, nitrogen, and mixed guest gases. In the case of water-insoluble promoters, data were collected for cyclopentane, cyclohexane, methylcyclohexane, dimethylcyclohexane (DMCH), methylcyclopentane, neopentane, and cyclopentane derivatives at excess promoter concentration. Semi-clathrate promoters comprised data from TBAB, TBAC, TBAF, TBANO₃, TBPB, and isopropylamine for guest gases, namely methane, carbon dioxide, hydrogen, nitrogen, and mixed systems. Gas-phase promoters included binary and ternary gas combinations of methane-ethane and methane-ethane-propane, where ethane and propane acted as the thermodynamic promoters.

3.1.2 Dataset Composition and Size

The total number of data points in the cleaned final dataset was roughly 700 for the water-soluble type, 450 for the water-insoluble type, 620 for the semi-clathrate type, and 150 for the gas phase type, adding up to roughly 1,900 data points in total. The number of data points present for each promoter compound is not equal due to the lack of experimental focus on some compounds compared to others in scientific literature — THF and TBAB, being the

benchmark compounds for the types they represent, are much more commonly observed than smaller molecules like THP or isopropylamine.

Table 6 summarizes the dataset composition by category.

Table 6. Summary of compiled datasets by promoter category.

Category	No. of promoter compounds	Approximate data points	Guest gases included
Water-soluble	7	700	CH ₄ , CO ₂ , H ₂ , N ₂ , O ₂ , mixed
Water-insoluble	6	450	CH ₄ , CO ₂ , H ₂ , N ₂ , O ₂ , Kr, CH ₃ F
Semi-clathrate	6	620	CH ₄ , CO ₂ , H ₂ , N ₂ , mixed
Gas-phase	2 gas systems	150	CH ₄ /C ₂ H ₆ , CH ₄ /C ₂ H ₆ /C ₃ H ₈
Total	21	1,920	

3.1.3 Raw Variable Structure

Before feature engineering, each data point in the dataset included the following raw variables:

- Temperature (T, K) – the temperature corresponding to the hydrate equilibrium condition used to determine the phase boundary;
- Pressure (P, MPa) – the pressure corresponding to the hydrate equilibrium condition and serving as the target variable in all machine learning algorithms developed in this thesis;
- Promoter concentration – measured as either weight percent (wt%) or mole percent (mol%), whichever was stated in the source; all values were converted to mol% before model training, using the molecular weights of the promoter substance and water;

- Gas composition – the mole fraction of each gas constituent in the system, reported in the source article.
- The gas-phase promoter class, wherein no additive is used, employed the promoter concentration variable as the mole fraction of the most abundant hydrocarbon compound in the gas mixture, namely C_2H_6 in binary methane/ethane hydrate systems and C_3H_8 in ternary methane/ethane/propane hydrate systems. This variable can serve as an approximation of the thermodynamic promotion effect exerted by the heavy hydrocarbon.

3.2 Data Preprocessing

3.2.1 Data Cleaning

Before the process of feature engineering and model training took place, a thorough cleaning process was conducted for all datasets to ensure data quality and consistency. Three main data cleaning processes were carried out.

Removal of duplicate data - Data points that were present in different research papers containing the same experimental data were found and removed, such that only one data point for each (T, P, concentration, gas mixture, promoter compound) combination was left. Duplicates usually occurred because of the presence of data points taken from reviews of primary experimental research.

Removal of outlier data points - Data points that had abnormal pressure and temperature readings, which did not follow physical expectations for the system, such as equilibrium pressures being abnormally high or low compared to the expected pressure trend given the promoter compound and gas mixture used, were found using P-T diagrams for all promoter compounds and then removed. This was done very carefully to avoid removing points that could naturally occur in experimental conditions.

Data points missing required information - Any data points where one or more of the required variables (T, P, concentration, gas mixture) could not be found in the source material and therefore could not be extracted with any precision were removed from the dataset. This happened rarely due to a lack of tabular data in the sources, while only graphs were provided.

3.2.2 Unit Standardization and Concentration Conversion

The concentrations of the promoters were given in various units in the source publications, such as mol%, wt%, and mole fractions. Concentrations were standardized to the unit of mol% based on the following relation:

$$x_{promoter}(mol\%) = \frac{\frac{m_{promoter}}{M_{promoter}}}{\frac{m_{promoter}}{M_{promoter}} + \frac{m_{water}}{M_{water}}} * 100$$

where $m_{promoter}$ and m_{water} are the masses of promoter and water, respectively, and $M_{promoter}$ and M_{water} are the corresponding molecular weights (18.015 g/mol for water). This conversion was applied to all wt% values prior to inclusion in the dataset.

In terms of the second dataset, the water-insoluble promoters, most experimental data available in the literature have been obtained in excess promoter conditions where an additional liquid hydrocarbon phase exists alongside the hydrate and aqueous phases. In this case, the concentration of the promoter cannot be considered as a free parameter. Accordingly, all data points with the excess condition have been disregarded in the concentration-based feature space. Thus, in the case of the water-insoluble dataset, promoter concentration is not included among the input variables; rather, the only information regarding the promoter is represented by its descriptors (see section 3.2.3).

3.2.3 Feature Engineering

Feature engineering involves designing informative input variables from raw data to make the machine learning model capable of discovering patterns and generalizing over a range of varying conditions within the system. In regard to the present work, the main issue related to feature engineering is how to represent the promoter's identity and physicochemical properties through continuous numerical input variables that can help discriminate between various promoters and allow generalization over unseen conditions.

The following features have been engineered for each observation in the dataset:

Temperature (T, K) - preserved as-is from the raw data as a continuous numerical input variable.

Promoter concentration (mol%) - preserved as-is from the raw data for the water-soluble and semi-clathrate categories of promoters. Not used in case of the water-insoluble category (excess conditions), see Section 3.2.2. In the case of the gas phase category, the mole fraction of the heaviest hydrocarbon was used as explained in Section 3.1.3.

Gas mole fractions - the mole fraction of each gas component present in the system was used as an additional input numerical variable. Multi-component gas systems had their own individual gas component columns, while single-component gas systems were assumed to be composed of only one gas former, whose mole fraction was taken as 1.0.

Molecular weight of the promoter (MW, g/mol) - this feature was assigned to each promoter compound as a constant value corresponding to its molecular weight. Molecular weights have been obtained using the NIST Chemistry WebBook database and its equivalent databases. This feature represents the molecular size of the promoter.

Dipole moment of the promoter (D, Debye) - this feature was assigned to water-soluble promoter category compounds as a constant value, using its dipole moment value as provided by the NIST database and its equivalents. Dipole moment measures the polar nature of a given molecule. As explained in Chapter 2, the dipole moment of the guest molecules is an important factor affecting the energy of interaction within the hydrate lattice as well as lattice strain imposed by the promoter.

LogP (octanol-water partition coefficient) - this feature was assigned to each promoter compound in water-soluble and water-insoluble categories as a constant value obtained from the NIST database and its equivalent databases. LogP is a conventional hydrophobicity measure; positive values correspond to water-insoluble (hydrophobic) promoters, while negative values are characteristic of water-soluble (hydrophilic) promoters. This feature allows for quantitatively distinguishing between different promoter behaviors without the need to employ a categorical representation of the promoter category.

Additional features for the semi-clathrate category:

- Cation number — the number of alkyl chains attached to the central cation (nitrogen or phosphorus), reflecting the size and geometry of the cationic component

- Anion number — a numerical identifier encoding the identity of the counter-anion (bromide, chloride, fluoride, or nitrate), reflecting the influence of anion type on semi-clathrate lattice geometry and thermodynamic stability
- Carbon chain length — the number of carbon atoms in each alkyl chain of the cation, reflecting the steric contribution of the cationic component to the semi-clathrate host lattice

In relation to the gas phase promoter's classification, a binary variable was used to encode the two different gas phase systems investigated; a value of zero was assigned for a binary methane-ethane system (with ethane acting as the promoter), while a value of one was encoded for a ternary methane-ethane-propane system (with propane being the dominant promoter). The above classification helps capture the structural difference between the two gas phase systems, namely sII promotion with ethane vs. sII promotion with propane.

Table 7 summarizes the complete feature set used for each promoter category.

Table 7. Input features used for each promoter category.

Feature	Water-soluble	Water-insoluble	Semi-clathrate	Gas-phase
Temperature (K)	✓	✓	✓	✓
Promoter concentration (mol%)	✓	✓	✓	✓ (heavy HC mol%)
Gas mole fractions	✓	✓	✓	✓
Molecular weight (g/mol)	✓	✓	✓	✓
Dipole moment (D)	✓	—	—	—
LogP	✓	✓	—	—
Cation number	—	—	✓	—
Anion number	—	—	✓	—
Carbon chain length	—	—	✓	—

Binary promoter type (0/1)	—	—	—	✓
----------------------------	---	---	---	---

3.2.4 Feature Scaling

The application of feature scaling as a pre-processing step before modeling was analyzed. Both scaling options, namely standardization (to zero mean and unit variance with StandardScaler) and no scaling, were compared for each category and all four algorithms. The results showed no apparent improvement in any of the selected performance metrics (R^2 , RMSE, AIC, BIC) when using scaled data compared to non-scaled data for any of the combinations of category-algorithm. This result aligns with the known characteristics of tree-based ensemble techniques (RF, XGBoost, CatBoost) that are scale-invariant since they do not assume any scale for the features used (Breiman, 2001; Prokhorenkova et al., 2018). In turn, for the SVR algorithm, which theoretically relies on scaled data because of the distance computations associated with the chosen kernel function (Smola & Schölkopf, 2004), the observed improvements were too minor to be included in the final pipeline. Therefore, feature scaling was abandoned for all algorithms and categories, and models were trained on non-scaled data only.

3.2.5 Feature Selection and Exploratory Data Analysis

Feature Selection

Before determining the selected input feature set, as indicated in Table 7, an expanded pool of candidate molecular and physicochemical descriptors was first generated for each category of promoter, based on a wider range of physical and structural information obtainable from databases. Such candidate inputs involved descriptors of ring structure (cyclic indicator, ring size), molecular geometry, as well as further ionic descriptors in the case of the semi-clathrate category. The need for generating such an expanded list of candidates before narrowing down to a more parsimonious feature set was to provide a thorough representation of the physicochemical nature of the promoters before conducting any selection process.

The inclusion of many features as inputs to a machine learning model, especially when the amount of data used for analysis is modest, poses challenges such as increasing model complexity, introducing noise, and inducing multicollinearity that negatively affect the training of the model (González et al., 2020). To overcome these issues, a methodical visual and statistical approach of feature selection was undertaken before training the model in each dataset using three diagnostic methods:

Pairplots - Pairwise scatterplots for all possible combinations of features were investigated to identify irrelevant features and determine whether there were any non-linear dependencies or clusters among the features.

Heatmaps of correlation values - Pearson's correlation coefficient for all features and the output variable were calculated and represented as a heatmap. This approach quantitatively identified redundant features (features that have a high correlation with one another) and irrelevant features (features that have near-zero correlation with pressure).

Based on these analyses, irrelevant features that had high multicollinearity with other relevant features ($|r| \geq 0.7$) and near-zero correlation with the target variable were excluded from the dataset. The results of this analysis are presented in Table 7. The selected features provide an adequate description without overfitting the model. Features such as cyclic indicators and ring size descriptors contained information already provided by molecular weight, lipophilicity, and dipole moment.

Exploratory Data Analysis

After feature selection, Pearson correlation matrices were calculated for each finalized category dataset to quantify linear relationships between input features and the output. Below is a summary of the results for each category as depicted in Figures 3 through 6.

Water-Soluble Promoters (Figure 3) – Temperature is the sole input variable with significant linear correlation with equilibrium pressure ($r \approx +0.34$), corresponding to the physical intuition that higher temperatures necessitate greater pressures for hydrate stability. The negative correlation between CO₂ mol% and pressure ($r \approx -0.25$) is due to the relatively low stability pressure of CO₂ hydrate compared to other gases such as methane and hydrogen at the same temperatures. All other variables, including molecular descriptors, display weak linear relationships with pressure ($|r| < 0.25$), implying that the pressure-input relationships are mainly nonlinear for this category. The presence of nonlinear relationships provides a strong rationale for selecting ensemble tree-based ML methods, as well as kernel algorithms, which are explicitly capable of modeling nonlinearity. As for the retained promoter descriptors, no problem of multicollinearity was detected between MW and LogP; therefore, the final feature set is valid after removing the ring and cyclic variables due to their redundancy ($r \approx +0.79$ and $+0.76$, respectively).

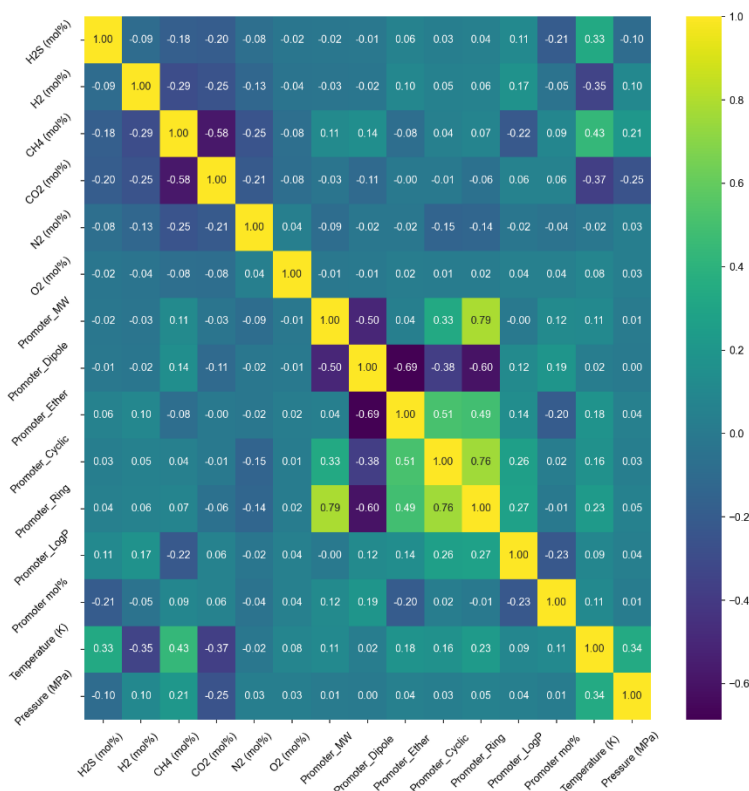


Figure 3. Pearson Correlation Heatmap for Water-soluble promoter dataset.

Water-Insoluble Promoters (Figure 4) – All correlations between input variables and equilibrium pressure are highly weak ($|r| \leq 0.21$), indicating almost nonexistent dominant linear determinants of pressure in this dataset. This is in stark contrast to physical intuition, as a negative weak correlation exists between temperature and pressure ($r \approx -0.21$). This is explained by the great structural heterogeneity of the water-insoluble category, which includes sII forming compounds (cyclopentane, cyclohexane, methylcyclopentane, neopentane) as well as sH forming molecules (methylcyclohexane, DMCH), both of which possess completely different P-T relations. As a result of the high multicollinearity of the initial structural descriptors between cyclic (ring $\approx +0.93$), ring vs. LogP ($\approx +0.74$), the step of feature selection proved to be necessary, as the existence of the three together would lead to significant redundancy. Upon removing the two highly-correlated descriptors, the remaining features (MW, LogP, dipole moment) are sufficiently uncorrelated to be considered independent inputs. With the complete lack of pressure-linear predictors as well as the structural heterogeneity, this category will pose the biggest challenge out of the four.

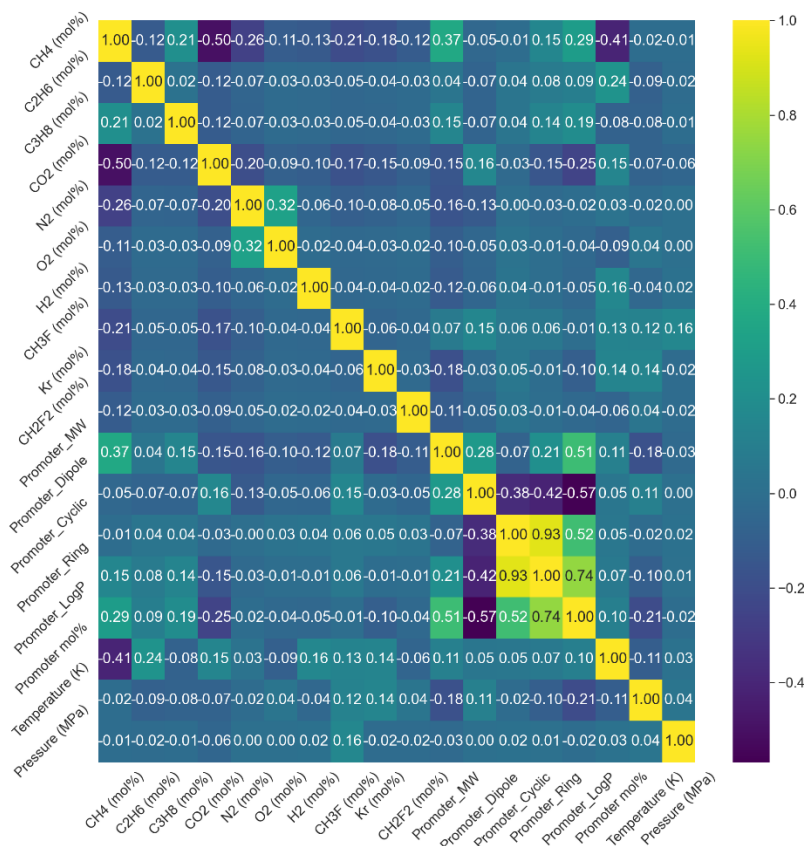


Figure 4. Pearson Correlation Heatmap for Water-insoluble promoter dataset.

Semi-Clathrate Promoters (Figure 5) – The semi-clathrate data is characterized by a highly significant negative correlation between temperature and equilibrium pressure ($r \approx -0.98$), which can be attributed to the inverse relationship between these two input variables in semi-clathrates. Specifically, the presence of a promoter causes a dramatic reduction of the equilibrium pressure of hydrates, resulting in a nearly deterministic linear relationship. This implies that in the semi-clathrate category, almost 96% of the linear variability of pressure is explained by temperature – thus, the ability of models to perform exceptionally well in this category can be expected. Regarding correlations between promoter descriptor inputs, MW and anion number display a significant correlation ($r \approx -0.80$) because various tetrabutylammonium salts have varying molecular weight depending on anion. In addition, carbon chain length and alkyl chain length are significantly correlated with each other ($r \approx +0.76$), corroborating redundancy leading to their removal during the process of feature selection. The chosen features for the semi-clathrate category – cation number, anion number, carbon chain length, MW,

dipole moment, LogP – effectively represent the structural properties of the compounds in this category.

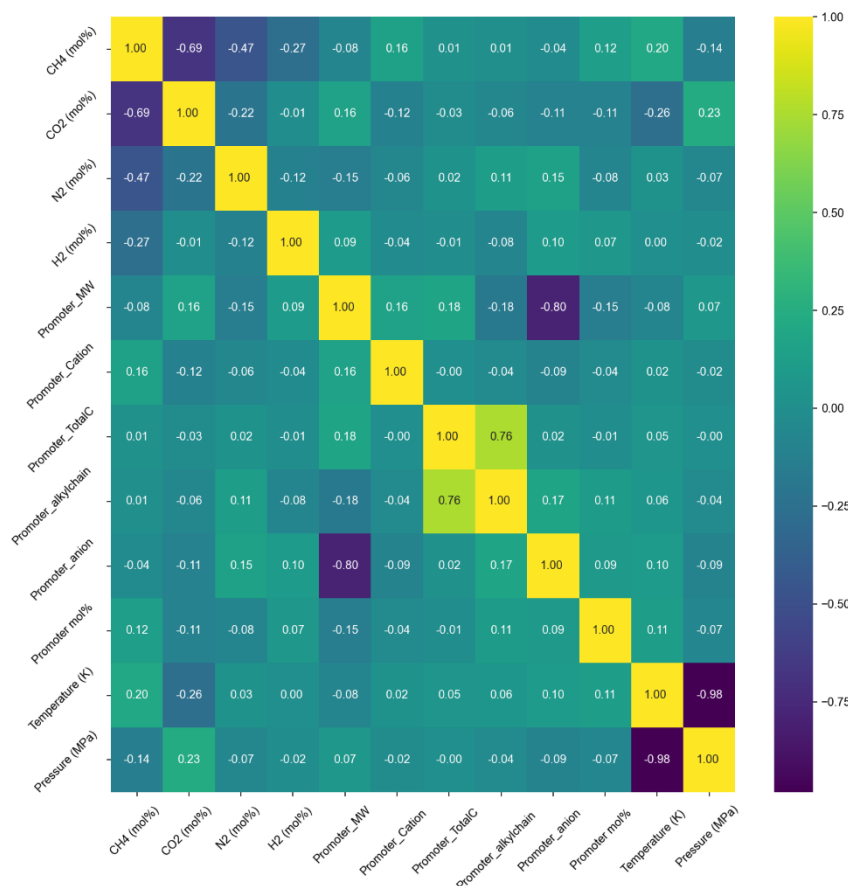


Figure 5. Pearson Correlation Heatmap for Semi-clathrate promoter dataset.

Gas Phase Promoters (Figure 6) – Gas phase is defined by its unique correlation matrix, which is primarily caused by the constant composition of the mixture. Indeed, temperature shows a significant positive correlation with pressure ($r \approx +0.70$), which follows the established physical characteristics of sII hydrate systems – pressure monotonically increases with temperature in the equilibrium curve. On the other hand, there exists a significant negative correlation between promoter mol% (heaviest hydrocarbon fraction) and equilibrium pressure ($r \approx -0.42$); this is explained by the promoter effect, which reduces the needed equilibrium pressure at a higher proportion of the promoter. What sets the gas phase category apart from the rest is its extremely high level of multicollinearity among gas composition variables and promoter mol%. For instance, there exist perfect correlations among $\text{CH}_4 \leftrightarrow \text{C}_2\text{H}_6$ ($r \approx -1.00$), $\text{CH}_4 \leftrightarrow \text{promoter mol\%}$ ($r \approx -1.00$), and $\text{C}_2\text{H}_6 \leftrightarrow \text{promoter mol\%}$ ($r \approx +1.00$). This is caused by

the requirement of total mole fractions of gases to equal unity – if the fraction of one gas species is known, the rest of them are completely determined. Therefore, several of the input variables offer no additional predictive value to the model due to the extreme collinearity; additionally, the effective dimensionality of the gas phase feature space is considerably lower than what could be inferred from the number of variables. Together with the small sample size (≈ 150 points), this characteristic of the gas phase feature space will be an important source of problems in chapters 4 and 5.

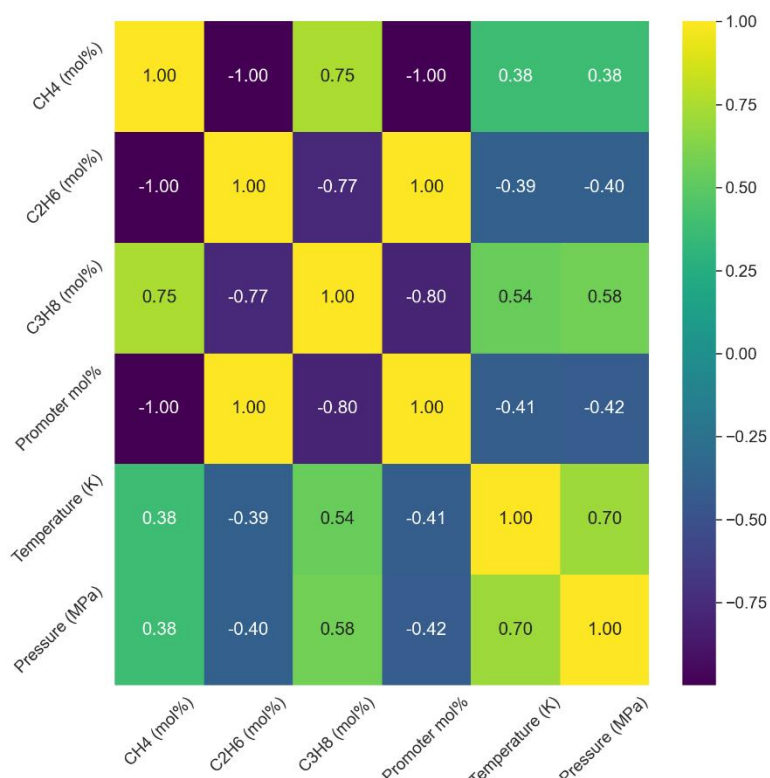


Figure 6. Pearson Correlation Heatmap for Gas promoters dataset.

3.3 Machine Learning Algorithms

In terms of machine learning algorithms, Random Forest (RF), Extreme Gradient Boosting (XGBoost), CatBoost, and Support Vector Regression (SVR) were used for this thesis. Their principles, objective functions, and hyperparameters are discussed in Chapter 2, Section 2.8.2 through 2.8.5. Therefore, the current section describes only implementation-specific details that are needed to describe the methodology of the current thesis.

Random Forest, XGBoost, CatBoost, and SVR were programmed and tested in Spyder, a Python 3.13 IDE. The following libraries were used for these implementations:

- scikit-learn (version 1.x) — for RF and SVR implementation, split into train/test datasets, cross-validation, and evaluation of the prediction error;
- xgboost — for XGBoost implementation;
- catboost — for CatBoost implementation;
- pandas and numpy — for dataset manipulation, preprocessing, and calculations;
- matplotlib and seaborn — for graphical representation of the obtained results.

The same algorithm implementations were used for all four categories of promoters. The only differences were in the selected feature set (see Table 7) and hyperparameters.

3.4 Model Training and Validation

3.4.1 Train-Test Split

For every category, the dataset was split randomly into two parts – training and hold-out test sets, using an 80/20 split ratio (80% training and 20% test set). The random split was done separately for each category with a predetermined seed for random numbers to make sure that results were reproducible. All training procedures and hyperparameter optimization were done only on the training set, and the test set was completely untouched until the end, when the test set metrics were calculated to determine final model performance.

3.4.2 Cross-Validation

To estimate model performance stability and mitigate the potential risk of overfitting during hyperparameter optimization, 5-fold cross-validation was employed for training of all models for all category-algorithm pairs. During 5-fold cross-validation, the training set is first randomly partitioned into 5 equal subsets (folds). Then, 4-folds are used as the training data while one fold is reserved as the validation data, and the whole process repeats 5 times (for 5 different combinations of training/validation splits), thus each fold is used as the validation fold exactly once. The mean and standard deviation of cross-validation scores across 5-folds give an unbiased estimation of expected model generalization ability and act as a crucial hyperparameter selection metric (González et al., 2020).

3.4.3 Hyperparameter Tuning

Hyperparameter optimization was done via the RandomizedSearchCV method implemented in the scikit-learn library. The algorithm randomly selects hyperparameters to sample from search space, instead of trying every possible combination (as in Grid Search),

which makes it computationally much cheaper to do when there is a need to explore many combinations, especially in the case of algorithms like XGBoost and CatBoost, which have lots of hyperparameters (González et al., 2020). RandomizedSearchCV was done on the training set by performing 5-fold cross-validation and choosing RMSE as the optimization criterion.

For every algorithm and category pair, the list of hyperparameters optimized was selected considering algorithm properties. For RF, these were `n_estimators`, `max_depth`, `min_samples_split`, `max_features`. For XGBoost, `n_estimators`, `eta`, `max_depth`, `subsample`, `colsample_bytree`. For CatBoost, `iterations`, `learning_rate`, `max_depth`, `L2_leaf_reg`. And for SVR, `C`, `ε`, `γ`. Final hyperparameter combinations for every category-algorithm pair were found in Chapter 4.

3.4.4 Model Evaluation Metrics

Model performance was estimated by using 4 metrics, calculated on the hold-out test set:

The coefficient of determination (R^2) measures the proportion of variance in the equilibrium pressure explained by the model:

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

The Root Mean Squared Error (RMSE) quantifies the average magnitude of prediction error in MPa:

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

The Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) are information-theoretic metrics that penalize model complexity — specifically the effective number of model parameters — alongside goodness of fit, providing a basis for comparing models of different complexity:

$$AIC = 2k - 2\ln(\hat{L})$$

$$BIC = k\ln(n) - 2\ln(\hat{L})$$

where k is the number of model parameters, n is the number of data points, and $\hat{L}$ is the maximized likelihood of the model. For regression models with normally distributed residuals, these simplify to:

$$AIC = n \ln \left(\frac{RSS}{n} \right) + 2k$$

$$BIC = n \ln \left(\frac{RSS}{n} \right) + k \ln(n)$$

where RSS is the residual sum of squares. Lower AIC and BIC values indicate better model performance relative to model complexity, with BIC imposing a stronger penalty on complexity than AIC for datasets with more than approximately eight data points (González et al., 2020). The inclusion of AIC and BIC alongside R^2 and RMSE addresses the known limitation of R^2 when used in isolation — specifically its inability to penalize model overfitting — and provides a more complete assessment of model quality, particularly when comparing algorithms of different complexity, such as SVR versus gradient boosting ensembles.

3.4.5 Internal and External Validation

Apart from the traditional train/test evaluation metrics, the external validation of the best-performing model for each category was additionally conducted through the analysis of the phase equilibrium data that was not used during the process of learning/training and test data assessment. These external validation data points were derived from the publications found among the general literature and containing phase equilibrium data for the studied promoter systems, but which did not enter the dataset that was used as input for learning the machine learning models. Thus, the external validation was meant to assess how the best-performing model can predict the equilibrium pressure for conditions and data sets that were never presented to the model in any form.

The best-performing model for each category (based on test set R^2 , RMSE, AIC, and BIC) was serialized with the help of the Python library pickle and saved as .pkl file. After serialization, the best-performing model for each category was loaded and applied to the internal and external validation data sets. The performance of internal validation was checked on the ability of the best model to understand the general trend and predictive ability. The performance of external validation was estimated in terms of R^2 , RMSE, and MAE:

$$MAE = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i|$$

Inclusion of MAE in addition to RMSE for the case of external validation is justified due to the physical sense of the problem – the prediction of equilibrium pressure in MPa, and because the MAE metric is less sensitive to extreme values of prediction error that can occur because of differences in the experimental technique that has produced the input dataset.

3.5 Comparative Evaluation Framework

Ultimately, this thesis aims to perform a comparative data-based analysis of thermodynamic promoter performance in each of the four categories. The comparative analysis includes two dimensions.

In terms of intra-category comparison, four models (RF, XGBoost, CatBoost, SVR) are comparatively analyzed against each other with respect to their performance according to R^2 , RMSE, AIC, and BIC metrics calculated from the test set. The algorithm with the highest aggregate score across all metrics becomes the best-performing model, used to calculate metrics for further comparative analysis (external validation, physical analysis).

In terms of inter-category comparison, performance metrics calculated for the best performing model in each category are comparatively analyzed among themselves with the aim of determining which category is more predictable from a thermodynamic perspective. The hypothesis underlying such comparison implies that the category characterized by higher physical consistency of its phase equilibrium behavior will demonstrate higher performance metrics, while the category with high thermodynamic variability or a limited amount of data, low ones. Such interpretation of machine learning algorithm performance metrics allows treating them as measures of thermodynamic regularity and reliability of datasets associated with a certain category – an approach, unprecedented in the hydrate promoter field and constituting a core value of this research.

4. RESULTS

4.1 Overview

The findings of the machine learning modeling analysis conducted in this thesis for all four classes of thermodynamic hydrate promoters are provided in this chapter. For each class of promoters, the performance on the holdout test set of all four algorithms used in this work – RF, XGBoost, CatBoost, and SVR – is given in terms of R^2 , RMSE, AIC, and BIC. The best performing model per class of promoters and its performance on the external validation set are

given. This chapter is organized based on the classes of promoters, starting with water-soluble, followed by water-insoluble, semi-clathrate, and gas-phase promoters. Comparative analyses between different classes of promoters are given in Chapter 5 (Discussion).

The numerical values provided in this chapter refer to the performance of the models on the holdout test set (20% of the entire dataset), unless otherwise stated. Hyperparameter tuning information of all models is presented in the tables presented in each section. Pressure prediction plots are shown for all models in each section.

4.2 Water-Soluble Promoters

4.2.1 Model Performance

Table 8 summarizes the test set performance metrics for all four algorithms applied to the water-soluble promoter dataset (700 data points, 7 promoter compounds, 80/20 split).

Table 8. Test set performance metrics for water-soluble promoter models.

Algorithm	R ²	RMSE (MPa)	AIC	BIC
Random Forest	0.87	1.16	62.2	91.9
XGBoost	0.87	1.17	63.0	93.5
CatBoost	0.92	0.91	47.6	77.2
SVR	0.85	1.26	86.2	115.8

Table 9 summarizes the hyperparameters that were used during testing and training of the ML models on the water-soluble dataset.

Table 9. Hyperparameters summary for the water-soluble category for all models

Algorithm	Hyperparameters	Optimal hyperparameters
Random Forest	n_estimators = [100, 200, 300]; max_features = [sqrt, None, auto]; max_depth = [2, 4, 6, 8, 10]; min_samples_split = [2]; min_samples_leaf = [1]	n_estimators = 200 max_features = 'sqrt' max_depth = 8 min_samples_split = 2 min_samples_leaf = 1
XGBoost	n_estimators = [100, 200, 300]; max_depth = [2, 3, 4, 5, 6]; learning_rate = [0.01, 0.1, 0.2, 0.3];	n_estimators = 200 max_depth = 4 learning_rate = 0.1

	min_child_weight = [2, 3, 4, 5, 6]; gamma = [0]; subsample = [1]; colsample_bytree = [1]; alpha = [0]; lambda = [1]	min_child_weight = 3 gamma = 0 subsample = 1 colsample_bytree = 1 alpha = 0 lambda = 1
CatBoost	'iterations' : randint(200, 1500), 'depth' : randint(3, 8), 'learning_rate' : loguniform(1e-3, 1e-1), 'l2_leaf_reg' : loguniform(1e0, 1e2), 'min_data_in_leaf' : randint(5, 30), 'subsample' : uniform(0.6, 0.4), 'colsample_bylevel' : uniform(0.6, 0.4), 'random_strength' : uniform(0.5, 4.5), 'bagging_temperature' : uniform(0, 2.0),	iterations = 864 depth = 6 learning_rate = 0.041 l2_leaf_reg = 7.35 min_data_in_leaf = 14 subsample = 0.83 colsample_bylevel = 0.74 random_strength = 2.61 bagging_temperature = 1.18
SVR	'svr__kernel': ['linear', 'rbf'], 'svr__C': loguniform(1e-1, 1e3), 'svr__epsilon': loguniform(1e-4, 1e-1), 'svr__gamma': ['scale', 'auto']	svr__kernel = 'rbf' svr__C = 42.7 svr__epsilon = 0.0063 svr__gamma = 'scale'

Figures 7-14 represent the scatter plots for all the ML models for the water-soluble dataset for training and testing.

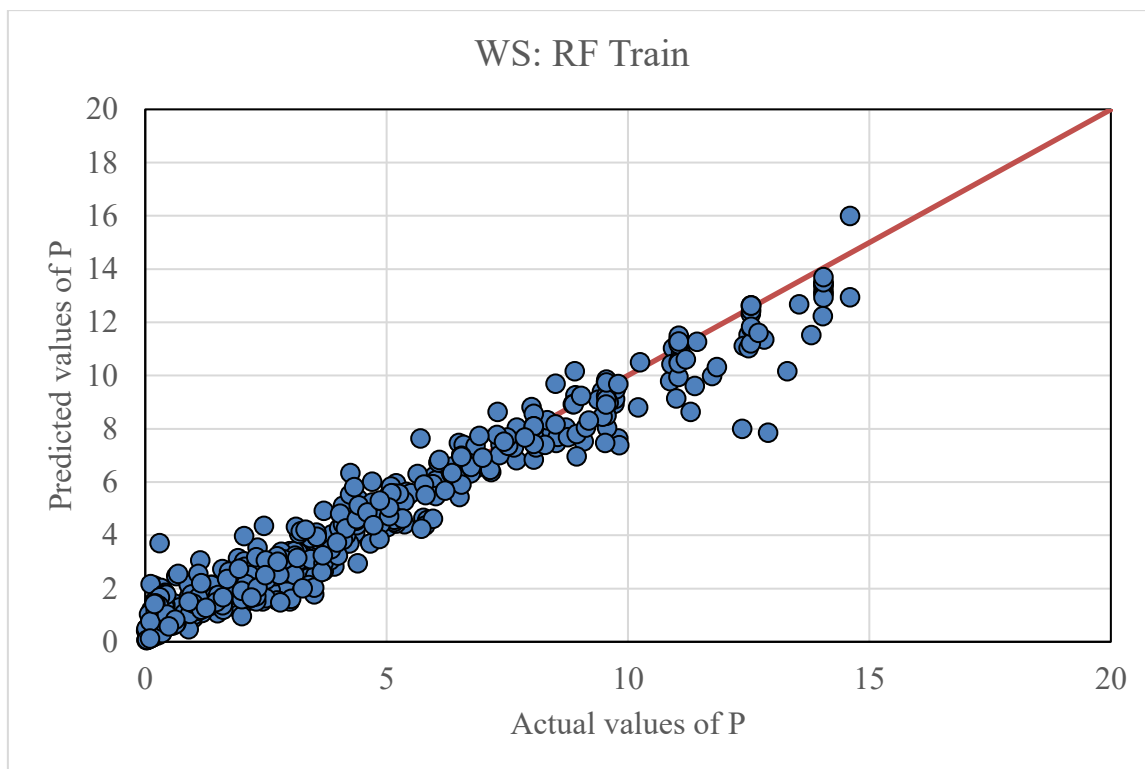


Figure 7. RF Scatter Plot for Water-soluble training dataset

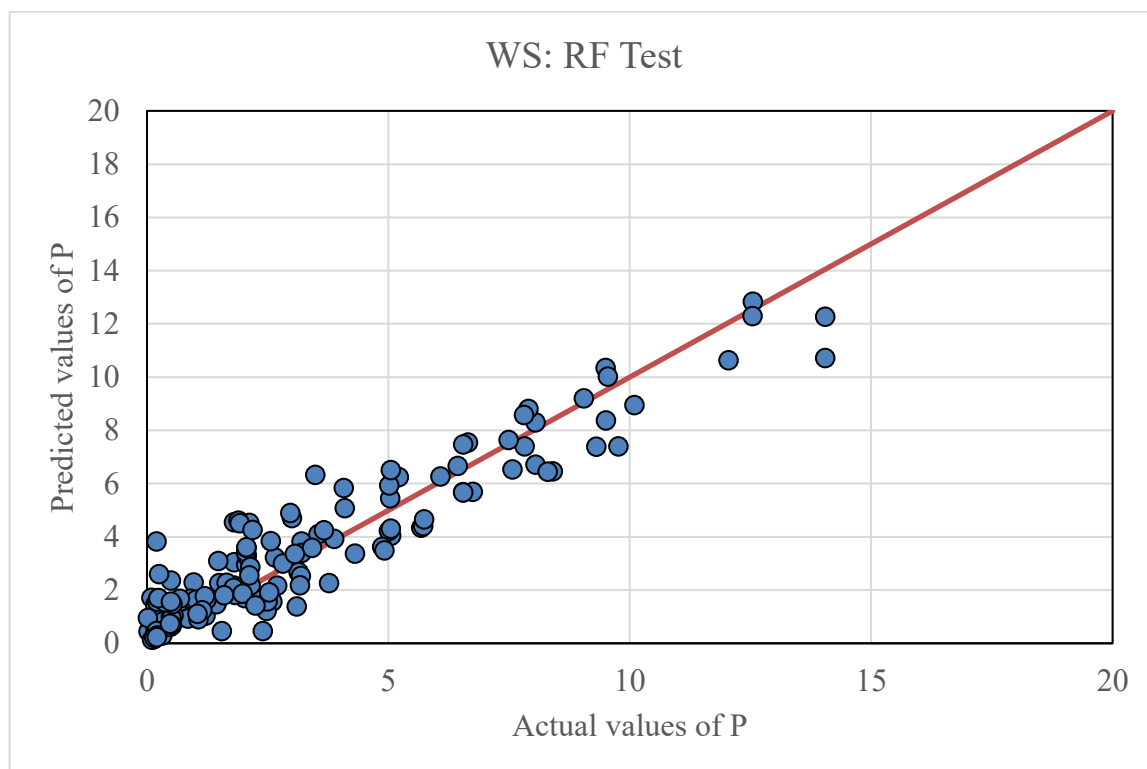


Figure 8. RF Scatter Plot for Water-soluble testing dataset.

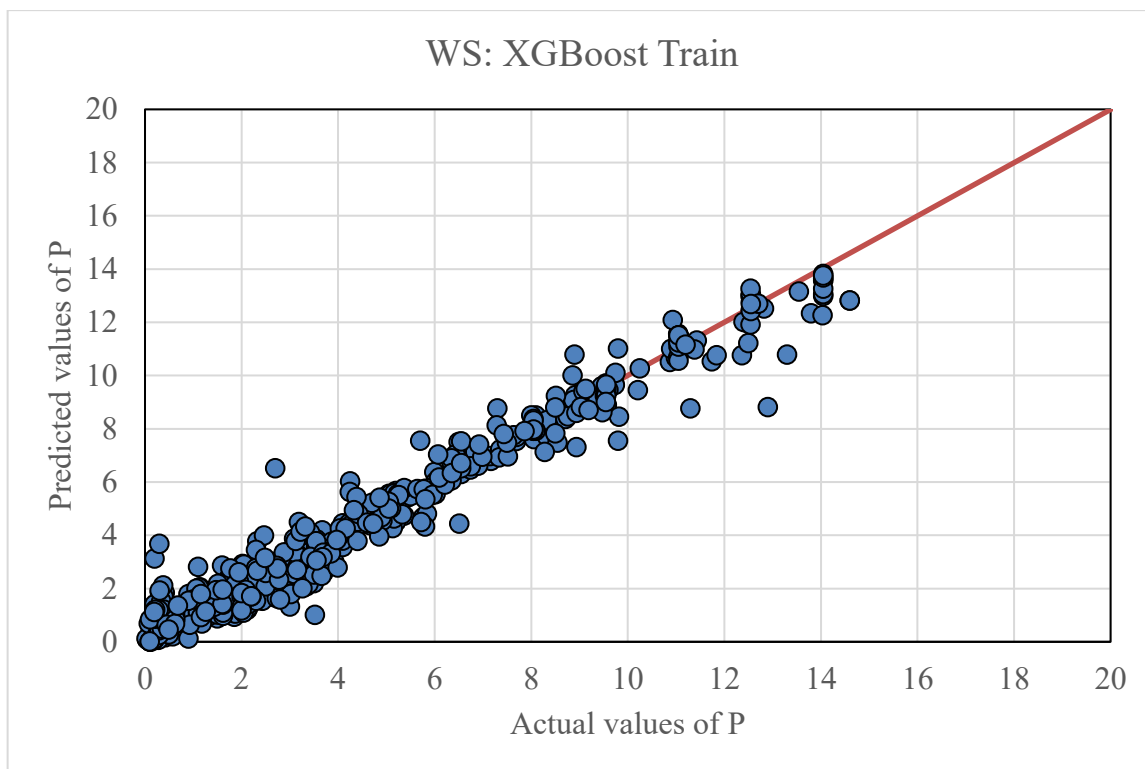


Figure 9. XGBoost Scatter Plot for Water-soluble training dataset.

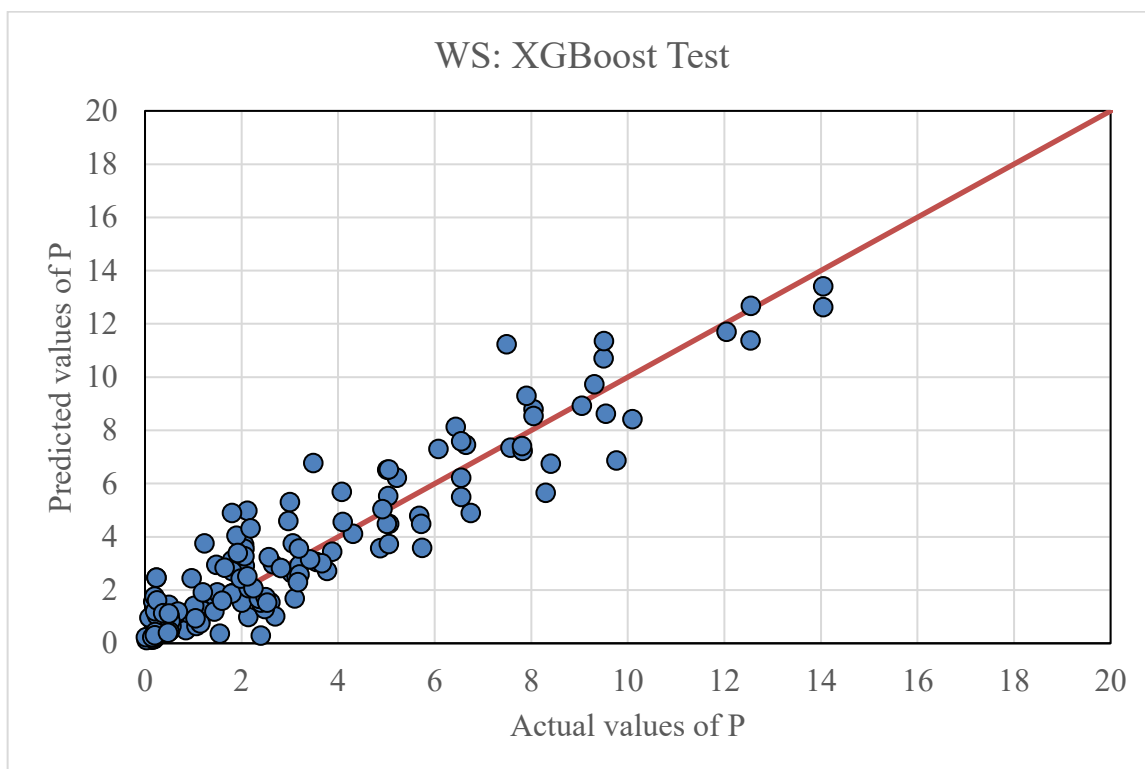


Figure 10. XGBoost Scatter Plot for Water-soluble testing dataset.

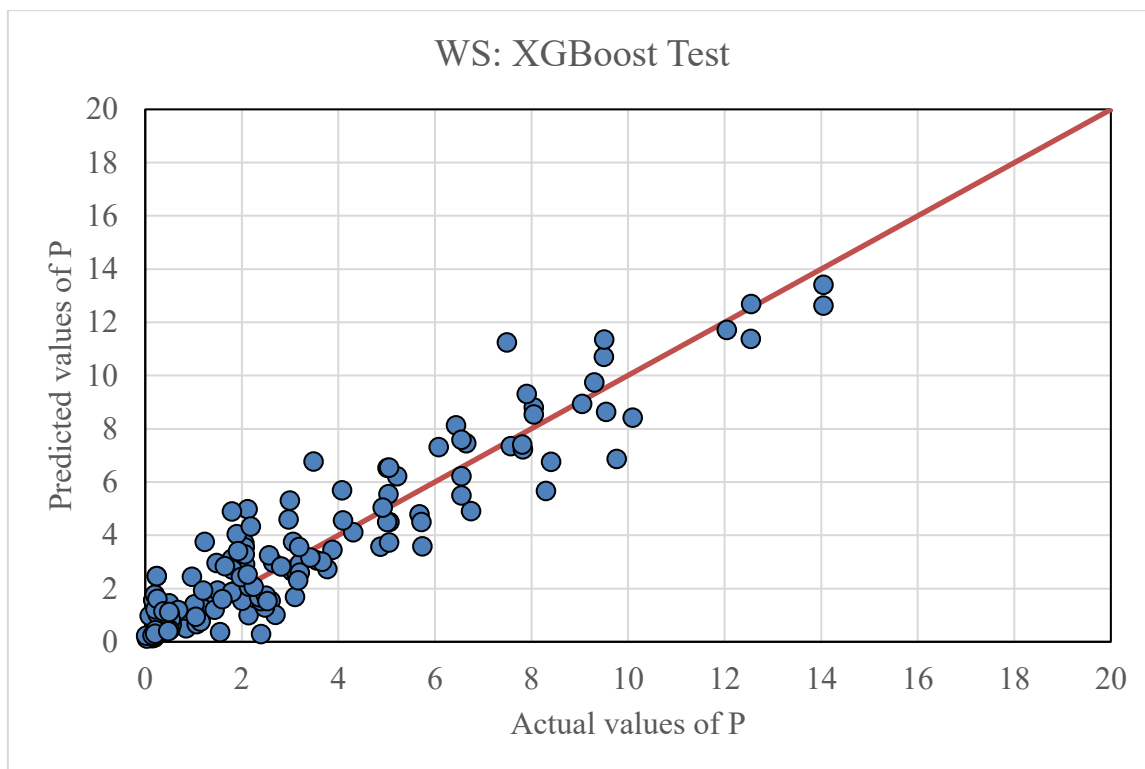


Figure 11. CatBoost Scatter Plot for Water-soluble training dataset

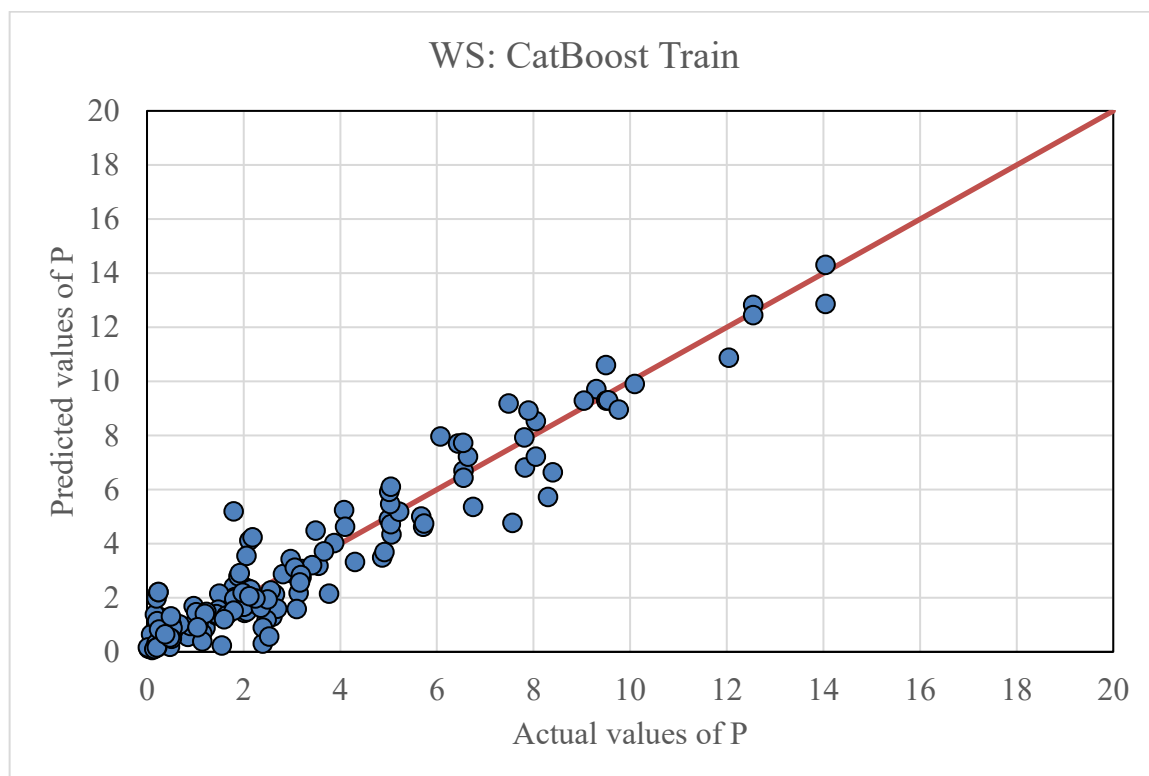


Figure 12. CatBoost Scatter Plot for Water-soluble testing dataset

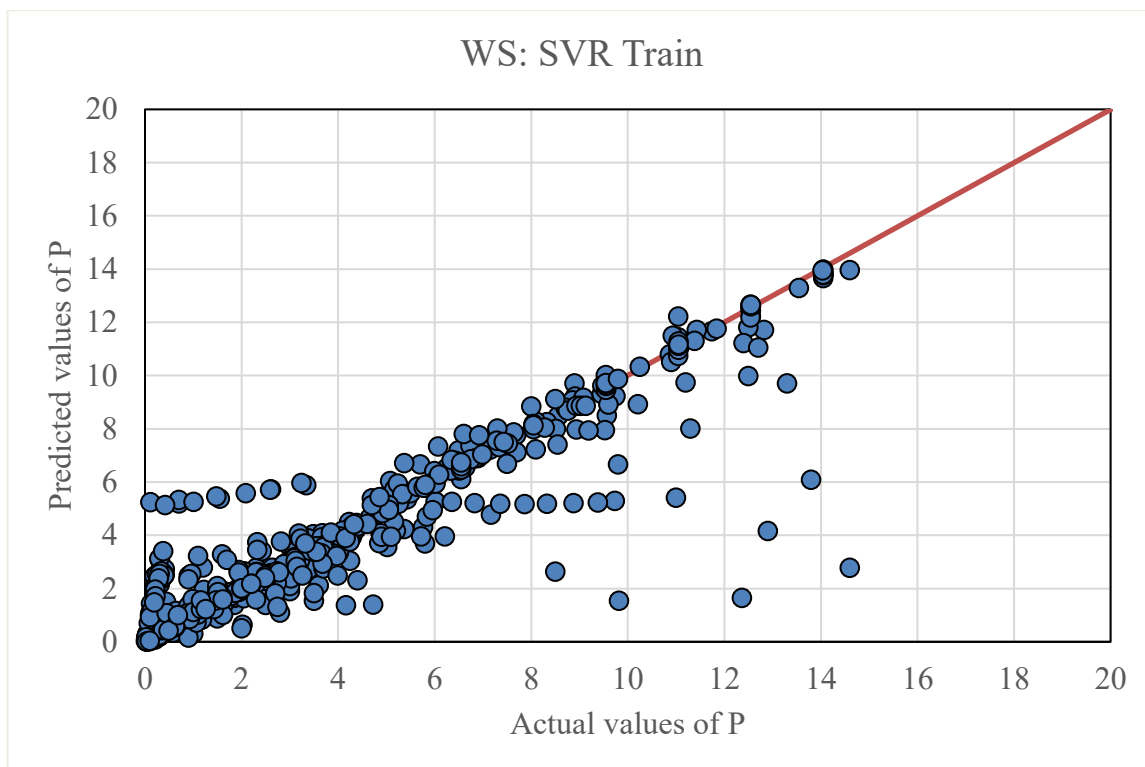


Figure 13. SVR Scatter Plot for Water-soluble training dataset

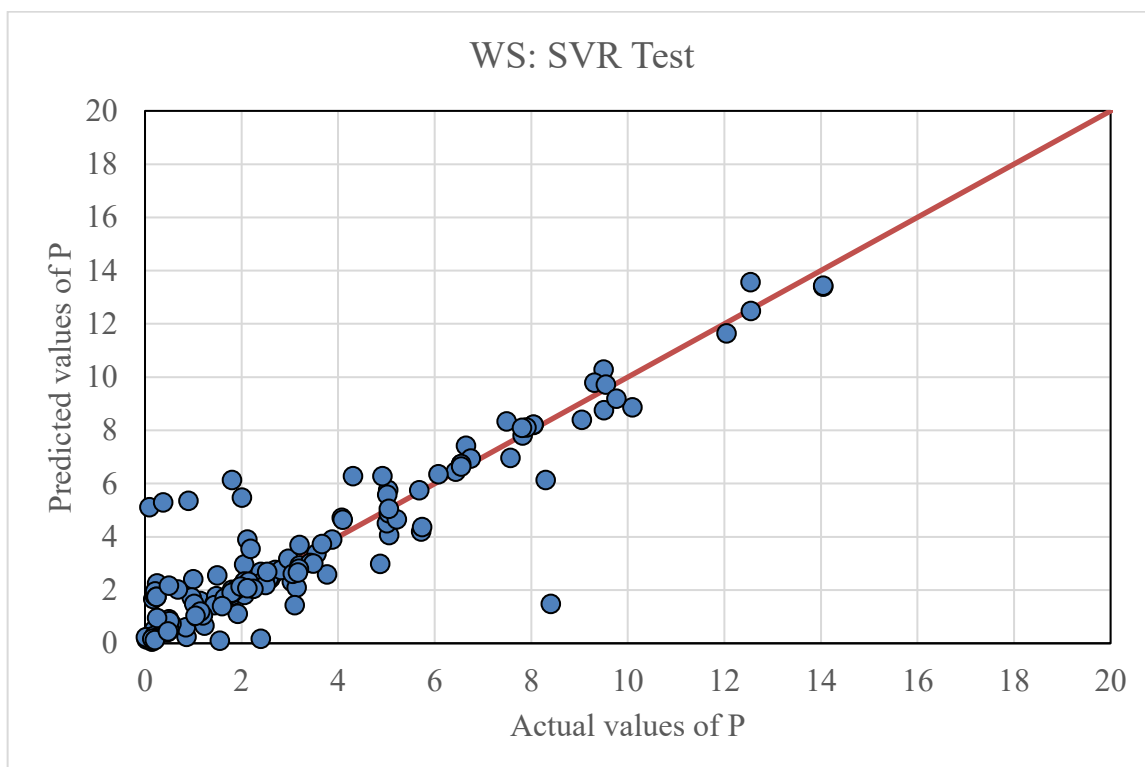


Figure 14. SVR Scatter Plot for Water-soluble testing dataset

Among all four algorithms, CatBoost demonstrated superior performance metrics by achieving $R^2 = 0.92$, $RMSE = 0.91$ MPa, as well as the smallest AIC and BIC values equal to 47.6 and 77.2, respectively. RF and XGBoost yielded very similar results ($R^2 = 0.87$, $RMSE \approx$

1.16–1.17 MPa), suggesting that the standard bagging-based model and gradient boosting algorithm deliver approximately the same degree of predictive accuracy for this particular class of algorithms. The smallest R^2 (0.85) and the largest RMSE (1.26 MPa) among all four regression techniques belonged to SVR; moreover, AIC and BIC of SVR were significantly higher (86.2 and 115.8, respectively) than those of other three algorithms. The reason behind the consistent superiority of CatBoost over RF and XGBoost can be explained by CatBoost's ordered boosting procedure and symmetric trees, leading to lower overfitting rates when applied to medium datasets with different promoter compositions (Prokhorenkova et al., 2018).

As it follows from the results, all four regression techniques can be successfully used on the data on water-soluble promoters. The relatively high values of R^2 , ranging between 0.85 and 0.92, imply that machine learning regressions can explain about 85–92% of the variation in equilibrium pressure. RMSE values ranging from 0.91 to 1.26 MPa can be considered reasonable errors given the pressure range of 0.1–14 MPa covered by the water-soluble dataset. The higher values of RMSE are anticipated at higher pressure values where the curve becomes steeper; the reason is the presence of a nonlinear relationship between concentration and equilibrium pressure near THF stoichiometry discussed in Chapter 3.

4.2.2 Best Model - CatBoost

CatBoost was identified as the best-performing model for the water-soluble category based on its superior performance across all four evaluation metrics. The model's strong performance is consistent with the moderate but nonlinear feature-pressure relationships identified in the correlation heatmap (Figure 3), where temperature was the dominant linear predictor ($r \approx +0.34$) but overall correlations were weak, indicating a dataset in which the ordered boosting and second-order gradient estimation of CatBoost can effectively capture the nonlinear interactions between temperature, promoter concentration, molecular descriptors, and gas composition.

4.2.3 Internal Validation – CatBoost

To confirm the model's ability to understand the trend and evaluate its general predictive ability, the digital model was used on the initial dataset, and the pressure values were predicted with it. The results were presented for different combinations of input parameters, namely: different concentrations of promoter, different gas compositions, and different promoters at the same concentrations. The results are presented in Figures 15–17.

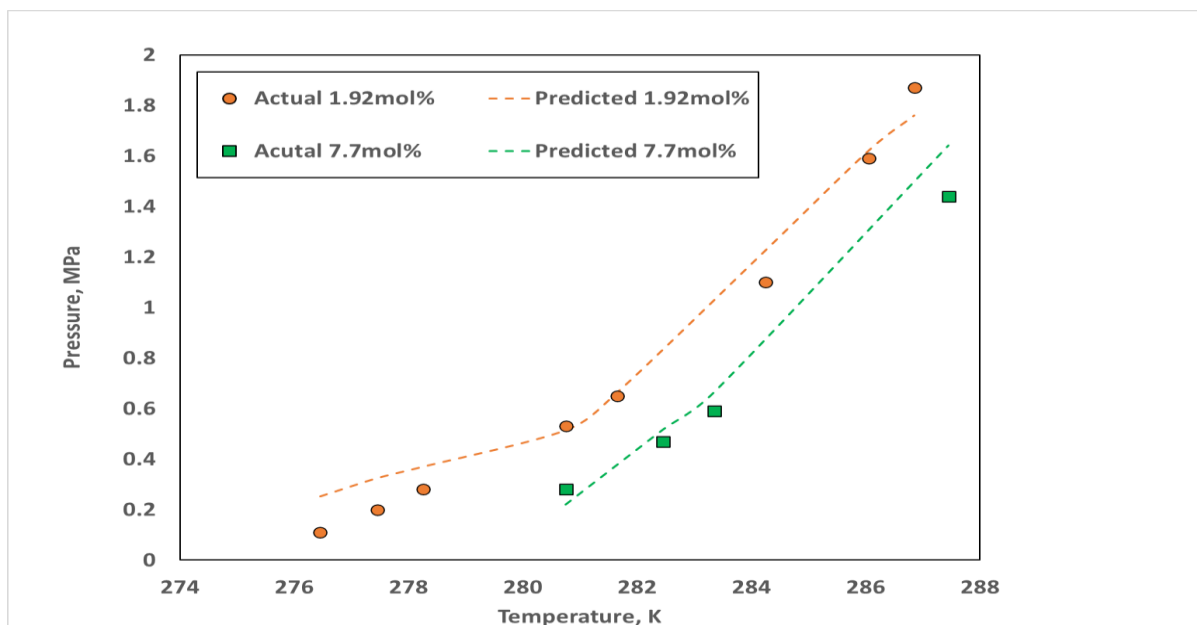


Figure 15. CatBoost predicted vs actual for different promoter concentrations for WS.

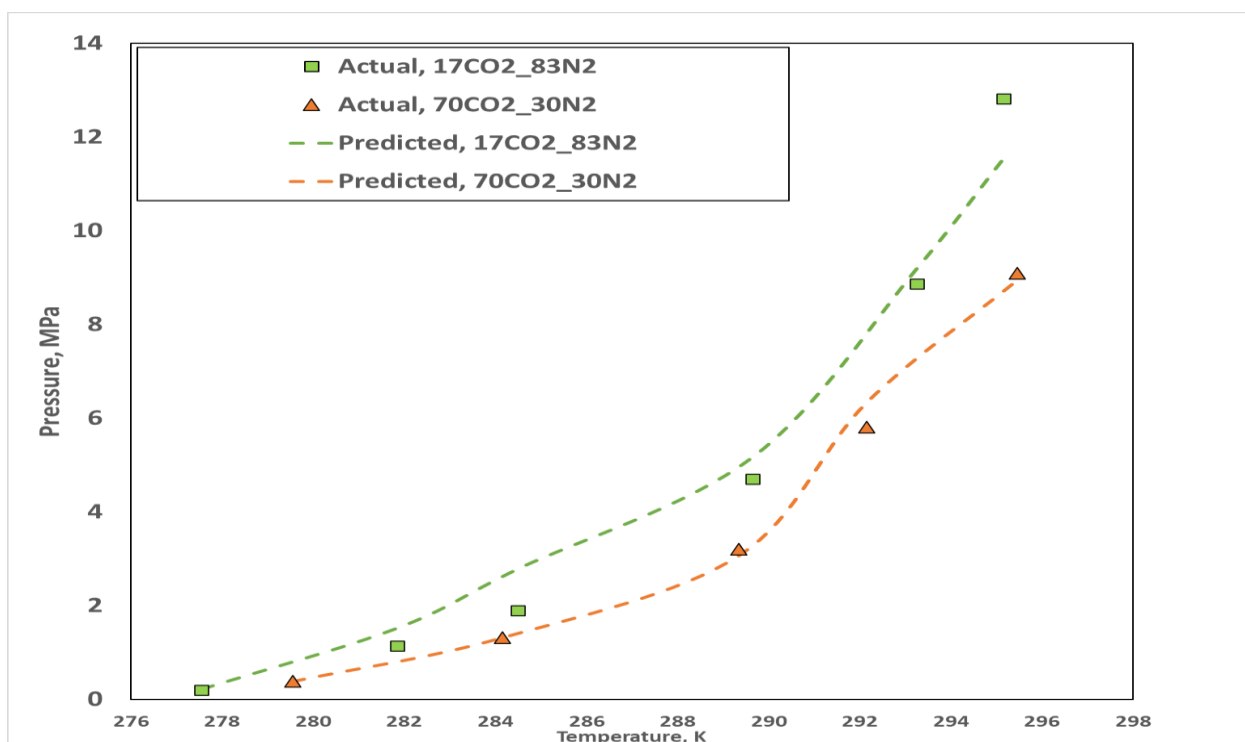


Figure 16. CatBoost predicted vs actual for different gas compositions for WS.

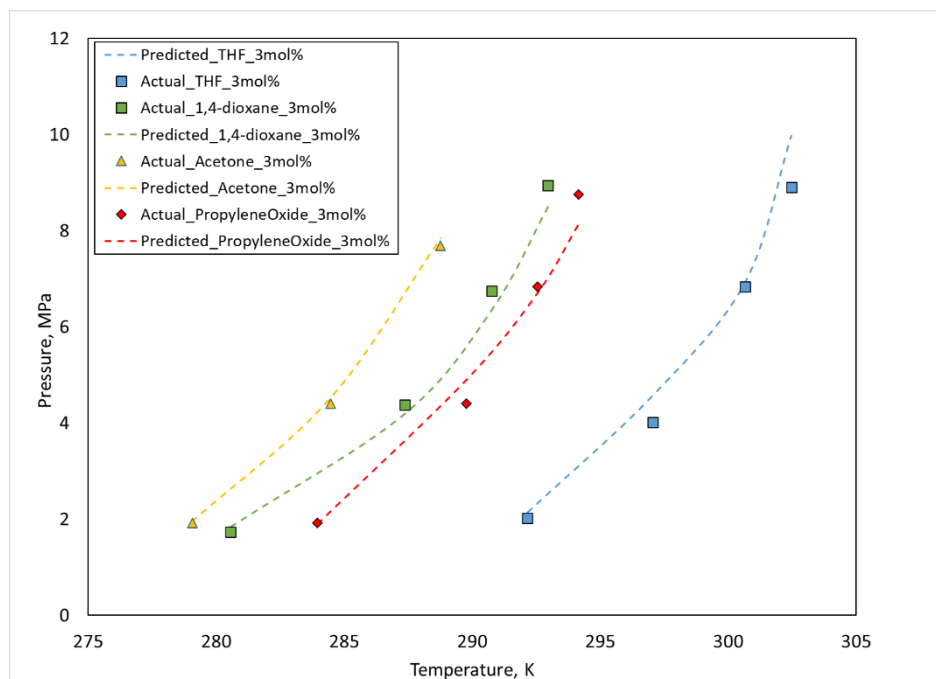


Figure 17. CatBoost predicted vs actual for different promoters at the same concentration for WS.

4.2.4 External Validation - CatBoost

The CatBoost model trained on the water-soluble dataset was applied to an external validation set of 9 data points sourced from a paper not included in the training or testing dataset. The results to evaluate the performance of the predictive ability of the model are shown in Figure 18. This figure represents the difference between the actual data from the paper and the predicted values from the model.

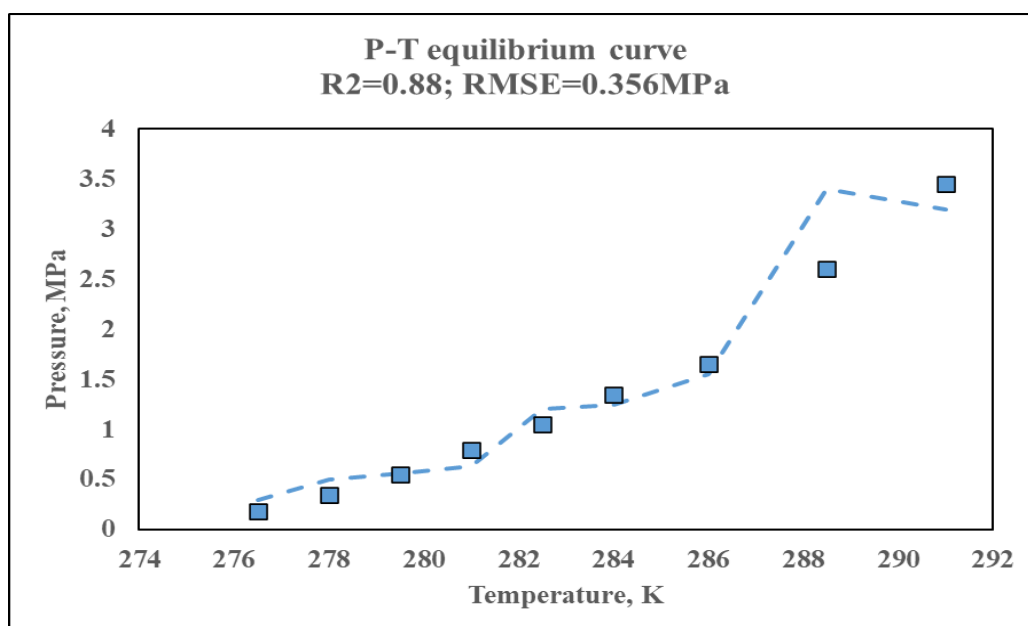


Figure 18. P-T Equilibrium curve for experimental and predicted values with WS CatBoost model.

Table 10 reports the external validation performance.

Table 10. External validation performance - water-soluble CatBoost model.

Metric	Value
R ²	0.88
RMSE	0.356 MPa
MAE	0.237 MPa
No. of external data points	9

Regarding the model's external validation performance, the very high external validation R² value of 0.88 and low RMSE value of 0.356 MPa indicate excellent predictivity about unseen phase equilibrium data. Importantly, the external validation RMSE value of 0.356 MPa is significantly smaller than the test set RMSE value of 0.91 MPa, which could imply that the external validation points are in the center of the data points distribution (near-stoichiometric THF concentration values and moderate pressure). This is because the model should predict such data better due to its higher confidence in making predictions for central rather than more extreme data points (which are not well-represented in the training set). The MAE of 0.237 MPa implies that, on average, the equilibrium pressure prediction error does not exceed 0.24 MPa, which represents an acceptable error range from the physical viewpoint.

4.3 Water-Insoluble Promoters

4.3.1 Model Performance

Table 11 summarizes the test set performance metrics for all four algorithms applied to the water-insoluble promoter dataset (450 data points, 6 promoter compounds, 80/20 split).

Table 11. Test set performance metrics for water-insoluble promoter models.

Algorithm	R²	RMSE (MPa)	AIC	BIC
Random Forest	0.81	1.16	43.8	67.9
XGBoost	0.68	1.51	87.5	111.6
CatBoost	0.83	1.12	37.9	62.0
SVR	0.84	1.12	37.9	62.0

Table 12 summarizes the hyperparameters that were used during testing and training of the ML models on water-insoluble dataset.

Table 12. Hyperparameters summary for water-insoluble category for all models

Algorithm	Hyperparameters	Optimal hyperparameters
Random Forest	n_estimators = [100, 200, 300] max_features = ['sqrt', 'log2', 0.7] max_depth = [2, 4, 6, 8, 10] min_samples_split = [2] min_samples_leaf = [1] bootstrap = [True]	n_estimators = 200 max_features = 'sqrt' max_depth = 8 min_samples_split = 2 min_samples_leaf = 1 bootstrap = True
XGBoost	'xgb__n_estimators': randint(100, 800), 'xgb__max_depth': randint(2, 8), 'xgb__learning_rate': loguniform(1e-3, 3e-1), 'xgb__subsample': uniform(0.5, 0.5), 'xgb__colsample_bytree': uniform(0.5, 0.5), 'xgb__reg_alpha': loguniform(1e-3, 1e1), 'xgb__reg_lambda': loguniform(1e-3, 1e1), 'xgb__min_child_weight': randint(1, 15), 'xgb__gamma': uniform(0, 1.0)	xgb__n_estimators = 472 xgb__max_depth = 5 xgb__learning_rate = 0.083 xgb__subsample = 0.82 xgb__colsample_bytree = 0.77 xgb__reg_alpha = 0.56 xgb__reg_lambda = 2.3 xgb__min_child_weight = 7 xgb__gamma = 0.18
CatBoost	Learning_rate = [0.01, 0.1, 0.2, 0.3]; depth = [2, 4, 6, 8, 10]; l2_leaf_reg = [2, 4, 6, 8, 10]; loss_function = [RMSE]	learning_rate = 0.067 depth = 6 l2_leaf_reg = 7 loss_function = 'RMSE'
SVR	SVR_c = [loguniform (15, 100)]; SVR_epsilon = [loguniform(1e-6, 2e-2)]; SVR_gamma = [loguniform (7e-2, 1.5e-1)]	SVR_C = 38.5 SVR_epsilon = 0.0042 SVR_gamma = 0.093

Figures 19-26 represent the scatter plots for all the ML models for the water-insoluble dataset for training and testing.

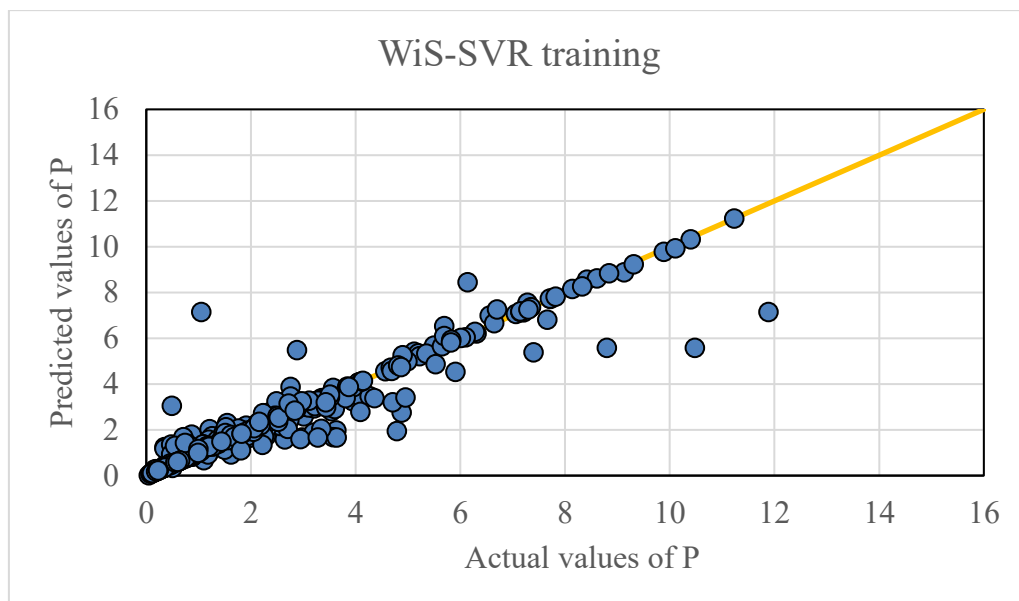


Figure 19. SVR Scatter Plot for Water-insoluble training dataset

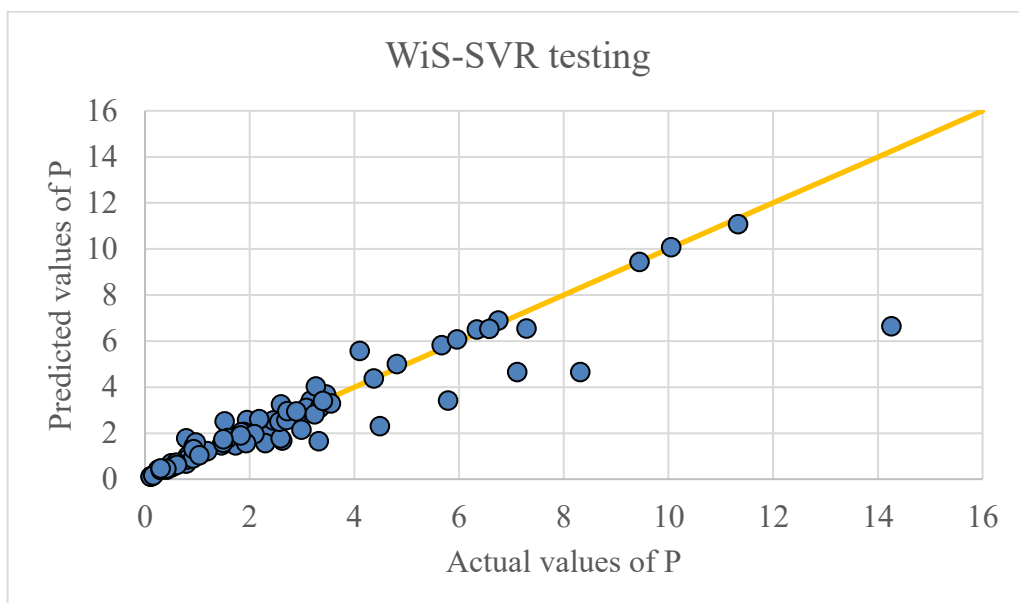


Figure 20. SVR Scatter Plot for Water-insoluble testing dataset

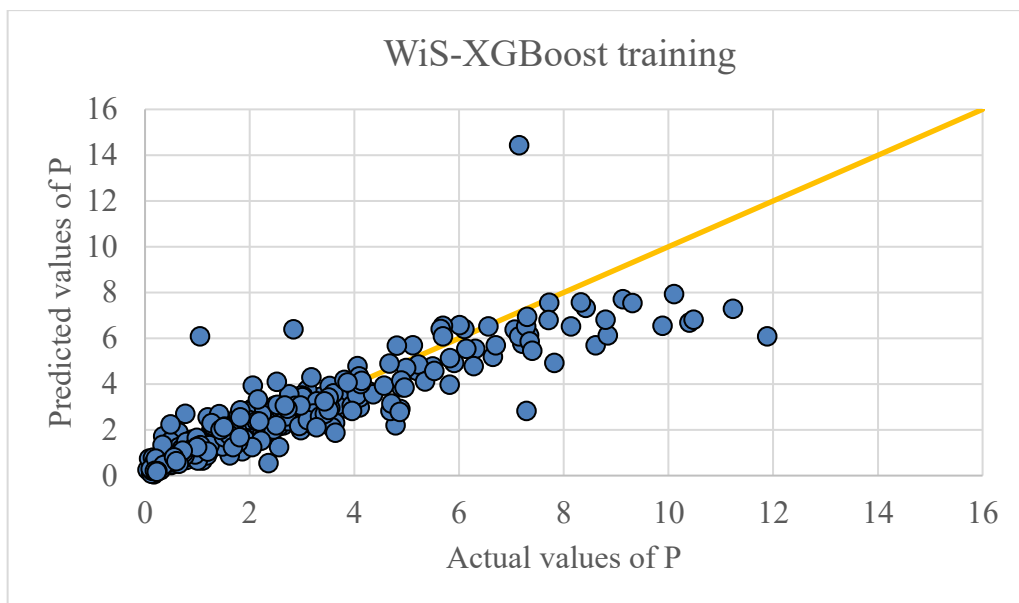


Figure 21. XGBoost Scatter Plot for Water-insoluble training dataset

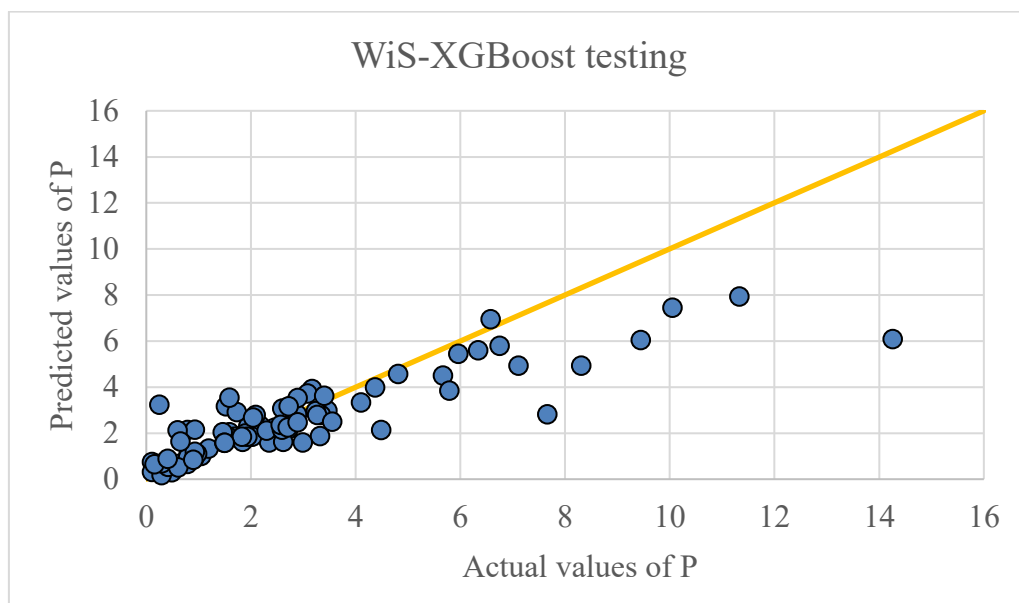


Figure 22. XGBoost Scatter Plot for Water-insoluble testing dataset

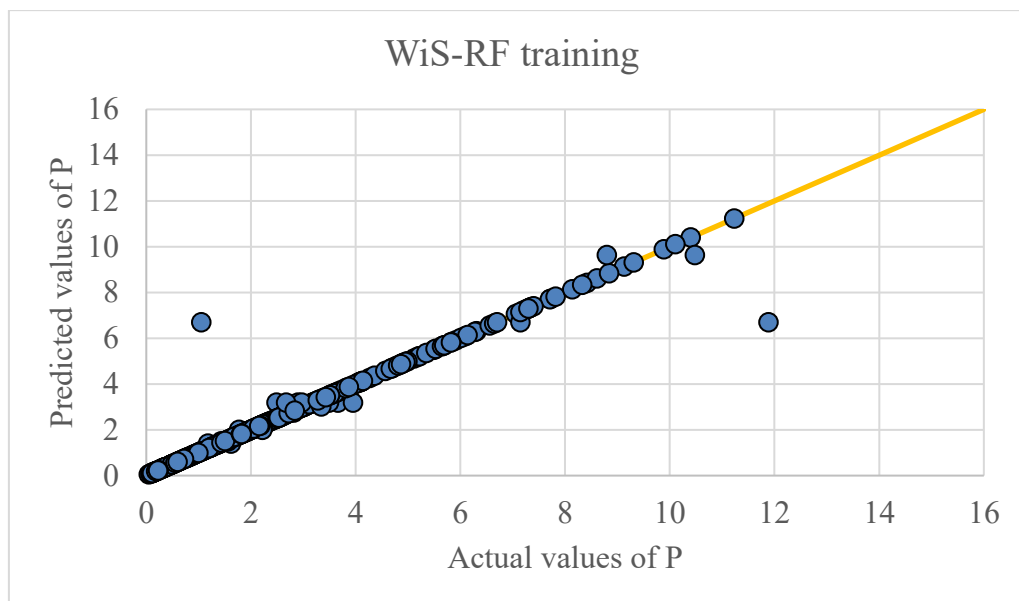


Figure 23. RF Scatter Plot for Water-insoluble training dataset

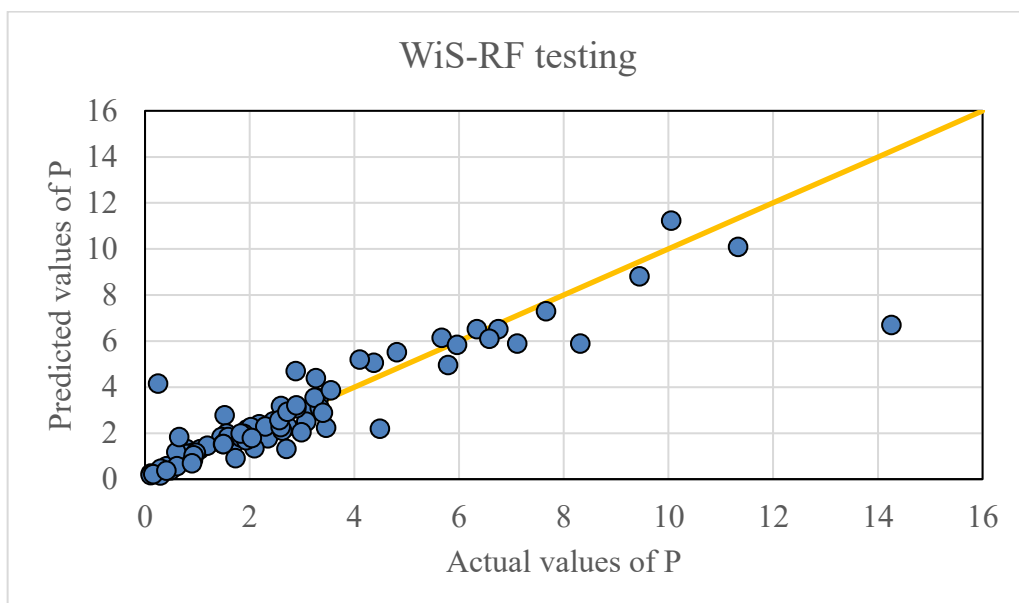


Figure 24. RF Scatter Plot for Water-insoluble testing dataset

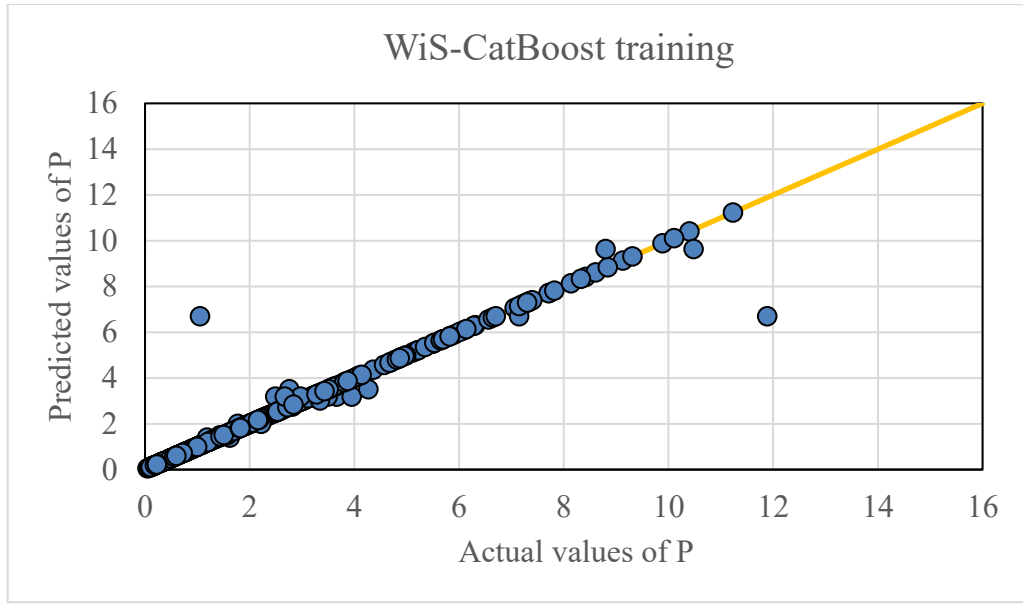


Figure 25. CatBoost Scatter Plot for Water-insoluble training dataset

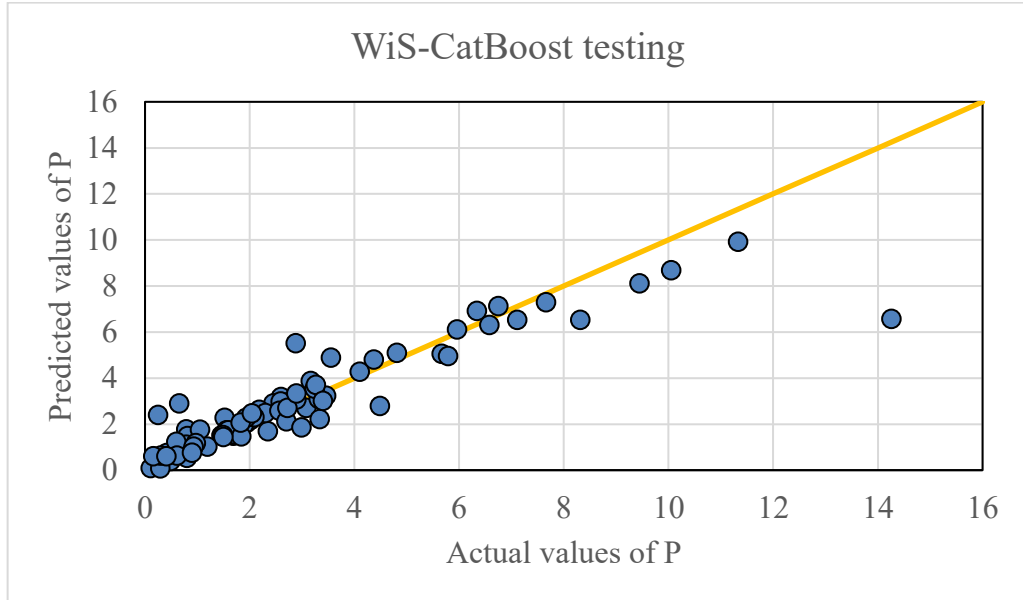


Figure 26. CatBoost Scatter Plot for Water-insoluble testing dataset

SVR showed the most accurate results for the water-insoluble category, as it had the largest R^2 (0.84) and smallest RMSE (1.12 MPa). At the same time, SVR has AIC and BIC values equal to 37.9 and 62.0, respectively, which coincide with CatBoost. Although SVR and CatBoost show only slight differences in terms of R^2 , RMSE, AIC, and BIC, the scatter plots demonstrate that SVR has better predictability behavior compared to CatBoost in terms of scatter plot distribution and prediction accuracy, generally in the dataset pressure range. Both CatBoost and SVR show much better results than RF ($R^2 = 0.81$) and significantly better results

compared to XGBoost ($R^2 = 0.68$), which was the worst performer among the four algorithms in predicting water-insoluble compounds. Such a relatively weak performance of XGBoost can be explained by its sensitivity to overfitting for small non-homogeneous datasets, which is mitigated with ordered boosting in CatBoost and the symmetrical tree construction method used in the algorithm (Prokhorenkova et al., 2018).

Overall, the R^2 values' range for the water-insoluble category is slightly lower, 0.68–0.84 compared to the range in the case of water-soluble compounds, which is in line with higher heterogeneity revealed in the exploratory data analysis. As was mentioned above, in Chapter 3, the water-insoluble compounds dataset consists of two different types of compounds: those which form a type II structure (sII) and those forming type H structure (sH), which implies different equilibrium pressure characteristics, resulting in a lack of strong correlation with any of the input variables ($|r| \leq 0.21$; Figure 4). Thus, such a low R^2 range can be justified in the context of the data modeling problem posed. Furthermore, RMSE values of 1.12–1.51 MPa may be considered somewhat high considering the pressure range used in this study.

4.3.2 Best Model — SVR

SVR was identified as the best performing machine learning model for the water insoluble category owing to its highest value of R^2 (0.84), the lowest value of RMSE (1.12 MPa, equal to that of CatBoost), and the most optimal scatter plot distribution. It is important to highlight that the better performance of SVR compared to the decision tree-based models is remarkable, especially given the ability of the Radial Basis Function (RBF) kernel to establish the nonlinear relationship between pressures and features without being limited to only a few structural patterns (Smola & Schölkopf, 2004).

4.3.3 Internal Validation – SVR

To confirm the model's ability to understand the trend and evaluate its general predictive ability, the digital model was used on the initial dataset, and the pressure values were predicted with it. The results were presented for different combinations of input parameters, namely: different concentrations of promoter, different gas compositions, and different promoters at the same concentrations. The results are presented in Figures 27-29.

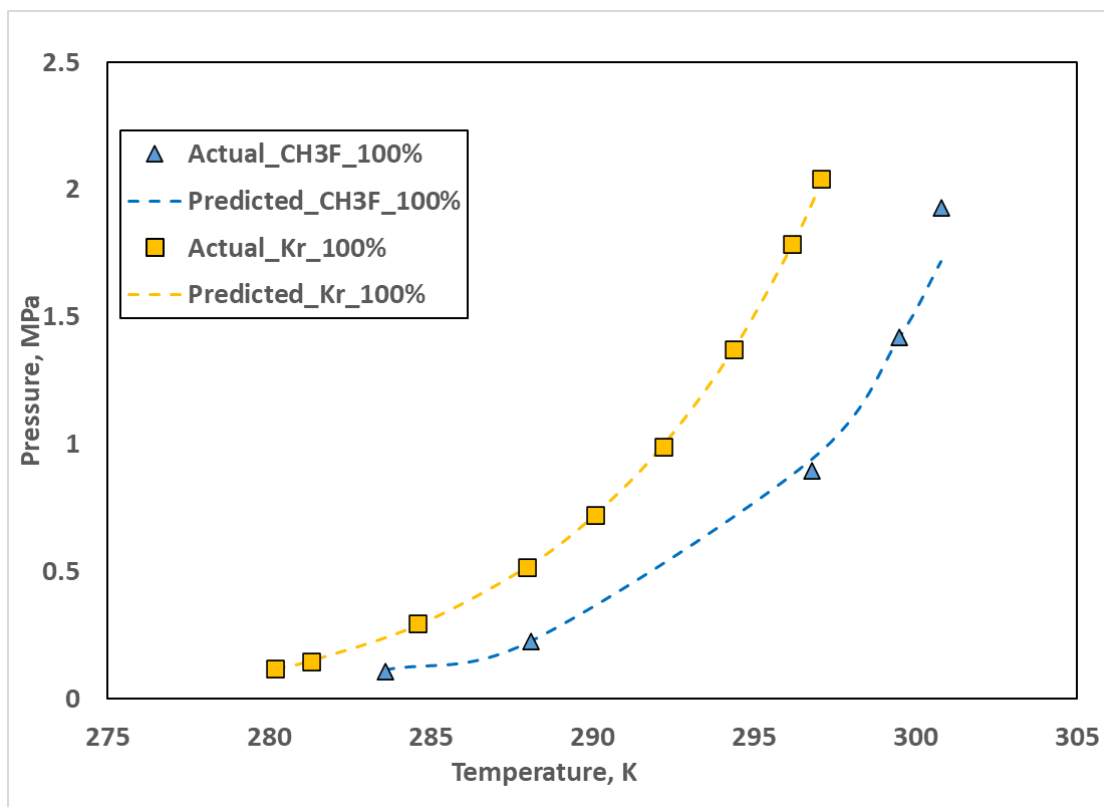


Figure 27. SVR predicted vs actual for different gas compositions for WiS.

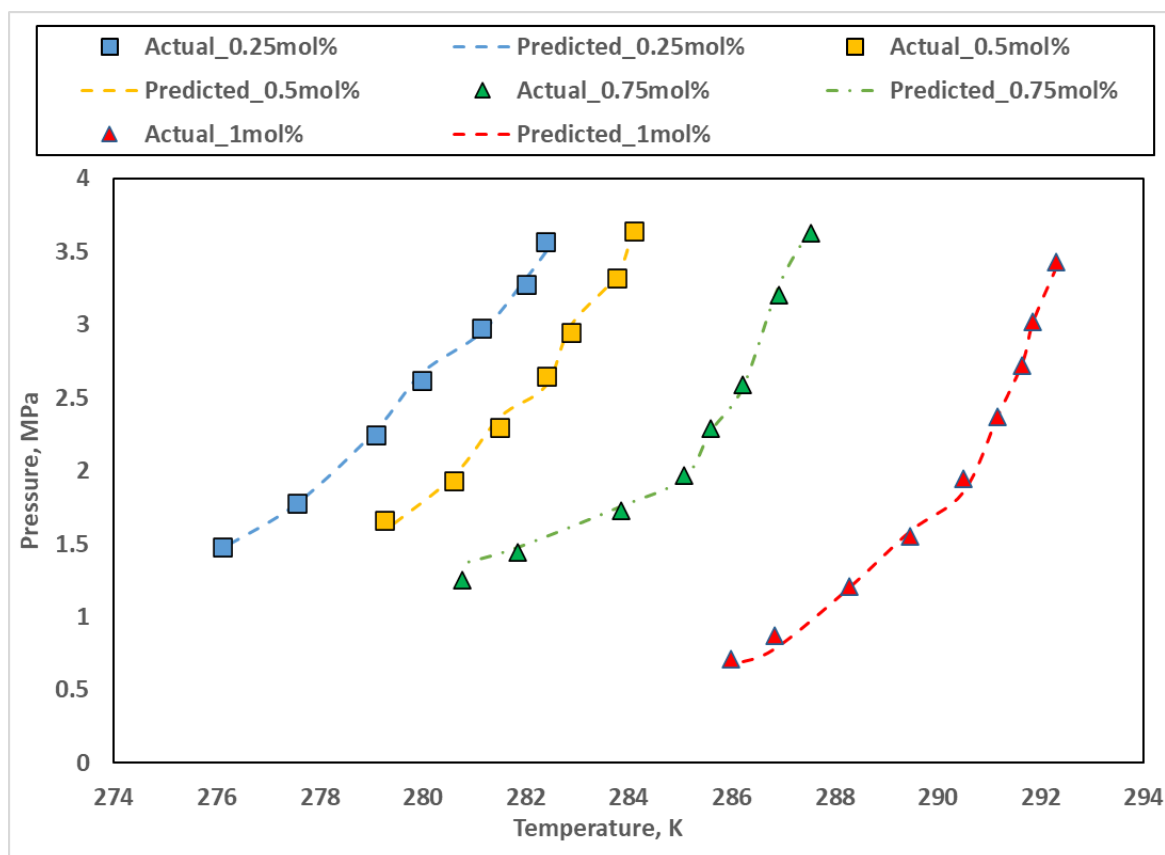


Figure 28. SVR predicted vs actual for different promoter concentrations for WiS.

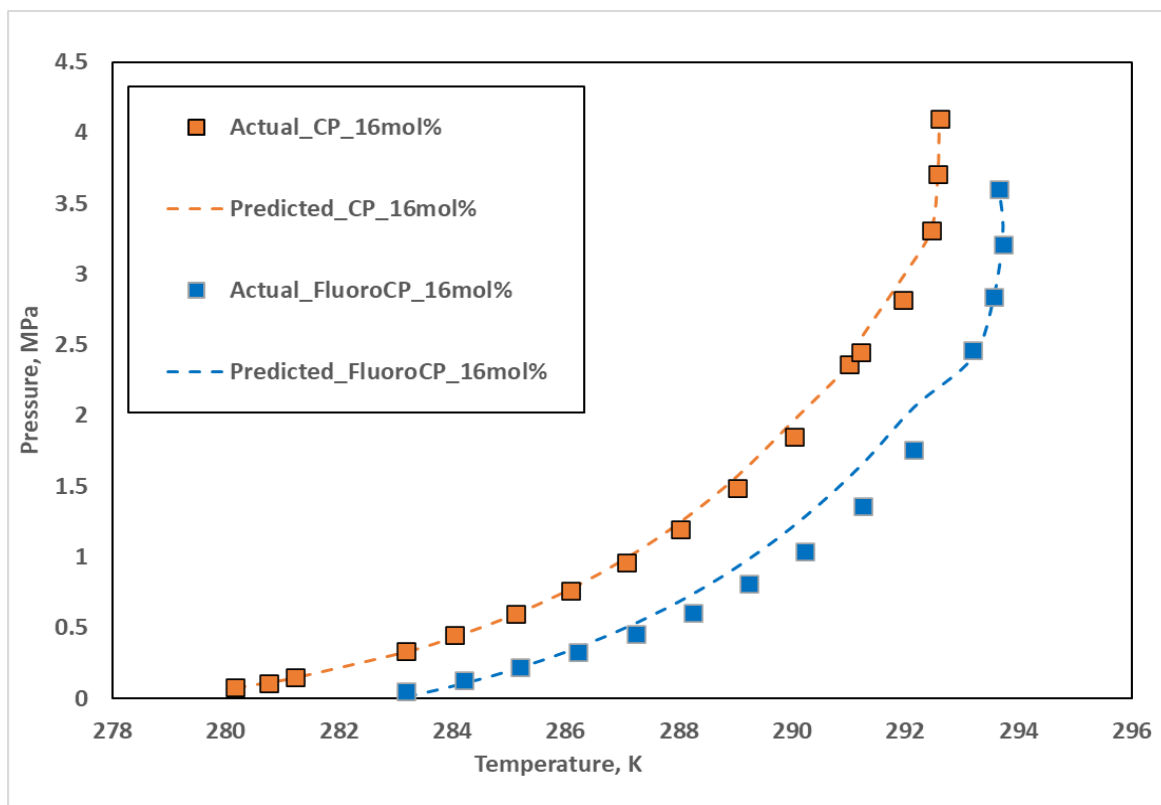


Figure 29. SVR predicted vs actual for different promoters at the same concentration for WiS.

4.3.4 External Validation — SVR

The SVR model trained on the water-insoluble dataset was applied to an external validation set of 9 data points sourced from a paper not included in the training or testing dataset. The results to evaluate the performance of predictive ability of the model are shown in Figure 30. This figure represents the difference between the actual data from the paper and the predicted values from the model.

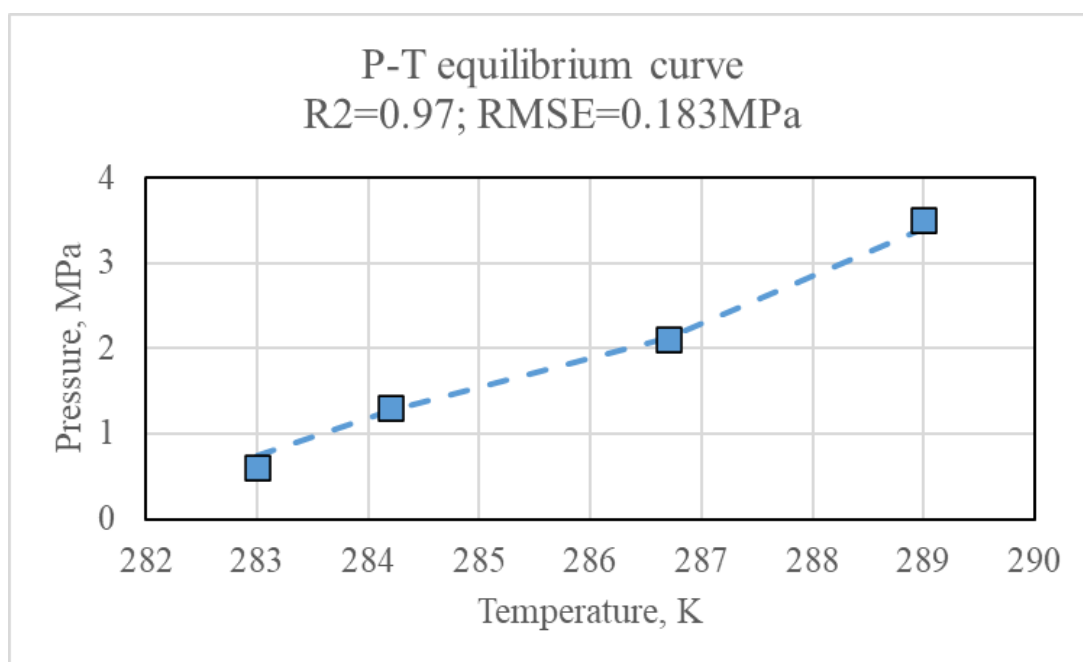


Figure 30. P-T Equilibrium curve for experimental and predicted values with WiS SVR model.

The SVR model trained on the water-insoluble dataset was applied to an external validation set of 13 data points. Table 13 reports the external validation performance.

Table 13. External validation performance — water-insoluble SVR model.

Metric	Value
R ²	0.97
RMSE	0.183 MPa
MAE	0.147 MPa
No. of external data points	4

The performance statistics of the water-insoluble SVR in external validation are impressive: $R^2 = 0.97$ and $RMSE = 0.183$ MPa – much more accurate compared to test data performance statistics: $R^2 = 0.84$, $RMSE = 1.12$ MPa. The MAE value of 0.147 MPa means that the model, on average, is predicting equilibrium pressure with an accuracy of about 0.15 MPa from the actual experimental value obtained from the external validation set. However, one should consider the fact that only four points were used for external validation – it may not be representative enough for a whole system. Despite the small number of external validation points, the accuracy attained by the water-insoluble SVR is quite high, thus demonstrating that the SVR model was able to capture adequately the relations between pressure, temperature, and molecular descriptors in water-insoluble promoter systems adequately. The high R^2 on the

external validation set in comparison with the test set is also notable: probably because the points in the external validation set are in the feature space area where there are enough training points, namely, in cyclopentane or cyclohexane with methane at medium pressure, the largest subset of water-insoluble points.

4.4 Semi-Clathrate Promoters

4.4.1 Model Performance

Table 14 summarizes the test set performance metrics for all four algorithms applied to the semi-clathrate promoter dataset (620 data points, 6 promoter compounds, 80/20 split).

Table 14. Test set performance metrics for semi-clathrate promoter models.

Algorithm	R ²	RMSE (MPa)	AIC	BIC
Random Forest	0.54	1.88	175.5	206.0
XGBoost	0.71	1.50	120.5	151.0
CatBoost	0.63	1.69	150.7	181.5
SVR	0.934	0.724	37.9	62.05

Table 15 summarizes the hyperparameters that were used during testing and training of the ML models on the semi-clathrate dataset.

Table 15. Hyperparameters summary for semi-clathrate category for all models.

Algorithm	Hyperparameters	Optimal hyperparameters
Random Forest	n_estimators = [100, 200, 300] max_features = ['sqrt', 'log2', 0.7, 1] max_depth = [2, 4, 6, 8, 10] min_samples_split = [2] min_samples_leaf = [1] bootstrap = [True]	n_estimators = 300 max_features = 0.7 max_depth = 6 min_samples_split = 2 min_samples_leaf = 1 bootstrap = True
XGBoost	n_estimators = [300, 500, 700]; max_depth = [3, 4, 5, 6]; learning_rate = [0.03, 0.05, 0.08]; min_child_weight = [1, 2, 4];	n_estimators = 500 max_depth = 4 learning_rate = 0.05 min_child_weight = 2

	$\gamma = [0, 0.1, 0.3];$ $\text{subsample} = [0.7, 0.85, 1];$ $\text{colsample_bytree} = [0.7, 0.85, 1];$ $\text{reg_alpha} = [0, 0.01, 0.1];$ $\text{reg_lambda} = [1, 2, 5]$	$\gamma = 0.1$ $\text{subsample} = 0.85$ $\text{colsample_bytree} = 0.7$ $\text{reg_alpha} = 0.01$ $\text{reg_lambda} = 2$
CatBoost	$\text{Learning_rate} = [0.01, 0.1, 0.2, 0.3];$ $\text{depth} = [2, 4, 6, 8, 10];$ $\text{l2_leaf_reg} = [2, 4, 6, 8, 10];$ $\text{loss_function} = [\text{RMSE}]$	$\text{learning_rate} = 0.1$ $\text{depth} = 6$ $\text{l2_leaf_reg} = 6$ $\text{loss_function} = \text{'RMSE'}$
SVR	$\text{SVR_c} = [\text{loguniform}(15, 100)];$ $\text{SVR_epsilon} = [\text{loguniform}(1\text{e-}6, 3\text{e-}2)];$ $\text{SVR_gamma} = [\text{loguniform}(7\text{e-}2, 5\text{e-}1)]$	$\text{SVR_C} = 57.3$ $\text{SVR_epsilon} = 0.0087$ $\text{SVR_gamma} = 0.21$

Figures 31-38 represent the scatter plots for all the ML models for water-insoluble dataset for training and testing.

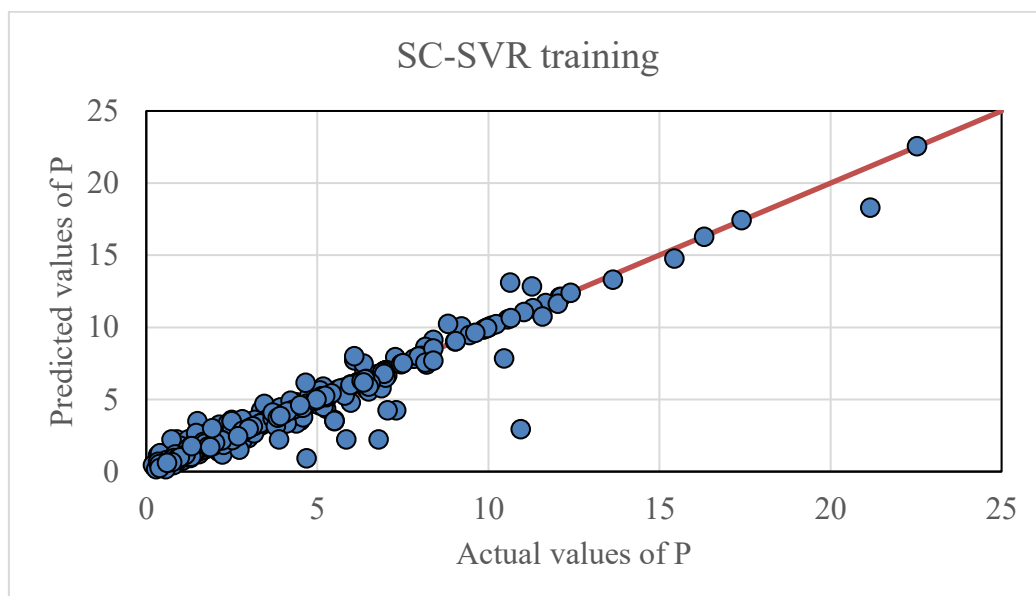


Figure 31. SVR Scatter Plot for Semi-clathrate training dataset

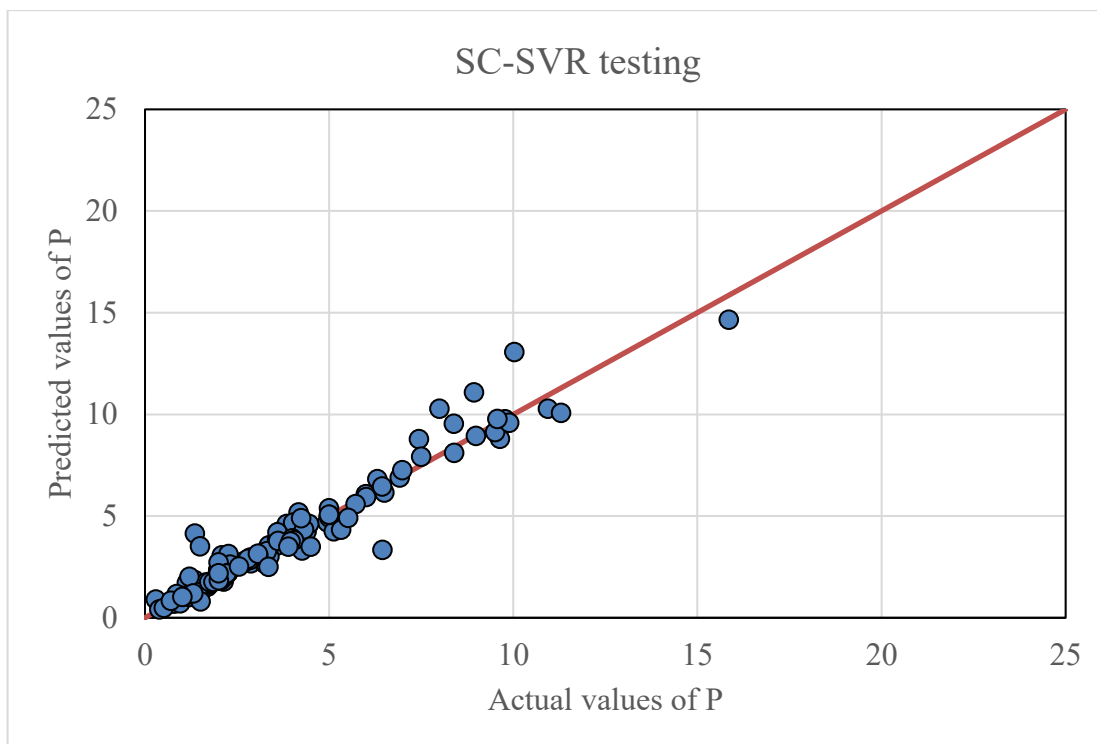


Figure 32. SVR Scatter Plot for Semi-clathrate testing dataset

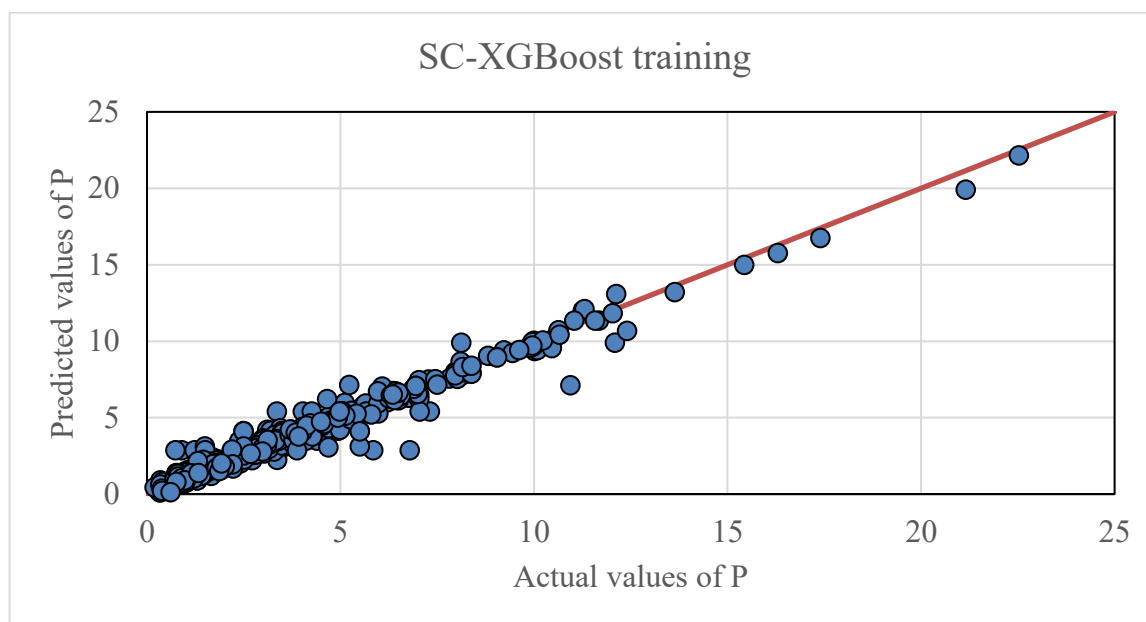


Figure 33. XGBoost Scatter Plot for Semi-clathrate training dataset

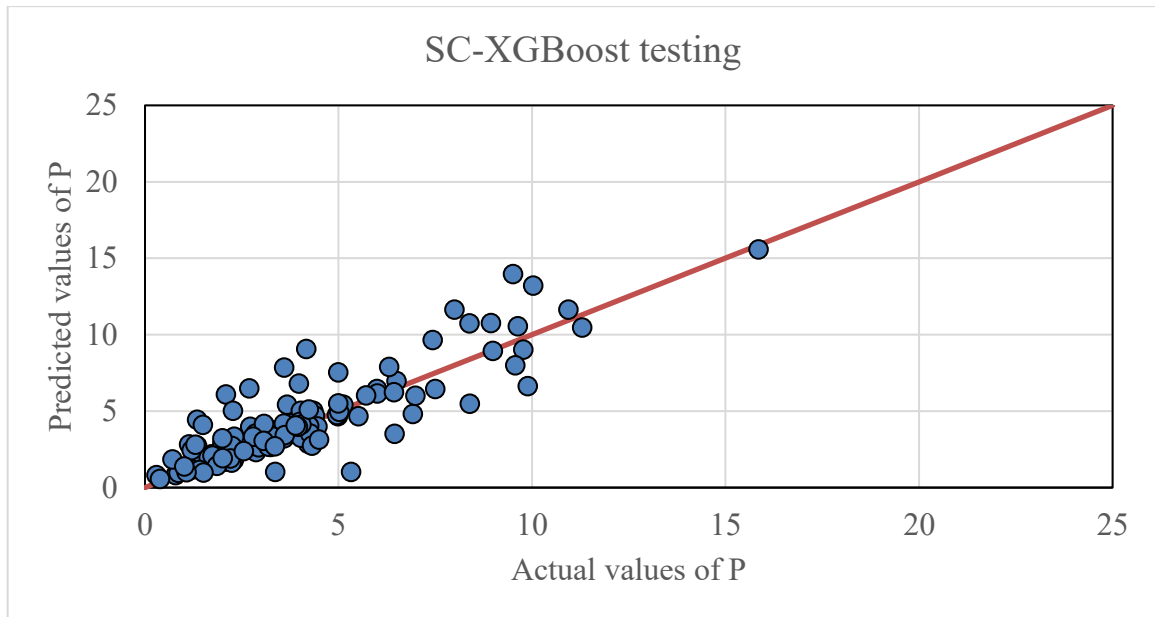


Figure 34. XGBoost Scatter Plot for Semi-clathrate testing dataset

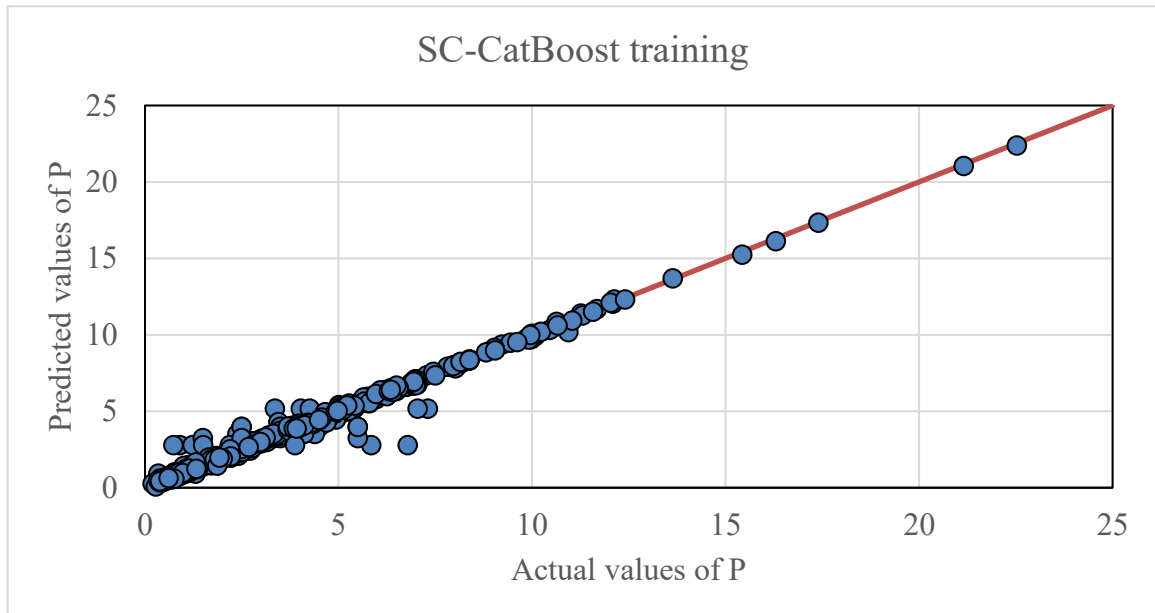


Figure 35. CatBoost Scatter Plot for Semi-clathrate training dataset

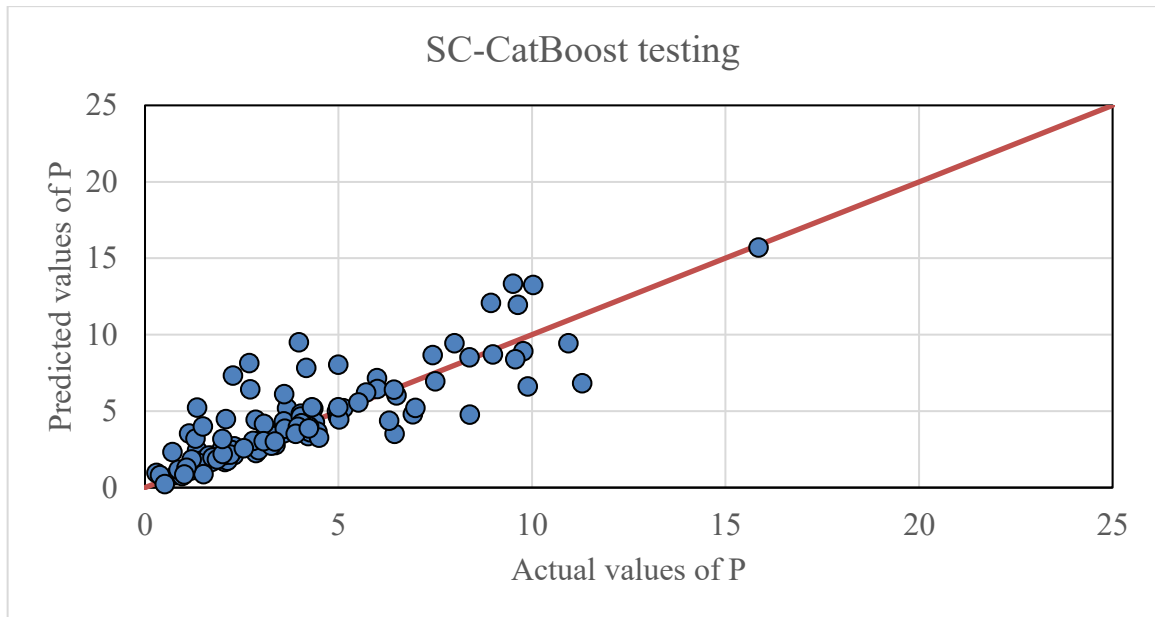


Figure 36. CatBoost Scatter Plot for Semi-clathrate testing dataset

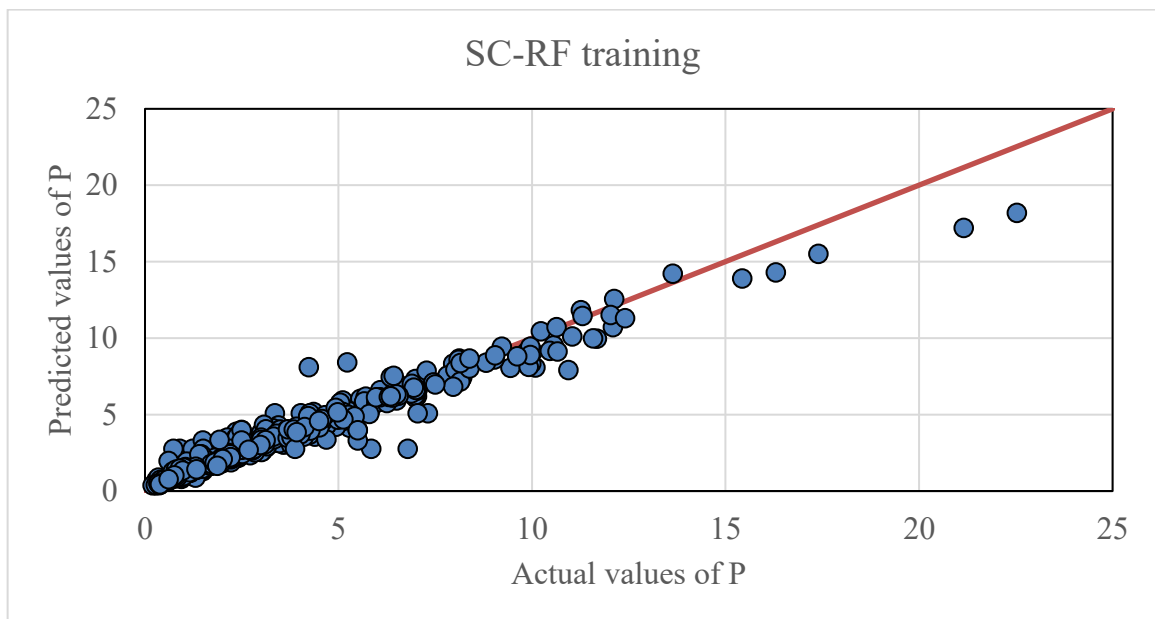


Figure 37. RF Scatter Plot for Semi-clathrate training dataset

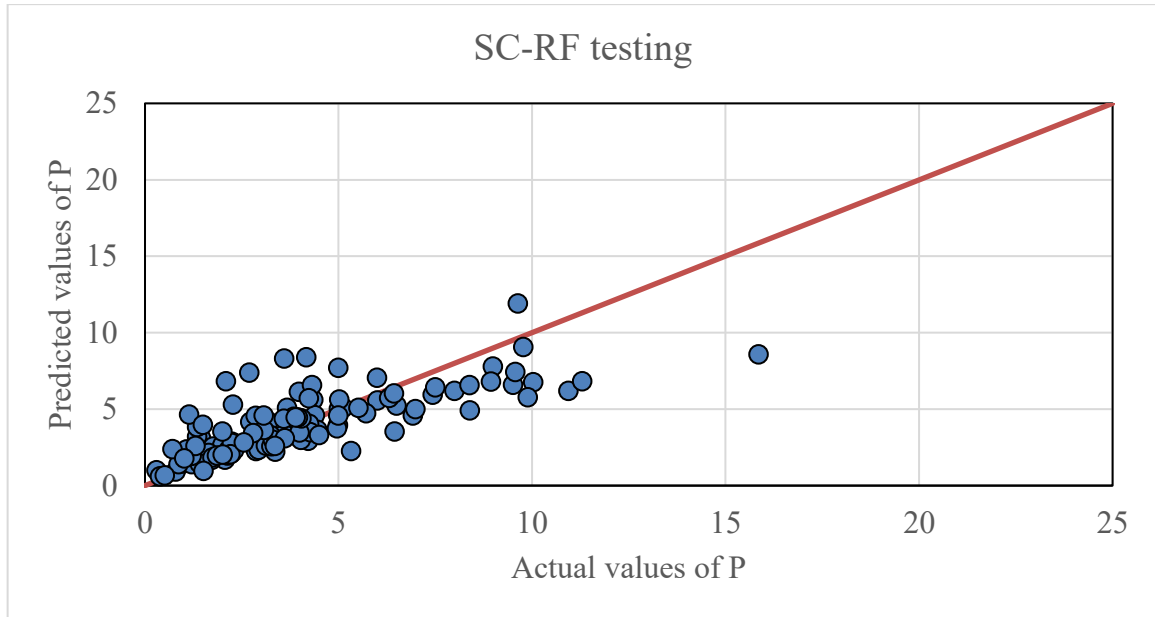


Figure 38. CatBoost Scatter Plot for Semi-clathrate testing dataset

The most interesting divergence in algorithm performance was observed in the semi-clathrate category. SVR significantly outperformed all three tree-based ensembles in all metrics considered, including $R^2 = 0.934$, RMSE = 0.724 MPa, AIC = 37.9, and BIC = 62.05. This is a substantially better performance compared even to the best of the ensemble methods — namely, XGBoost — which had $R^2 = 0.71$ and RMSE = 1.50 MPa. The superiority of SVR compared to ensemble methods is especially pronounced in terms of RMSE (which is about 40% of that for RF) as well as AIC and BIC (which are much smaller for SVR than for all three ensembles — 37.9 and 62.05 vs. 175.5 and 206.0 for RF). As for tree-based algorithms, XGBoost performed best (with $R^2 = 0.71$), then CatBoost (with $R^2 = 0.63$), and finally RF with the worst performance among them ($R^2 = 0.54$), but the latter two algorithms still performed relatively well on other datasets.

The reason behind the exceptional performance of SVR in the semi-clathrate category becomes obvious when considering the results from the exploratory data analysis described in Chapter 3. Namely, as seen in Figure 5, the semi-clathrate dataset has a very pronounced negative linear correlation between temperature and pressure (with $r \approx -0.98$), which indicates a near-deterministic inverse relationship that explains 96% of linear variance in the dependent variable. As noted above, the RBF kernel of SVR is especially suited to modeling this kind of relationship due to the ability to capture both the strong linear component and minor residual nonlinearities in the data (Smola & Schölkopf, 2004). On the other hand, tree-based ensemble algorithms would require many more partitions to model the same function, which could be

affected by local features of the semi-clathrate dataset, like the presence of two different phase behaviors in semi-clathrates for TBAB Type A and Type B hydrates.

The fact that RF performed relatively poorly in the semi-clathrate category (with $R^2 = 0.54$), despite excellent performance in the water-soluble category, points towards the inadequacy of the feature space of the former for the algorithm. Namely, as mentioned above, the near-linearity and the heterogeneity of the structure of the semi-clathrate feature space, caused by the presence of six chemical compounds with six different semi-clathrate phase types, make it difficult for a random forest algorithm to make accurate predictions. This can also be inferred from much worse values of AIC and BIC (120.5–206.0 for tree-based algorithms vs. 37.9–62.05 for SVR) compared to SVR.

4.4.2 Best Model — SVR

The fact that SVR was the model that performed best in the semi-clathrates category was evident from all four criteria of evaluation. It can be said that the supremacy of SVR in relation to the ensemble techniques in the semi-clathrate category is one of the major findings of this thesis, which is analyzed in Chapter 5.

4.4.3 Internal Validation – SVR

To confirm the model's ability to understand the trend and evaluate its general predictive ability, the digital model was used on the initial dataset, and the pressure values were predicted with it. The results were presented for different combinations of input parameters, namely: different concentrations of promoter, different gas compositions, and different promoters at the same concentrations. The results are presented in Figures 39-41.

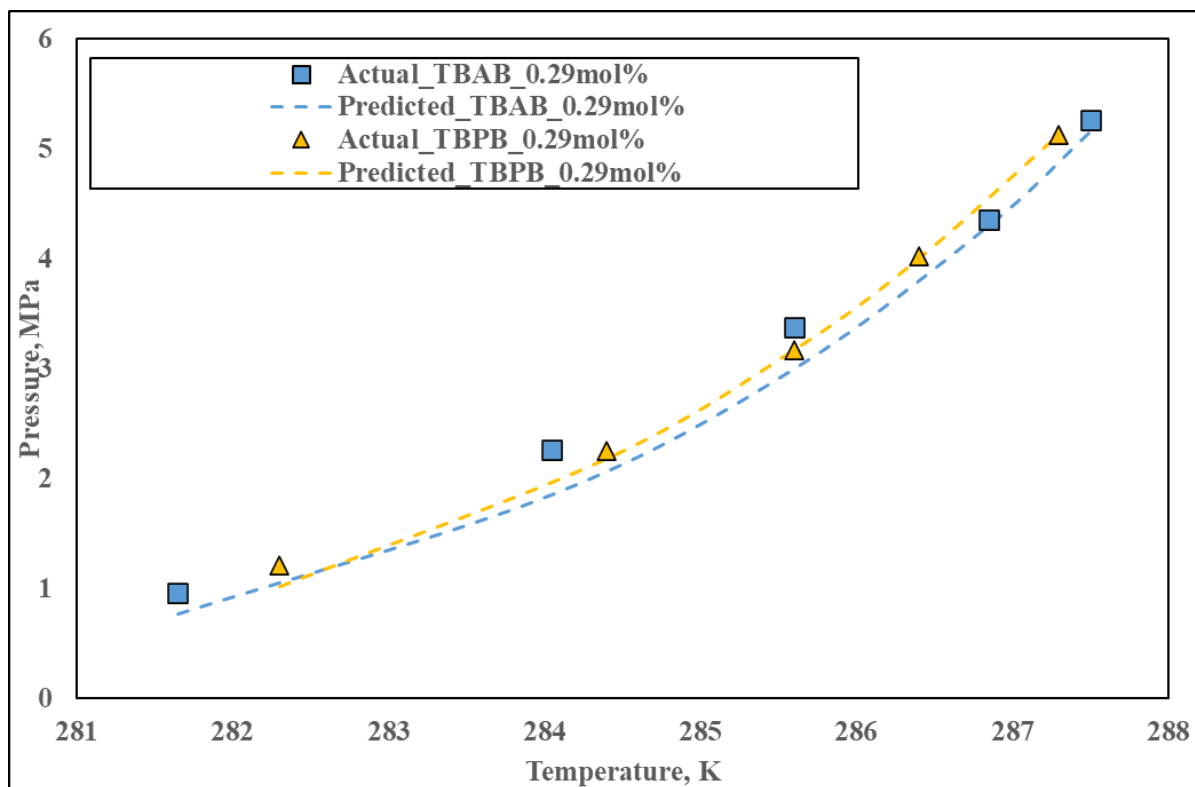


Figure 39. SVR predicted vs actual for different promoters at the same concentration for SC.

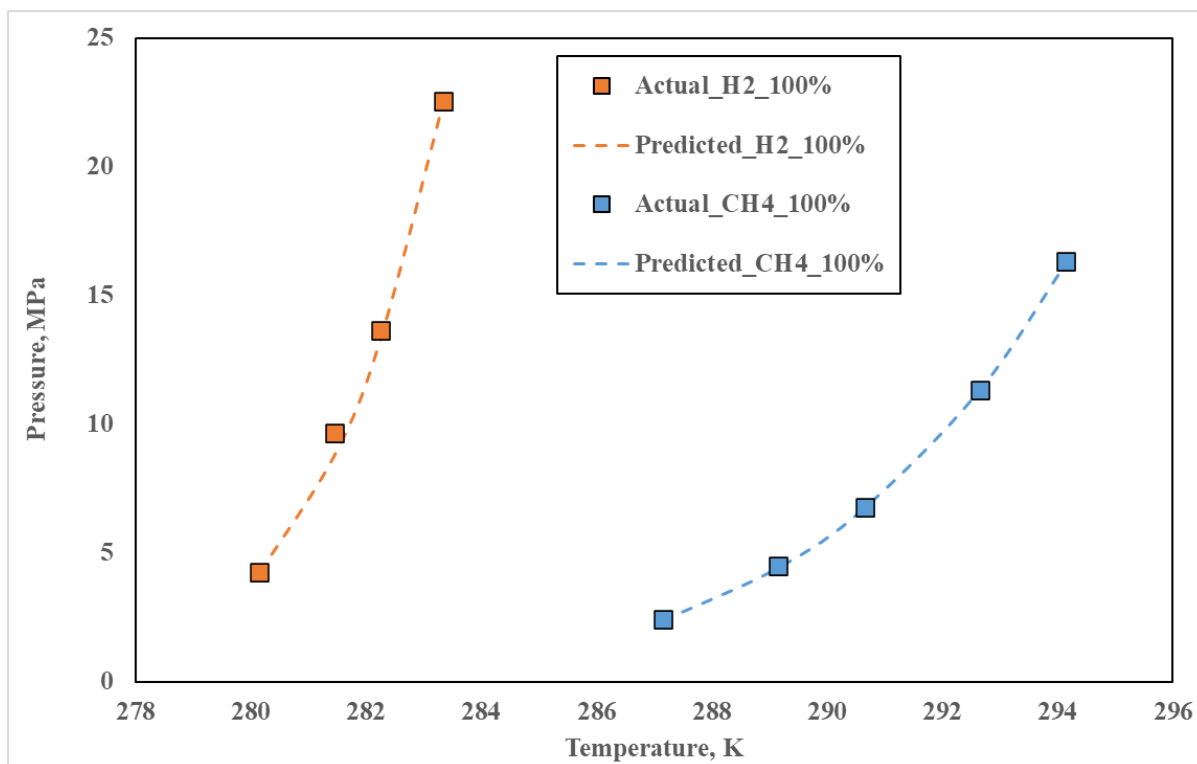


Figure 40. SVR predicted vs actual for different gas compositions for SC.

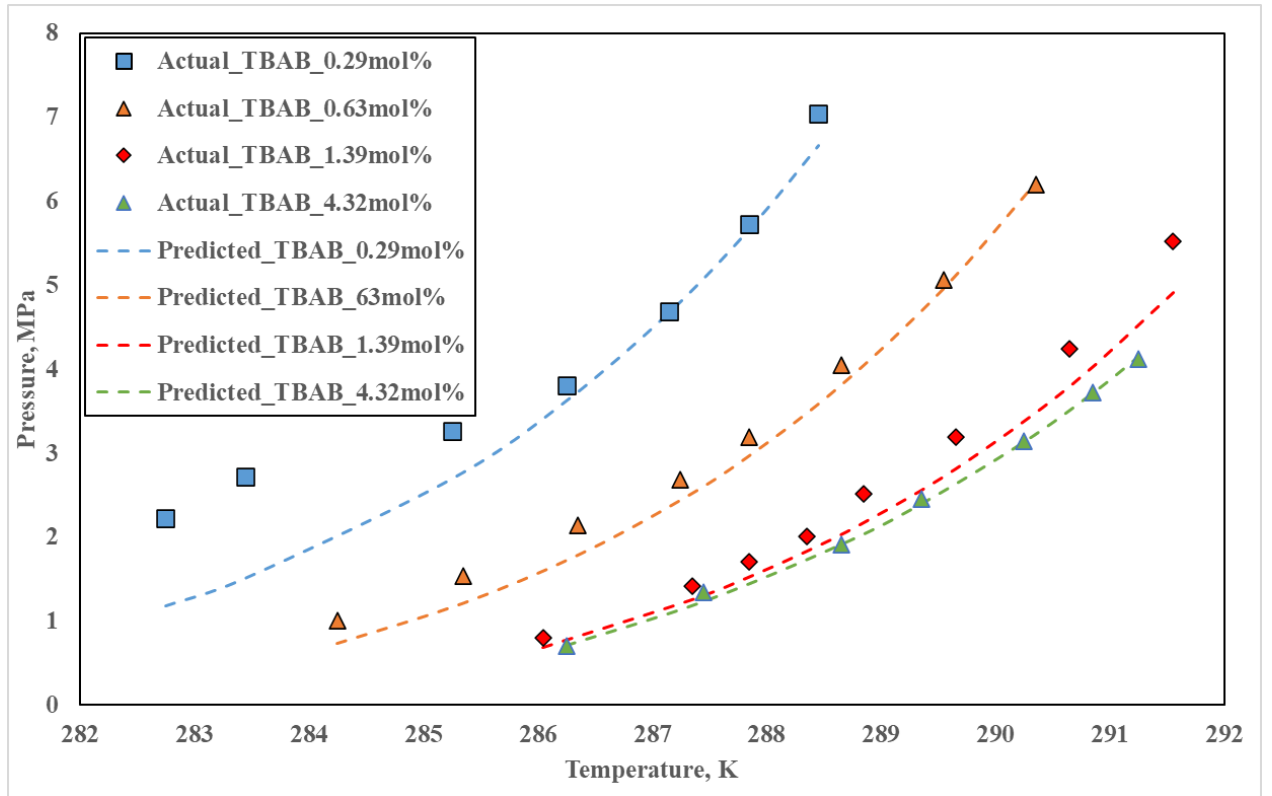


Figure 41. SVR predicted vs actual for different promoter concentrations for SC.

4.4.4 External Validation — SVR

The SVR model trained on the semi-clathrate dataset was applied to an external validation set of 7 data points sourced from a paper not included in the training or testing dataset. The results to evaluate the performance of predictive ability of the model are shown in Figure 42. This figure represents the difference between the actual data from the paper and the predicted values from the model.

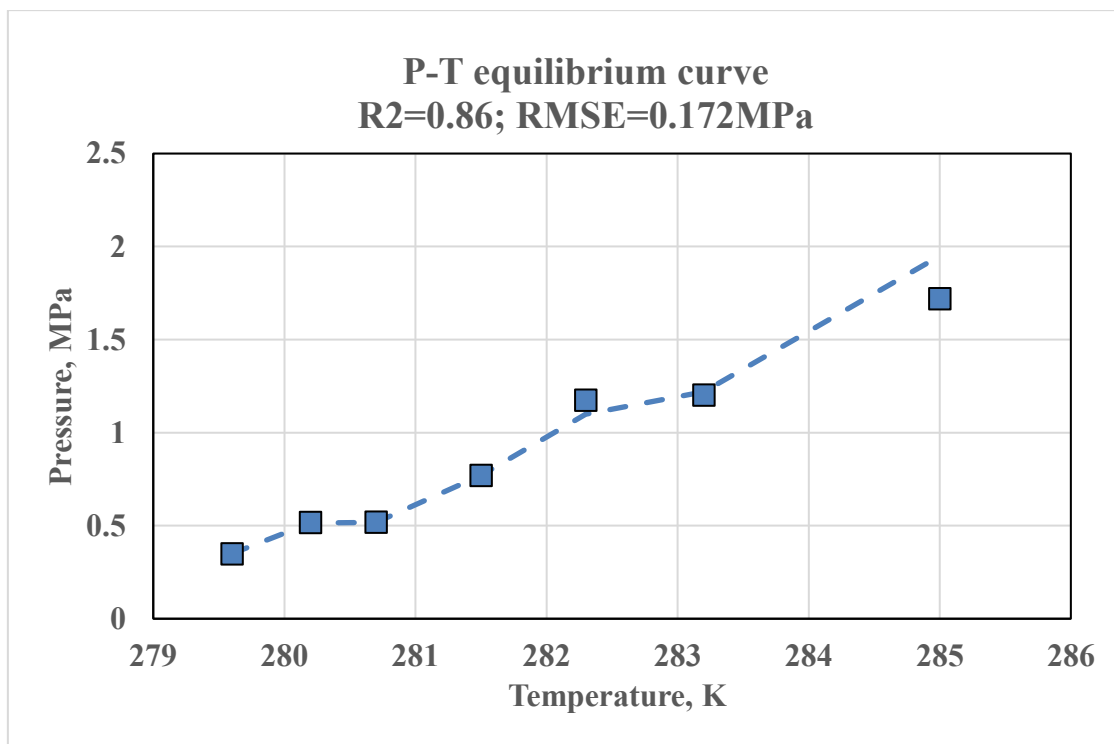


Figure 42. P-T Equilibrium curve for experimental and predicted values with SC SVR model.

The SVR model trained on the semi-clathrate dataset was applied to an external validation set of 7 data points. Table 16 reports the external validation performance.

Table 16. External validation performance - semi-clathrate SVR model.

Metric	Value
R ²	0.86
RMSE	0.172 MPa
MAE	0.114 MPa
No. of external data points	7

External validation for the semi-clathrate hydrate SVR model reveals excellent generalization performance with an R² score of 0.86 and an RMSE of 0.172 MPa for 7 unseen external points. A MAE of 0.114 MPa suggests that the SVR can predict equilibrium pressure to within ~0.11 MPa for the external points on average, which is physically significant in the context of semi-clathrate hydrate processing, where pressures significantly below those found for pure gas-water hydrates are common. Similar to the analysis for water-insoluble hydrates, the RMSE obtained through external validation (0.172 MPa) is substantially lower than that

obtained using the test set (0.724 MPa), suggesting that the external validation points are primarily located in feature space regions which are densely populated by training samples; most likely TBAB with methane or CO₂ at intermediate concentration levels, which forms the bulk of the semi-clathrate training data. A slightly reduced R² value (0.934 for the test set vs. 0.86 for external validation) can be attributed to potential differences in methodology between our training dataset and the external validation dataset.

4.5 Gas-Phase Promoters

4.5.1 Model Performance

Table 17 summarizes the test set performance metrics for all four algorithms applied to the gas-phase promoter dataset (150 data points, 2 gas systems, 80/20 split).

Table 17. Test set performance metrics for gas-phase promoter models.

Algorithm	R ²	RMSE (MPa)	AIC	BIC
Random Forest	0.56	1.43	31.5	39.3
XGBoost	0.56	1.44	31.7	39.5
CatBoost	0.56	1.44	31.9	39.6
SVR	0.67	1.25	24.2	31.9

Table 18 summarizes the hyperparameters that were used during testing and training of the ML models on the gas-phase dataset.

Table 18. Hyperparameters summary for gas-phase promoters' category for all models.

Algorithm	Hyperparameters	Optimal hyperparameters
Random Forest	n_estimators = [100, 200, 300] max_features = ['sqrt', 'log2', 0.7] max_depth = [2, 4, 6, 8, 10] min_samples_split = [2] min_samples_leaf = [1] bootstrap = [True]	n_estimators = 100 max_features = 'log2' max_depth = 6 min_samples_split = 2 min_samples_leaf = 1 bootstrap = True
XGBoost	'n_estimators': [300, 500, 800], 'max_depth': [3, 4, 5, 6],	n_estimators = 800 max_depth = 5

	'learning_rate': [0.03, 0.05, 0.1], 'min_child_weight': [1, 3, 5], 'gamma': [0, 0.1, 0.3], 'subsample': [0.8, 1.0], 'colsample_bytree': [0.8, 1.0], 'reg_alpha': [0, 0.1], 'reg_lambda': [1, 2]	learning_rate = 0.05 min_child_weight = 3 gamma = 0.3 subsample = 0.8 colsample_bytree = 1.0 reg_alpha = 0.1 reg_lambda = 2
CatBoost	'iterations': [500, 1000], 'learning_rate': [0.05, 0.08, 0.1], 'depth': [6, 7, 8], 'l2_leaf_reg': [3, 5, 7], 'bagging_temperature': [0, 1], 'random_strength': [1, 10], 'border_count': [64, 128], 'loss_function': ['RMSE']	iterations = 1000 learning_rate = 0.08 depth = 7 l2_leaf_reg = 5 bagging_temperature = 1 random_strength = 10 border_count = 128 loss_function = 'RMSE'
SVR	'svr__C': loguniform(15, 500), 'svr__epsilon': loguniform(1e-6, 3e-2), 'svr__gamma': loguniform(7e-2, 5e-1)	svr__C = 124.6 svr__epsilon = 0.0021 svr__gamma = 0.34

Figures 43-50 represent the scatter plots for all the ML models for the water-insoluble dataset for training and testing.

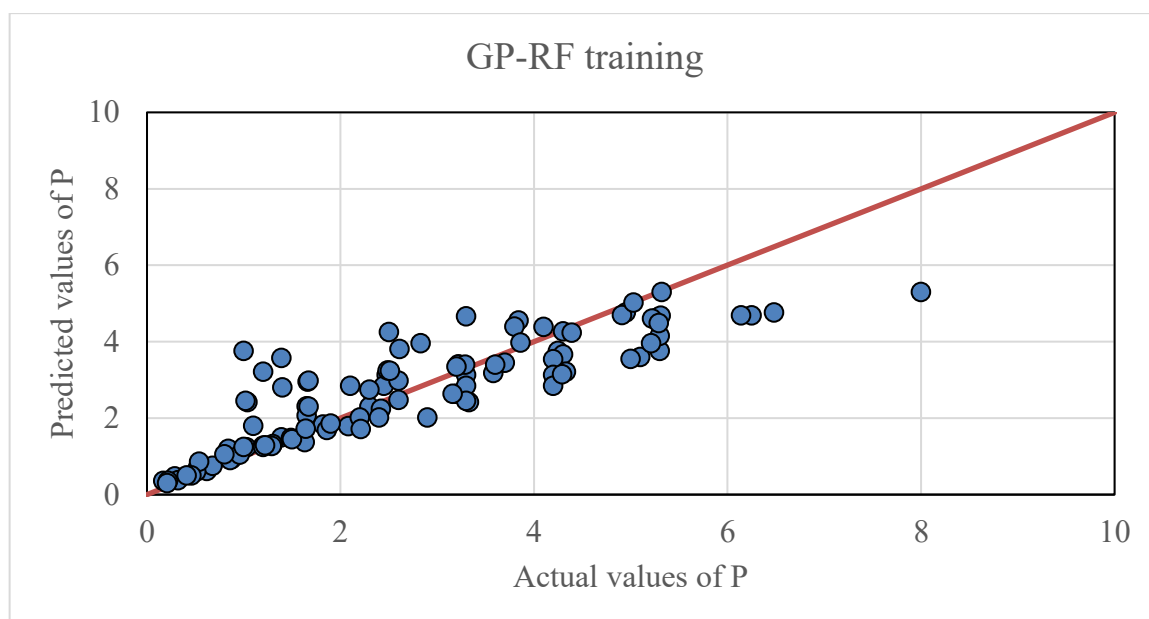


Figure 43. RF Scatter Plot for Gas-phase promoter training dataset

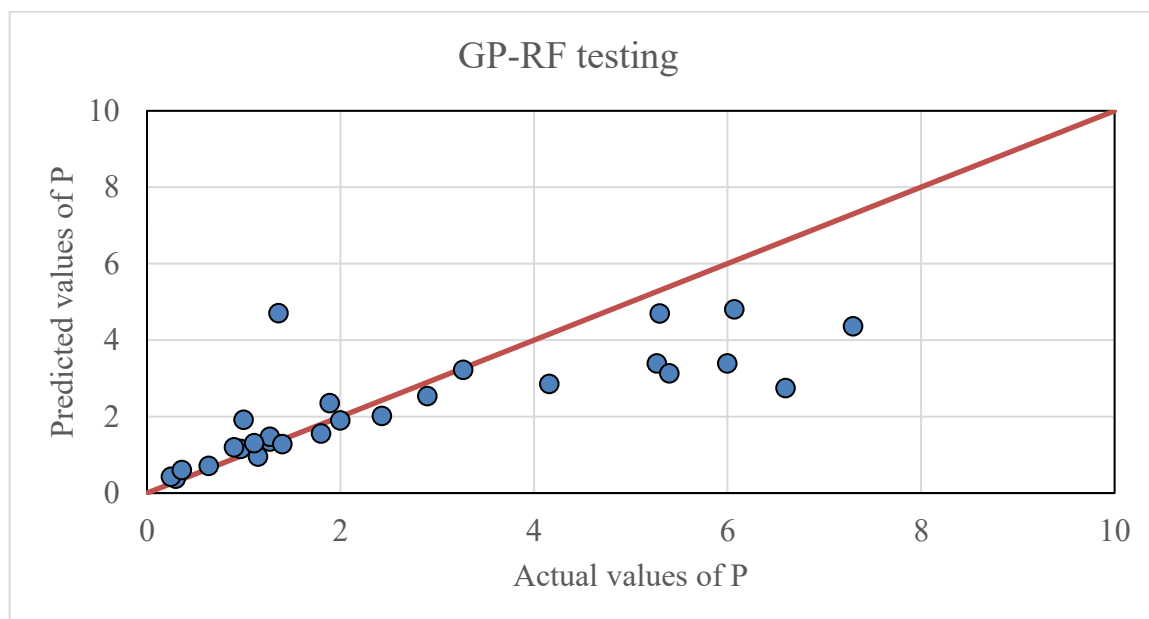


Figure 44. RF Scatter Plot for Gas-phase promoter testing dataset

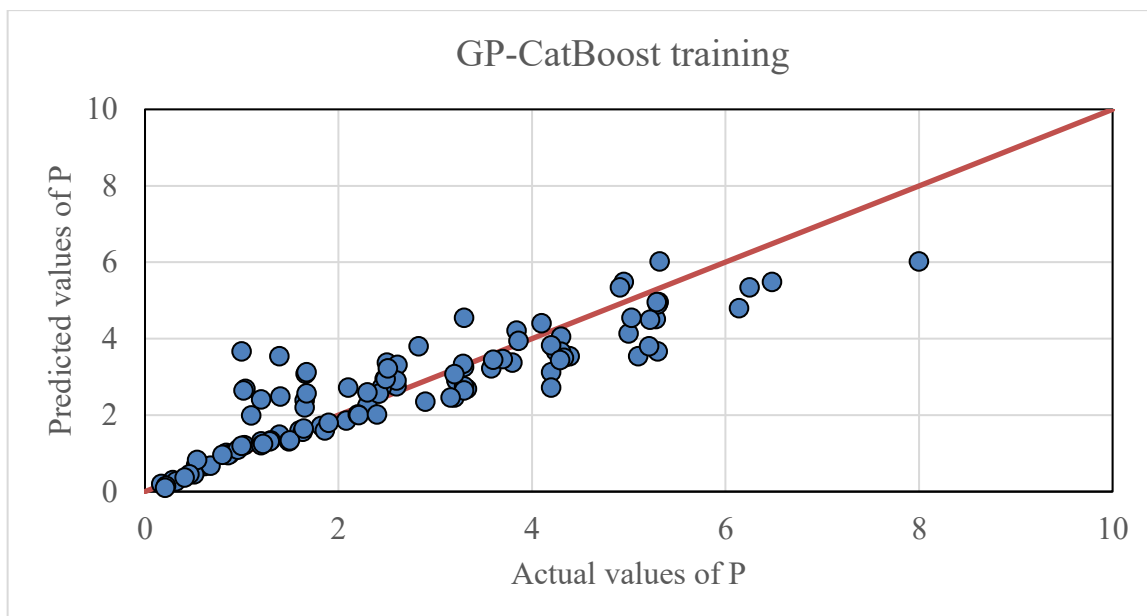


Figure 45. CatBoost Scatter Plot for Gas-phase promoter training dataset

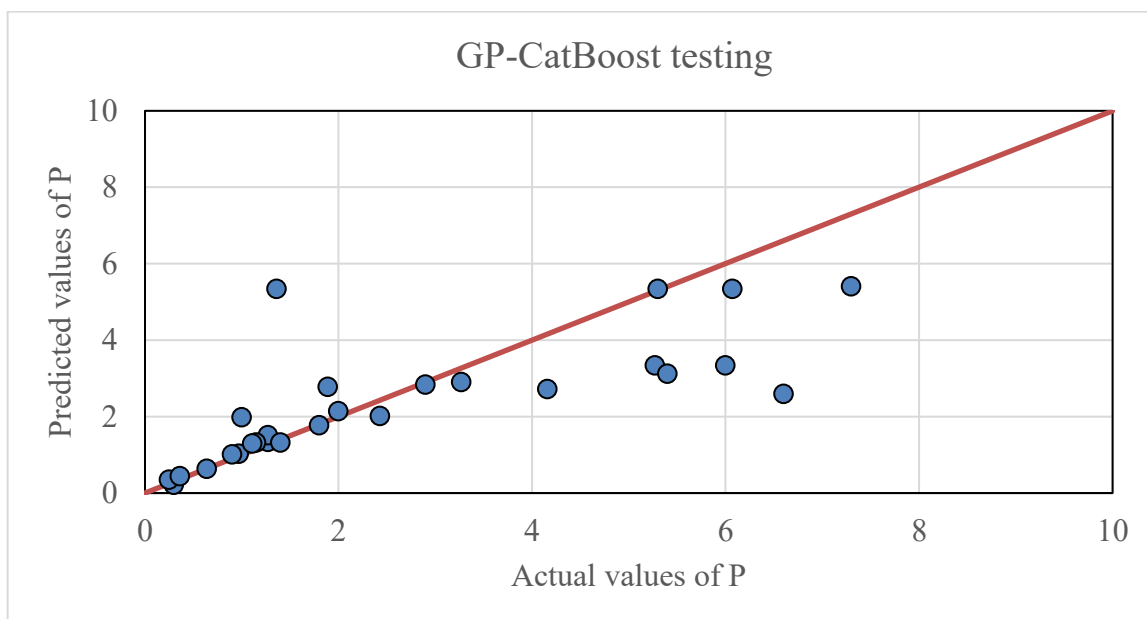


Figure 46. CatBoost Scatter Plot for Gas-phase promoter testing dataset

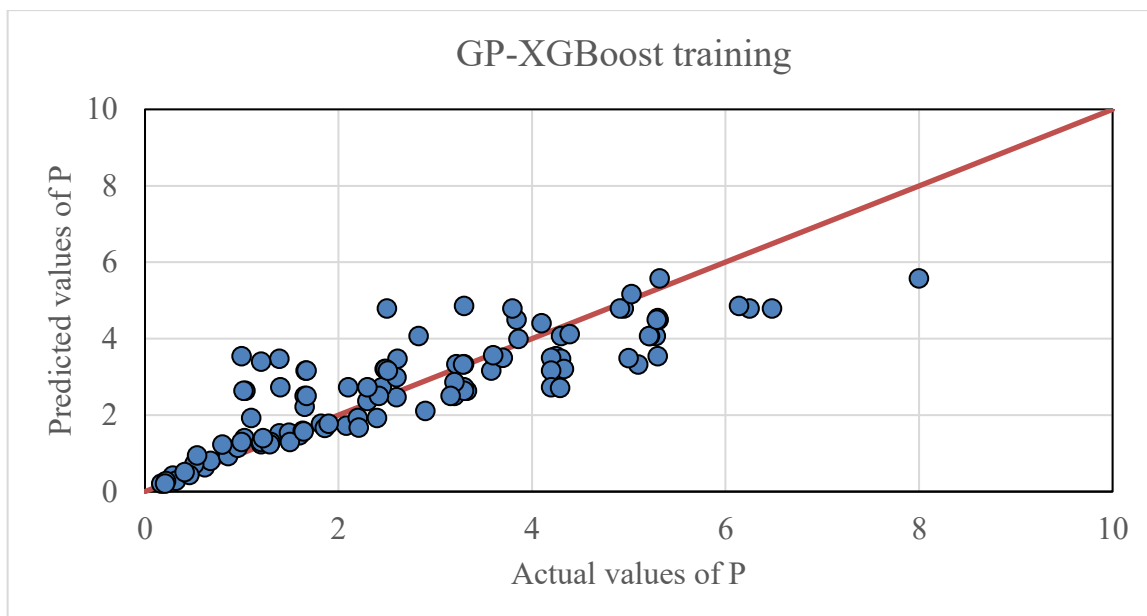


Figure 47. XGBoost Scatter Plot for Gas-phase promoter training dataset

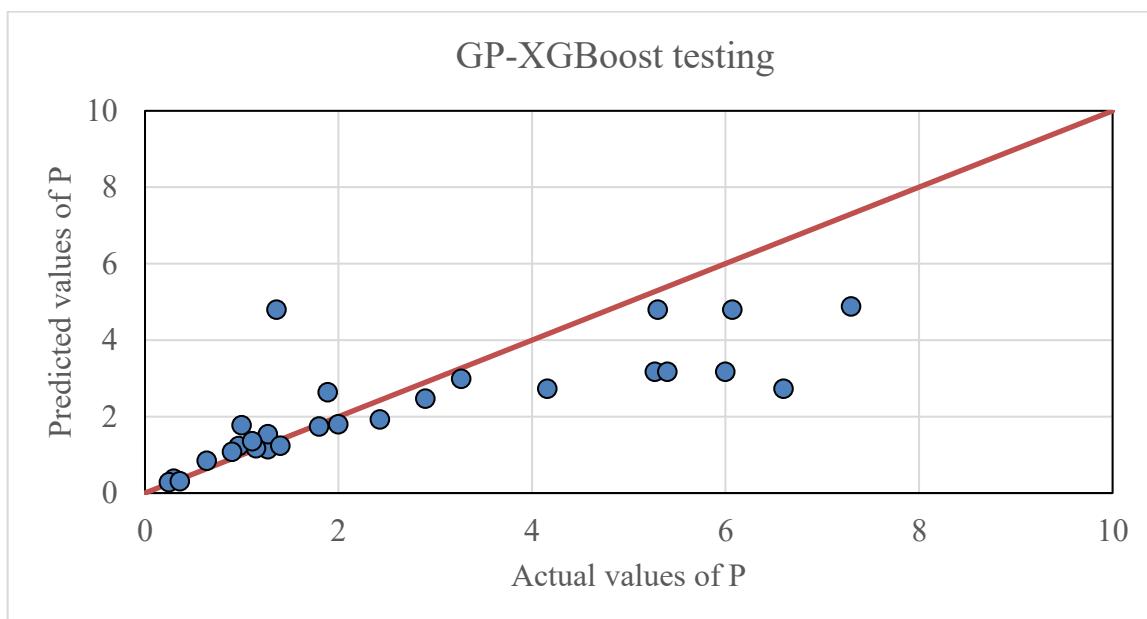


Figure 48. XGBoost Scatter Plot for Gas-phase promoter testing dataset

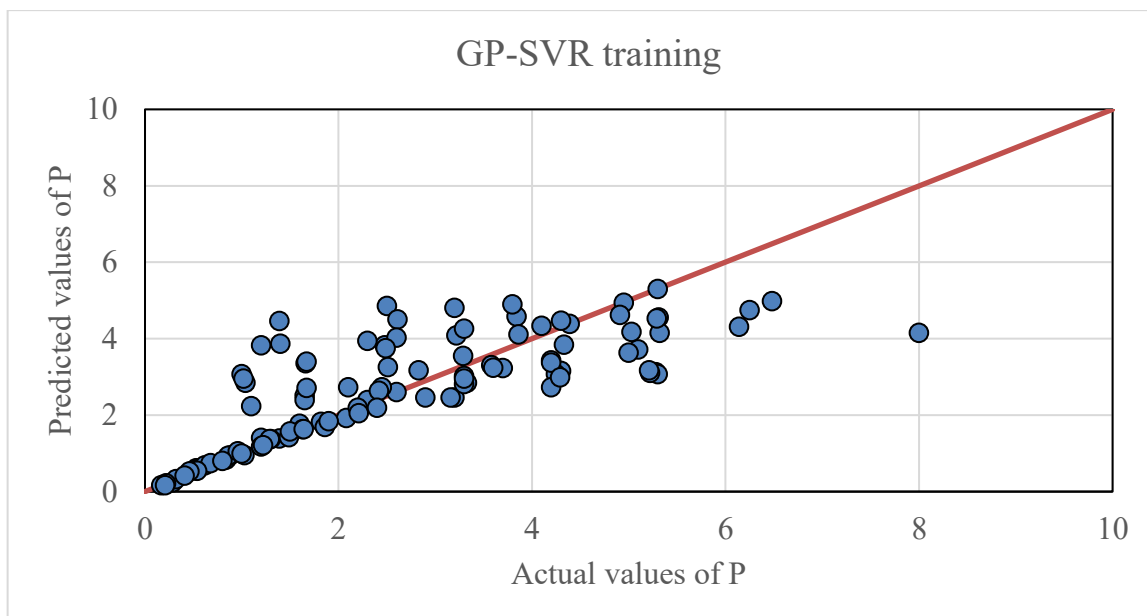


Figure 49. SVR Scatter Plot for Gas-phase promoter training dataset

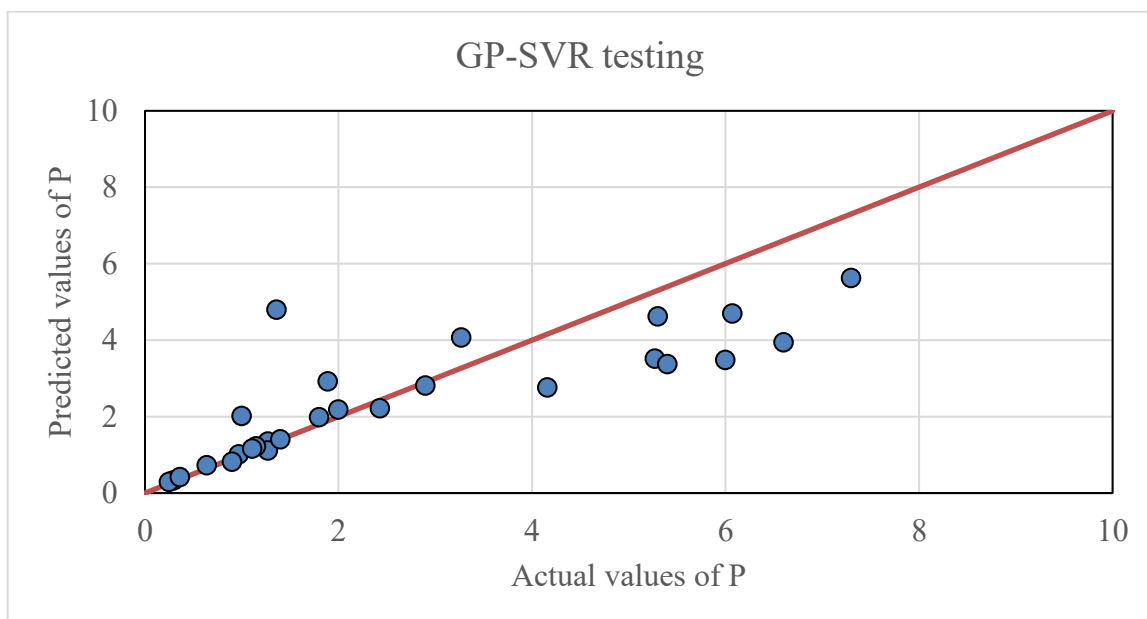


Figure 50. SVR Scatter Plot for Gas-phase promoter testing dataset

Among all categories of the dataset, the gas-phase promoter type resulted in the lowest performances, having R^2 ranging from 0.56 to 0.67 across all algorithms considered. RF, XGBoost, and CatBoost were very similar to each other in terms of performance ($R^2 = 0.56$, $RMSE \approx 1.43 - 1.44$ MPa), whereas SVR showed slightly better results ($R^2 = 0.67$, $RMSE = 1.25$ MPa). The fact that tree-based algorithms RF, XGBoost, and CatBoost have almost converged in their results implies that the highest level of prediction quality achievable using these algorithms is already achieved. Therefore, switching to a different tree-based algorithm or changing parameters of a particular algorithm cannot result in a significant improvement in

model performance. The reason why SVR shows slight superiority compared to tree-based algorithms is probably that the RBF kernel can exploit the moderately positive linear relationship between temperature and pressure in gas-phase systems ($r \approx +0.70$; see Fig. 3.5).

The universally low R^2 values observed in this case reflect the fact that the issues discussed earlier when analyzing the data and described in Chapter 3 directly affect model predictive power. There are at least three reasons for low predictive performance in this case. Firstly, the relatively small number of points (about 150) does not allow learning the underlying structure for any algorithm. Secondly, perfect multicollinearity in gas composition variables (CH_4 , C_2H_6 , and promoter mol%), meaning that $r \approx \pm 1.00$, makes some input variables essentially identical in information content, which reduces the dimensions of the feature space. Thirdly, the lack of a distinct molecule of the promoter, whose characteristics should be encoded in molecular descriptors, and thus the presence of just a binary categorical feature in place of such descriptors makes learning impossible. While AIC and BIC are quite low for the gas-phase category (24.2 – 39.6), the low R^2 values confirm that the models are accurate enough only moderately despite the potential complexity penalty control.

4.5.2 Best Model — SVR

According to the highest value of R^2 (0.67), lowest RMSE (1.25 MPa), and lowest AIC and BIC, SVR was found to be the best model for gas-phase hydrate. But there is a need to mention here that the superiority of the SVR model over tree-based models is quite small, while the value of R^2 , being 0.67, shows that the variance in the gas phase pressure can be explained by the model up to only 0.67 times, which is not enough to make predictions.

4.5.3 Internal Validation – SVR

To validate the SVR model trained on the GP dataset, the digital model was used to predict the pressure values for the same promoter with different concentrations and for different promoters at the same concentrations. The results are presented in Figures 51 and 52.

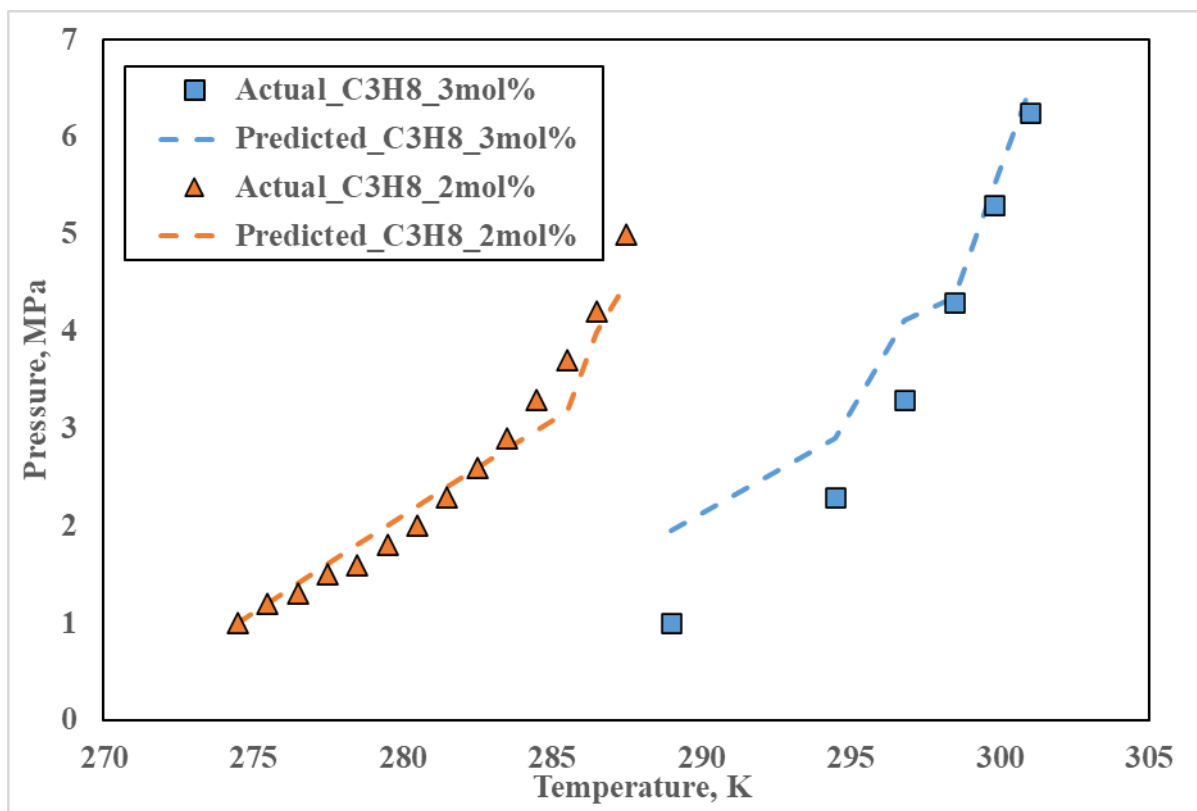


Figure 51. SVR predicted vs actual for same promoter at different concentrations for GP.

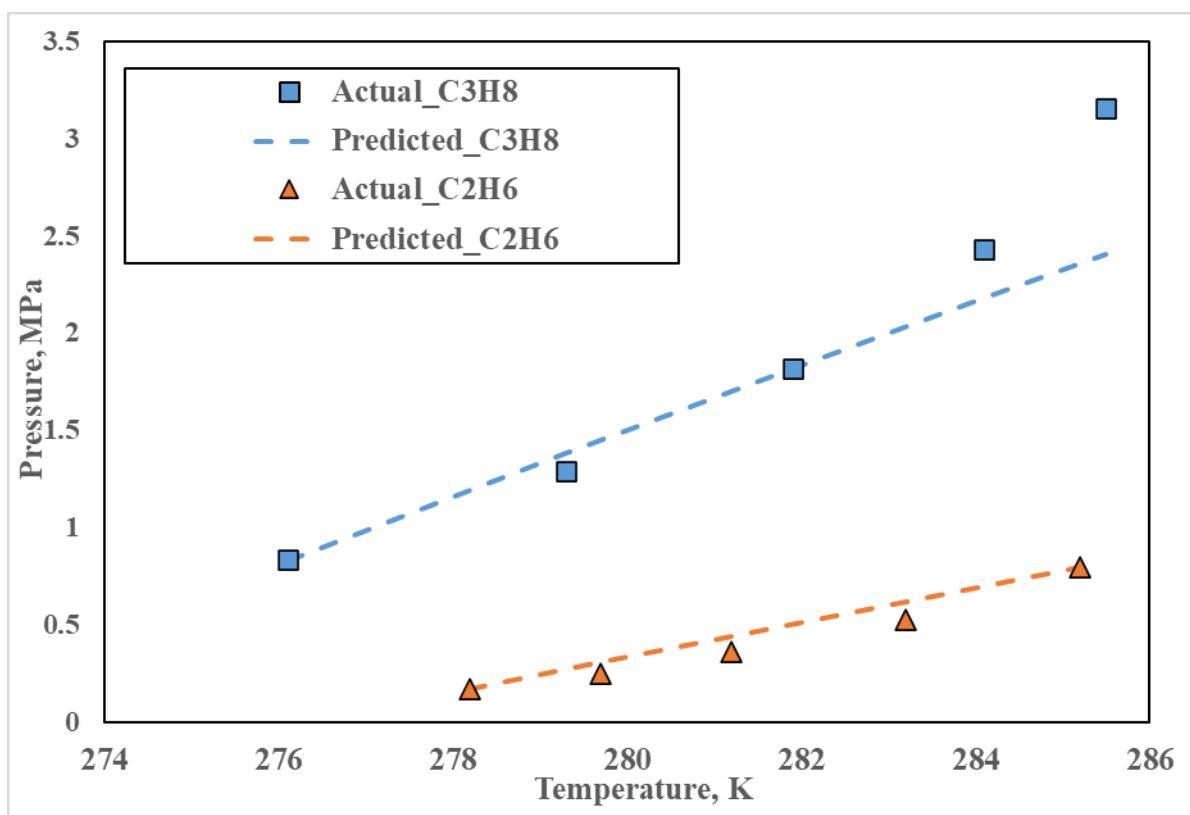


Figure 52. SVR predicted vs actual for different promoter concentrations for GP.

4.5.4 External Validation — Not Performed

No external validation was carried out for the gas-phase promoter category. Due to the small number of points in the gas-phase dataset (150 points for two gases only), it was not possible to determine enough points that were kept out of both the training and test datasets to conduct external validation. Performing external validation on a small number of data points would only lead to inaccurate and non-representative results concerning the validity of the developed predictive models. It may thus be regarded as a drawback of the current work that no external validation was done for this category of points, and more data on gas-phase hydrates is recommended to be gathered in the future, as mentioned in Chapter 6.

4.6 Summary of Results Across All Categories

Table 19 consolidates the test set and external validation performance of the best-performing model for each category, providing a cross-category overview that serves as the basis for the comparative discussion in Chapter 5.

Table 19. Summary of best model performance across all four promoter categories.

Category	Best model	Train R ²	Train RMSE (MPa)	Train AIC	Train BIC	Ext. R ²	Ext. RMSE (MPa)	Ext. MAE (MPa)
Water-soluble	CatBoost	0.92	0.91	47.6	77.2	0.88	0.356	0.237
Water-insoluble	SVR	0.84	1.12	37.9	62.0	0.97	0.183	0.147
Semi-clathrate	SVR	0.934	0.724	37.9	62.05	0.86	0.172	0.114
Gas-phase	SVR	0.67	1.25	24.2	31.9	N/A	N/A	N/A

Several important observations from Table 19 should be made before the discussion is taken further in Chapter 5. First, SVR is found to be the most successful algorithm for three of the four types: water-insoluble, semi-clathrate, and gas phase, whereas CatBoost is the most successful for the fourth type – water-soluble. The semi-clathrate type produces the highest test

set R^2 value (0.934) and the lowest test set RMSE (0.724 MPa), while the gas phase type produces the lowest test set R^2 value (0.67) and the highest test set RMSE (1.25 MPa). External validation proves the effectiveness of water-insoluble and semi-clathrate SVR models, especially the former (ext. $R^2 = 0.97$). Based on the test set R^2 ordering of the categories, which is semi-clathrate > water-soluble > water-insoluble > gas phase, the comparison of the categories' performance can be performed in Chapter 5.

5. DISCUSSION

5.1 Overview of Discussion Approach

The results presented in Chapter 4 are considered in this chapter. From a within-category perspective, the performance of the four ML algorithms is analyzed in terms of how it relates to the thermodynamic and structural properties of each promoter category, especially considering the insights provided by the internal and external validation results concerning the ability of the model to describe realistic pressure-temperature-composition relations. From a cross-category perspective, the performance of the best models for the four categories is compared to rank them in terms of their ability to generate coherent thermodynamic predictions for promoters and to identify the factors responsible for performance differences among them. The novel idea introduced in this thesis — the use of ML performance results as an indicator of thermodynamic regularity of promoter categories — will be explored throughout this chapter, considering the literature review in Chapter 2.

5.2 Water-Soluble Promoters

5.2.1 Algorithm Comparison and CatBoost Performance

The water-soluble promoter dataset generated the second-best average R^2 of the four datasets using the CatBoost algorithm (CatBoost, $R^2 = 0.92$). All four models had decent performances within the range $R^2 = 0.85$ – 0.92 . It is noteworthy that RF and XGBoost performed similarly ($R^2 = 0.87$) despite their difference boosting technique and were surpassed only slightly by CatBoost ($R^2 = 0.92$). This suggests some important properties about the water-soluble promoter dataset. The near-equivalent performance between the two different gradient boosting methods implies that there is no inherent limitation in either method due to the type of problem, and instead, the dataset itself is responsible for the plateaued performance. Therefore, the advantage of the CatBoost model can be attributed to its ordered boosting, which eliminates

the bias introduced through the prediction shift by calculating an unbiased gradient estimate, and its symmetric trees that generate more stable splits for moderately-sized heterogeneous datasets (Prokhorenkova et al., 2018). It reduced RMSE to approximately 0.91 MPa from ~1.16 MPa and AIC to 47.6 from 62-63.

On the other hand, the relatively poor performance of SVR for the water-soluble promoters ($R^2 = 0.85$, RMSE = 1.26 MPa, AIC = 86.2) contradicts its superiority in other categories, suggesting the need for further discussion. The water-soluble promoters cover a broad range of absolute pressures – from less than 0.2 MPa in THF-promoted CO₂ systems near the stoichiometric concentration of promoter to above 14 MPa in methane systems at sub-stoichiometric concentrations and high temperatures – and show considerable nonlinearity when varying promoter concentrations around the stoichiometric concentration point. While the RBF kernel used by SVR can capture smooth nonlinearities, it might be unable to handle sharp concentration changes in the water-soluble promoter dataset since the decision trees' partitioning can be better suited to such tasks. In addition, the significantly higher values of AIC (86.2) and BIC (115.8) for SVR compared to CatBoost (47.6 and 77.2, respectively) support that SVR achieved an R^2 of 0.85 at a much higher effective model complexity.

5.2.2 Internal Validation: Water-Soluble Promoters (CatBoost)

The validation process of the CatBoost predictive model used for water-soluble promoters consisted of comparing the predicted pressure value with the real experimental one. This comparison took place under three different conditions of analysis: variation in THF content; gas mixture composition at constant THF content; and promoter identity at constant concentration. The outcomes of this evaluation can be found in Figures 11, 12, and 13, respectively.

Effect of THF Concentration

Figure 15 illustrates prediction results of the CatBoost model for THF at two concentrations – 1.92 mol% and 7.7 mol% – at temperatures between ~276 and 288 K. The model correctly depicts the thermodynamic correlation between pressure and temperature, showing a qualitative trend towards an exponential increase in pressure as temperature rises. Moreover, the model accounts for the promotion effect that depends on concentration: equilibrium pressures are always lower in the case of a higher THF concentration (7.7 mol%) as compared to a lower one (1.92 mol%) at the same temperatures due to a stronger thermodynamic promotion at higher THF concentration reported earlier.

In the case of the 1.92 mol% series, pressure varies from approximately 0.10 MPa at 276 K to 1.87 MPa at 287 K. Here, the model overestimates the pressure at high temperatures, i.e., the predicted value is about 1.75 MPa at 287 K vs. the experimental result of 1.87 MPa. In turn, at moderate temperatures within ~281-283 K, the predicted pressure is approximately 0.50-0.65 MPa, which correlates well with real values of 0.53-0.65 MPa. As for the 7.7 mol% series, pressure varies between approximately 0.27 MPa at 281 K and 1.44 MPa at 287 K. In this case, the model again overestimates at high temperature, i.e., exceeds the actual value at 287 K by the predicted curve. However, at moderate temperatures of 282-283 K, model values of ~0.46-0.59 MPa correlate well with experimental values of ~0.46-0.59 MPa.

Effect of Gas Composition at Fixed THF Concentration

Predictions from Figure 16 show that the developed CatBoost model can predict hydrate equilibrium pressure of THF for two binary gas compositions, 17 mol% CO₂/83 mol% N₂ and 70 mol% CO₂/30 mol% N₂, within a temperature range of approximately 277 to 296 K and a pressure range of approximately 0.2 to 12.8 MPa. These predictions aim to determine how sensitive the model is to changes in the composition of gases.

In terms of the two binary gas compositions, the model has successfully distinguished between the two gas systems and shows that the composition with higher percentage content of nitrogen (17 mol% CO₂/83 mol% N₂) needs higher pressure than the composition with higher percentage content of carbon dioxide (70 mol% CO₂/30 mol% N₂) at equivalent temperatures; this finding correlates with the thermodynamics because nitrogen is less hydrate forming compared to carbon dioxide. The composition 17 mol% CO₂/83 mol% N₂ ranges from actual pressures of approximately 0.2 MPa at 277 K to 12.8 MPa at 295 K. In this case, the model predicts pressures accurately at low temperatures, which correspond to pressures near 277–282 K; at higher temperatures, the predictions gradually deviate, and predicted pressures are 11.5 MPa at 295 K, while the actual pressures are 12.8 MPa.

The other composition, 70 mol% CO₂/30 mol% N₂, ranges from actual pressures of approximately 0.35 MPa at 279 K to 9.1 MPa at 295 K. Like in the first composition, the CatBoost model predicts pressure accurately at lower temperatures; the predictions are near 3.2 MPa at around 289 K. At higher temperatures, the model slightly underpredicts the pressure, with predictions reaching near 8.8 MPa while the actual pressures are 9.1 MPa. These gradual deviations at higher temperatures in the two gas compositions suggest that the CatBoost model captures pressure dependence on gas composition in the mid-range pressures but underpredicts

near the high-pressure boundaries due to the exponential nature of the curves for predicting pressure of gas hydrate equilibria.

Effect of Promoter Identity at Fixed Concentration

As can be seen from Figure 17, predictions have been made for four water-soluble promoters – THF, 1,4-dioxane, acetone, and propylene oxide – with each of these promoters being considered at 3 mol% concentrations in an aggregate temperature interval between approximately 279–303 K and pressure interval between approximately 1.9–9 MPa. Such a comparison aims to analyze the performance of the proposed modeling approach in predicting the difference in promoters that share equal concentrations but possess different physicochemical properties.

The ranking order of promotion effectiveness of these four compounds is properly accounted for by the CatBoost model in question. Thus, the model correctly identifies acetone as being the most thermodynamically effective promoter (being able to perform in temperatures as low as approximately 279–284 K at the same pressures), while THF, with temperatures required to be as high as approximately 303 K to reach pressures similar to those that acetone reaches at approximately 284 K, represents the weakest promotion effect. As can be seen from Figure 13, the model considers the unique molecular structure of each promoter and utilizes corresponding molecular descriptors (including dipole moment, LogP, and molecular weight).

The accuracy of CatBoost model predictions on a per-compound basis is relatively good in terms of lower pressure values. Acetone has been shown to produce a prediction that agrees quite well with the actual value of approximately 2.0 MPa at 280 K; however, there is some overestimation of pressures in this case, with the model predicting a trend that goes further than the actual value of approximately 7.7 MPa near 284 K. The behavior of the model is quite similar when applied to 1,4-dioxane: while this promoter is accurately predicted at approximately 1.8 MPa at 280 K, there is an increasing deviation from the actual value at approximately 8.9 MPa near 292 K. Propylene oxide is characterized by closer approximation (the range is from approximately 2.0 to 8.7 MPa), though there is still slight underestimation at the highest actual value near 293 K. THF at 3 mol% demonstrates the most significant deviations both ways: with the model predicting pressures larger or smaller than actuals in the temperature range from approximately 291 to 303 K (actual pressure ranging from approximately 2.0 to 9.0 MPa).

5.2.3 External Validation Interpretation

The external validation performance metrics of the CatBoost model ($R^2 = 0.88$, RMSE = 0.356 MPa, MAE = 0.237 MPa for 9 points) show that the trained model generalizes well beyond the training and testing sets. The lower RMSE on the external validation set compared to the test set (0.356 MPa vs. 0.91 MPa) indicates that the external validation points lie in a part of the feature space for which the training data provides an accurate representation, which we speculate to be nearly stoichiometric THF and either methane or CO₂ as guest gases, forming the densest and least noisy part of the water-soluble dataset. An MAE of 0.237 MPa means that on average, the model predicts equilibrium pressure within approximately 0.24 MPa of experimentally measured values, which is physically meaningful because pressure drops of 1-10 MPa are common for the water-soluble promoter category relative to pure water.

5.2.4 Physical Interpretation

The excellent performance of machine learning models for the water-soluble promoter category ($R^2 = 0.85$ – 0.92) reflects the thermodynamically predictable nature of the category: all six promoter compounds form sII hydrates using the same cage occupancy mechanism, resulting in a dataset where the non-linear pressure-temperature-concentration relationships reflect a consistent physical reality that machine learning models can accurately predict using a training set of approximately 560 samples (80% of 700 points). The utility of the molecular descriptor features in capturing the physical meaning of the water-soluble promoters dataset, specifically using dipole moment and LogP as proxies for polarity and hydrophobicity, is demonstrated by the promoter comparison internal validation, showing that the models can rank-order promoters without categorical input. This result implies that three molecular descriptors provide enough physicochemical information for machine learning models to learn the ring geometry and polarity-stability properties described in Chapter 2 (Wu et al., 2023; Seo et al., 2001).

5.3 Water-Insoluble Promoters

5.3.1 Algorithm Comparison and SVR Performance

In terms of spread in performance, the water-insoluble promoters category showed a wider range of R^2 values than any other category, excluding semi-clathrate (XGBoost $R^2 = 0.68$ to SVR $R^2 = 0.84$). It is especially worth mentioning the poor performance of the XGBoost algorithm compared to CatBoost, LightGBM, and SVR. Despite the proven performance of

XGBoost in engineering problems and its competitive performance on the water-soluble category (as seen above), XGBoost achieves only $R^2 = 0.68$ and $RMSE = 1.51$ MPa for the water-insoluble data set. These results are considerably worse than those obtained by CatBoost ($R^2 = 0.83$) and SVR ($R^2 = 0.84$). This result is associated with the problem of overfitting by XGBoost, which uses aggressive tree growth and has low regularization in comparison with CatBoost (Bentéjac et al., 2019; Prokhorenkova et al., 2018). At the same time, the almost equal performance of CatBoost and SVR in RMSE (1.12 MPa), AIC (37.9), and BIC (62.0), with a small advantage of SVR in R^2 (0.84 versus 0.83) and scatter plot distribution suggest that these two algorithms reach the same performance limit on this data set in different ways.

The performance limit of $R^2 \approx 0.84$ for the water-insoluble category is lower than the one reached for the water-soluble category ($R^2 = 0.92$). This limitation is explained by the high heterogeneity of the water-insoluble data set. According to Figure 4 from Chapter 3 and the discussion provided in Chapter 3, the water-insoluble category includes both sII-forming compounds (cyclopentane, cyclohexane, methylcyclopentane, neopentane) and sH-forming compounds (methylcyclohexane, DMCH). Since the P-T relation for these structural groups is different (the P-T conditions of sH systems are always higher than those of sII due to the increased size of the large cages in sH systems (Nakamura et al., 2003)), learning a single model that would incorporate predictions for two different P-T behavior types is impossible without irreducible prediction errors.

5.3.2 Internal Validation: Water-Insoluble Promoters (SVR)

The internal validation of the SVR model for the water-insoluble promoters' class has been carried out under three experimental conditions; the first was one was the influence of gas composition on hydrate equilibrium behavior, the second one was the study of the influence of CP concentration on hydrate equilibrium pressure, and the third one was the influence of the promoter under the constant concentration condition. These results are presented in Figures 27, 28, and 29, respectively.

Effect of Gas Composition

Figure 27 shows predictions made by the SVR model of two different gas species in the gas phase – fluoromethane (CH_3F) and krypton (Kr) – each evaluated at 100 mol% in the range of 280–301 K and 0.10–2.05 MPa, respectively. In the water-insoluble dataset, the source of variability in the gas composition comes from the gas identity at constant promoter conditions,

since the guest gas molecule dictates the occupation of the hydrate cages and thermodynamic equilibria.

As seen from Figure 27, the SVR model is able to reproduce the thermodynamic properties of the two gas systems accurately in terms of their relative behaviour: Kr has higher pressures than CH₃F within the same range of temperatures due to their respective hydrate stability curves. Actual values of Kr pressures range from 0.11 MPa at 280 K to 2.05 MPa at 297 K. Predictions show good agreement with experiments at lower-to-mid temperatures, with the prediction curve following the experiment curve closely between 280–288 K. The predictions show some slight overpredictions at mid-to-high temperatures at intermediate pressures, but converge perfectly with the experiment curve at the last observation point (at 297 K ~2.05 MPa).

Actual pressures of CH₃F range from 0.10 MPa at 284 K to 1.93 MPa at 301 K. Predictions show accurate convergence with experiment results at lower-to-mid temperatures, with actual values ranging from 0.10–0.90 MPa being reproduced well. However, at the highest temperature end (near 301 K), predictions underestimate the experiment value, which is ~1.72 MPa against an actual value of ~1.93 MPa, resulting in a difference of ~0.21 MPa. Underestimation at the higher temperature limit of the data follows the trend observed in previous SVR models, since training points become fewer and kernel interpolation becomes less constrained at the limits of the feature space.

Effect of CP Concentration

Figure 28 illustrates the results of SVR modeling of CP at four different concentrations – 0.25, 0.50, 0.75, and 1.00 mol%. These values pertain to a temperature interval of about 276–292 K and to a pressure interval of about 0.70 to 3.65 MPa. The figure is considered to represent the best results of internal validation obtained among all other results about water-insoluble additives.

Predictions for 0.25 mol% are accurate within the entire range from about 1.50 MPa at 276 K to about 3.57 MPa at about 283 K, with the predicted curve following observations without significant deviations. Predicted values of 0.50 mol% are very consistent with observations in an interval between about 1.65 MPa at 279 K and about 3.65 MPa at 284 K. 0.75 mol% predictions are also quite accurate and follow the trend in observations, but have some small upward deviation of the predicted curve only at the highest point. The range covered by predictions of the 1.00 mol% case starts with the lowest value of about 0.70 MPa at 286 K

and ends with about 3.43 MPa at 292 K, which also corresponds to observations accurately. Overall, all predictions pertaining to the four-concentration series follow their corresponding observations quite accurately and show only minor deviations in each series, indicating successful learning of the CP concentration-pressure-temperature dependence.

Effect of Promoter Identity at Fixed Concentration

Figure 29 displays the pressure–temperature curves of CP and FluoroCP, both at 16 mol% in a temperature range of approximately 280 to 294 K and a pressure range of approximately 0.07 to 4.10 MPa. This plot examines the ability of the model to distinguish between the two similar compounds in terms of their different promotion effects, as encoded in the set of molecular descriptors LogP, dipole moment, and molecular weight.

As expected, the SVR model successfully distinguishes the qualitative behavior of the two promoters, with CP exhibiting lower equilibrium pressures compared to FluoroCP at equivalent temperatures. Thus, the thermodynamic promotion of CP is more effective, which is also reproduced by the proposed model. In terms of the CP pressure–temperature curve, the real pressures are distributed from approximately 0.10 MPa at 280 K to 4.10 MPa near 292–293 K. Here, the curve of the SVR predictions shows good agreement in the mid-range pressures: at approximately 285 K, both actual and predicted values are nearly 0.60 MPa, with an underestimation occurring only at the high-temperature end of the range near 292–293 K, when the actual values (3.70–4.10 MPa) exceed predicted ones.

Concerning the FluoroCP curve, the actual pressures vary from approximately 0.07 MPa at 283 K to 3.60 MPa at 294 K. In this case, too, the predicted values closely follow the actual ones over the mid-range (287–291 K, approximately 0.45–1.35 MPa), whereas a small underestimation occurs at both ends of the range. As for both promoters, the exponential increase of the pressure–temperature dependence is well captured by the SVR predictions, albeit with decreased accuracy near the maximum pressure values due to a known limitation of kernel methods based on the training data density.

5.3.3 External Validation Interpretation

The performance of the water-insoluble SVR model on the external validation dataset ($R^2 = 0.97$, RMSE = 0.183 MPa, MAE = 0.147 MPa for $n = 4$ points) is the best external validation performance among all categories investigated in this thesis. The R^2 value of 0.97 obtained on the external validation dataset is considerably better compared to the test R^2 of 0.84, although this may partially be caused by the low number of points used ($n = 4$), yet still

demonstrating that the SVR model accurately describes the underlying thermodynamic dependencies of water-insoluble promoters with a high level of accuracy for the conditions present within the external validation dataset. The RMSE of 0.183 MPa and MAE of 0.147 MPa are extremely good results considering the pressure range in the water-insoluble promoter dataset, implying that the external validation points are located mostly in the well-known regions of the feature space — most probably the systems of cyclopentane and cyclohexane with methane under moderate pressures, constituting the most significant part of the water-insoluble dataset. It should be emphasized that the external validation dataset consists of only four points, which is a weakness, but the obtained results can be viewed as very promising; however, more data are needed.

5.3.4 Physical Interpretation

The moderately successful predictions made by the water-insoluble category of models can be attributed to two problems: firstly, the structural diversity due to the mixture of both sII and sH-forming structures within the same dataset; secondly, the lack of promoter concentration as a free variable in most samples obtained in excess conditions. While the validity results prove that MW, dipole moment, and LogP make efficient molecular descriptors when discriminating between promoter molecules, the lower value of correlation coefficient in comparison to that of water-soluble models (0.84 vs 0.92) indicates that the addition of certain new structural descriptors, like an additional indicator of either sII or sH structure and cage molecular diameter, may provide better predictions for more diverse datasets. The excellent external validations obtained in the cyclopentane environment confirm that the highest prediction accuracy is achieved by the model trained on the subset demonstrating the highest degree of homogeneity, whereas the lower bound of $R^2 = 0.84$ observed when testing the whole category represents a limit imposed by the structural heterogeneity caused by the sH structures.

5.4 Semi-Clathrate Promoters

5.4.1 Algorithm Comparison and Dramatic SVR Dominance

Among all categories examined in this thesis, the semi-clathrate category generated the largest difference in algorithmic performances, with SVR delivering $R^2 = 0.934$ and RMSE = 0.724 MPa compared to XGBoost, the most successful tree-based method in this category, with $R^2 = 0.71$ and RMSE = 1.50 MPa. Thus, SVR delivered over twice the improvement over the baseline relative to tree-based algorithms as measured by the R^2 value and under half the prediction error as measured by the RMSE. Moreover, despite generating some of the poorest

predictions by tree-based algorithms for this category (RF, $R^2 = 0.54$), the semi-clathrate category achieves the highest R^2 for any category by SVR. This seemingly paradoxical result can be explained by examining the nature of correlations in this dataset.

Indeed, according to the exploratory data analysis (Fig. 5), the semi-clathrate category features a very strong negative linear relationship between the variables of temperature and pressure ($r \approx -0.98$). This deterministic relation implies that temperature accounts for almost 96% of pressure variance in terms of a smoothly declining monotonic function. Since SVR utilizes radial basis functions (RBFs) for kernel-based similarity assessment, the latter is capable of capturing this near-deterministic relationship between temperature and pressure by finding a globally smooth inverse function with minimum complexity (Smola & Schölkopf, 2004). In comparison, tree-based algorithms, which attempt to approximate this global monotonic inverse function in terms of piecewise constant or linear segments, would necessarily require many of them to fit this global function. Given RF's bagging approach with highly variable shallow trees, it would especially be difficult to approximate the smooth inverse T-P function via bagged averaging of the piecewise approximations. Therefore, SVR would be able to deliver superior results for this type of data compared to tree-based algorithms.

Finally, the performance difference between RF and XGBoost/CatBoost is also telling: the latter two generate much higher prediction scores (0.71 for XGBoost and 0.63 for CatBoost) compared to RF, despite all three being decision trees. While both algorithms employ a boosting strategy to fit the underlying data model, RF utilizes bagging, which fits many highly variable shallow trees in parallel. In comparison, XGBoost performs the fit sequentially, thus providing for continuous refinement of the fitted function from the previous step. Furthermore, CatBoost employs the same boosting strategy as XGBoost, with the key distinction of using symmetric trees, which are less variable than asymmetric trees but still perform approximate fitting in a piecewise manner (thus making them less suitable for this problem than SVR). This point is confirmed by the values of AIC (BIC) for tree-based and SVR models; the former are much larger (120.5 – 175.5 vs. 37.9 for SVR), implying higher effective model complexity.

5.4.2 Internal Validation: Semi-Clathrate Promoters (SVR)

The internal validation of the SVR model for the semi-clathrate promoter type was carried out under three different test scenarios: the influence of the TBAB concentration on the equilibrium pressure of the hydrates, the influence of different promoters at the same

concentration, and the influence of the gas mixture composition on the equilibrium behavior of the hydrates.

Effect of Promoter Identity at Fixed Concentration

Figure 39 compares the predictions for two semi-clathrate promoters – tetra-n-butylammonium bromide (TBAB) and tetra-n-butylphosphonium bromide (TBPB), both tested at 0.29 mol% within a temperature range of ~281.7 to 287.5 K and a pressure range of ~1.00 to 5.25 MPa. This comparison tests the model's ability to discriminate between similar compounds with different ionicities represented by semi-clathrate-specific descriptors: the number of cations, the number of anions, and the length of the carbon chain.

For both compounds, there is a very narrow spread in pressure versus temperature at this promoter concentration. The model can reproduce this similarity due to their similar thermodynamic promotional efficacy at this concentration level. The prediction for TBPB is quite accurate, as the prediction follows the experimental observations quite closely. For instance, at 282.3 K, the predicted value is about 1.10 MPa while the observed pressure is 1.20 MPa; at 284.2 K, the agreement is nearly perfect (~2.22 MPa); and finally, at the upper limit near 287.3 K, the predicted pressure is around 5.00 MPa compared to the observed pressure of 5.10 MPa. Thus, in each case, the deviation from the experimental value does not exceed 0.15 MPa.

At 0.29 mol%, TBAB shows some differences in its behavior when compared to TBPB. In particular, the model is unable to accurately reproduce the experimental observations at the lower end of the temperature range; in other words, the predicted pressures for TBAB significantly underestimate the actual values. Indeed, at 281.7 K, the predicted value is ~0.82 MPa while the observed pressure is 1.00 MPa; at 284 K, the predicted pressure is about 1.60 MPa while the observed pressure is 2.25 MPa; and at 285.7 K, the predicted pressure is approximately 2.80 MPa while the observed pressure is 3.55 MPa. However, as the temperature increases further, the accuracy of the predictions increases, and the curves converge at the high end (near 287.3–287.5 K; about 4.35–5.25 MPa).

Effect of Gas Composition

Figure 40 represents the predictions generated by the SVR model for two gas systems: H₂ and CH₄ (100 mol%), both in the presence of the TBAB semi-clathrate, covering the combined temperature ranges of ~280 to ~294 K and the pressure ranges of ~2.5 to ~22.5 MPa. This comparison reflects the most thermodynamically extreme internal validation scenario in

the family of semi-clathrates due to the opposite nature of the respective gases' hydrate formation abilities.

The model appropriately detects the qualitative difference between the two systems: the semi-clathrate system involving H_2 gas displays an exceptionally sharp pressure versus temperature relationship over a very narrow temperature range (~ 280 to ~ 283.3 K) where the equilibrium pressure increases from ~ 4.2 MPa to ~ 22.5 MPa over just 3.3 K. On the other hand, the CH_4 system functions over substantially higher temperatures (~ 287 to ~ 294 K) for an equivalent range of pressures, which can be attributed to the significantly lower hydrate formation capability of methane as compared to that of hydrogen.

Regarding the CH_4 system, the SVR model performs adequately throughout a substantial portion of the data range. For instance, the model predicts pressure to be ~ 2.80 MPa at 287 K, overestimating the actual value of ~ 2.50 MPa. Moreover, the prediction at 289.3 K (~ 4.60 MPa) correlates quite well with the actual value of ~ 4.50 MPa, as does that at 290.7 K (~ 6.50 MPa). At the upper part of the graph near 294 K, the model underpredicts the value by about ~ 0.90 MPa, predicting ~ 15.50 MPa, whereas the actual pressure was measured at ~ 16.40 MPa.

As far as the H_2 system is concerned, the sharpness of the pressure versus temperature curve constitutes a much more difficult problem for the model to cope with. The prediction near 280 K is relatively accurate (~ 4.80 MPa compared to ~ 4.20 MPa), as is the prediction at the midpoint of the graph at 282.2 K (~ 13.50 MPa predicted versus ~ 13.60 MPa actual). At the upper part of the graph near 283.3 K, the prediction of ~ 22.0 MPa is close to the actual value of ~ 22.5 MPa; however, the irregularity of interpolation across the points in this narrow temperature range is noticeable. The reason behind the difficulties faced by the model in handling this dataset is that it deals with the exceptional steepness of the pressure versus temperature relationship, which imposes significant requirements upon the SVR kernel function. Despite this limitation, the model succeeds in distinguishing the H_2 system from the CH_4 system qualitatively, thus encoding gas identity as a deterministic variable.

Effect of TBAB Concentration

Predictions by the SVR model at four different concentrations of TBAB (0.29, 0.63, 1.39, and 4.32 mol%) are depicted in Figure 41 over a temperature range of about 282.5 to 291.5 K and a pressure range of about 0.70 to 7.05 MPa. The prediction model faithfully replicates the concentration-dependent behavior of TBAB in the hydrate equilibria: with the

increase in concentration, the pressure increases systematically for the same temperature, in accordance with the well-known thermodynamic promotion effect of quaternary ammonium salts in semi-clathrate hydrate systems.

Model accuracy depends on the level of the studied concentration. For the highest concentration (4.32 mol%), the predicted hydrate equilibria curve demonstrates the best agreement with experiments over the whole temperature interval. The real pressures vary in the range between 0.70 and 4.10 MPa for temperatures from 286.2 K to 291.5 K, respectively, and all discrepancies are not larger than 0.10 MPa. At 1.39 mol%, similar accuracy can be achieved up to an overestimation of about 0.65 MPa, occurring only at the highest temperature in the studied range (near 291.5 K, 5.55 MPa experimentally). The model works reasonably for 0.63 mol%, too, with a good agreement observed from 285.5 K to almost 291.5 K and a moderate deviation at the very top of the series (at 290.5 K, the difference is about 0.20 MPa, whereas the actual pressure is around 6.20 MPa).

The largest deviations occur for 0.29 mol%. The model underestimates the value at the lowest temperature, where the discrepancy is of the order of 1.0 MPa. A much better match is achieved for intermediate temperatures, when both curves coincide in their values of about 3.25 MPa at 285.5 K. A slight underprediction is registered at the upper boundary, with the real value exceeding the predicted one by about 0.55 MPa (actual ~7.05, predicted ~6.50). Thus, we may assume that the SVR model performs poorly at low TBAB concentrations, especially at the lowest temperatures, because there is insufficient data for reliable interpolation in this part of the feature space.

5.4.3 External Validation Interpretation

The external validation performance of the semi-clathrate SVR model ($R^2 = 0.86$, RMSE = 0.172 MPa, MAE = 0.114 MPa for 7 external points) confirms high generalization ability of the trained SVR model to new data. In particular, the lowest external validation MAE of all categories (0.114 MPa) suggests that the semi-clathrate model is able to predict the semi-clathrate hydrate equilibrium pressure to an average absolute error of about 0.11 MPa. This is considered an accurate prediction because the typical reduction in equilibrium pressure due to semi-clathrate promoters varies between 1 to 10 MPa relative to the pure gas-water mixture. However, the relatively small difference in R^2 performance between the test set (0.934) and external validation (0.86) confirms the expected performance decline due to application of the model on data obtained from another experimental source with different methodological

conditions. Nevertheless, the external validation for seven points is a significantly stronger check for model generalization compared to the previous four points for the water-insoluble category. As such, the successful prediction for the semi-clathrate category can be interpreted as a successful learning of the semi-clathrate SVR model for the underlying thermodynamic relations from the training data.

5.4.4 Physical Interpretation

The outstanding performance of SVR in the case of the semi-clathrate category (test $R^2 = 0.934$) is clear evidence that semi-clathrate hydrate equilibrium pressure can be predicted deterministically from temperature, concentration, and ionic descriptors when an adequate learning algorithm is used. The near-deterministic relation between T and P found in the semi-clathrate data set is the result of the pronounced negative slope of the semi-clathrate equilibrium curve, which stems from the high dissociation enthalpy and strong ionic lattice stabilization of the semi-clathrate type hydrates (Lin et al., 2008; Zhou et al., 2018). This combination of factors makes semi-clathrate hydrates the most predictable and hence the most thermodynamically regular out of all four hydrate categories. Thus, the spectacular difference in accuracy achieved by tree-based and SVR algorithms on the semi-clathrate data set can be interpreted as a clear indication that the thermodynamics of semi-clathrate hydrates is smooth and globally regular enough to be adequately described using SVR.

5.5 Gas-Phase Promoters

5.5.1 Algorithm Convergence and Performance Limitations

Gas-phase models have shown the worst overall performance, with all R^2 values ranging between 0.56 (for RF, XGBoost, and CatBoost) and 0.67 (for SVR). Convergence of the performance of three completely different kinds of models – bagging, sequential boosting, and ordered boosting – towards the nearly identical $R^2 = 0.56$ and RMSE of about 1.43–1.44 MPa points to the fact that the best possible performance on this dataset has been achieved and the main issue preventing further improvement is related not to the choice of algorithm, but the underlying nature of the gas-phase data. If models with different inductive biases show similar performance, this points to the fact that there is no more information that can be extracted from the feature set in question (González et al., 2020; Talaei Khoei & Kaabouch, 2023).

Poor performance of the gas-phase models can be explained by the combination of three interrelated factors revealed in the exploratory data analysis carried out in Chapter 3. First, there

are not many observations in the dataset (~150 points), which implies relatively poor training signals, since the data refers to two types of gas phases with different thermodynamics: binary and ternary combinations. Second, perfect multicollinearity between gas phase components (CH_4 , C_2H_6 , and promoters in %mol) with correlations of $r \approx \pm 1.00$ significantly constrains the feature space, thereby restricting learning opportunities of models. Third, binary categorical features (0 for ethane, 1 for propane) provide little information about thermodynamic promotion in gas-phase hydrate formation and do not allow capturing the structural differences between ethane- and propane-promoted sII hydrates found in experimental data (Holder & Hand, 1982; de Menezes et al., 2019).

A marginal superiority of SVR over the tree-based models ($R^2 = 0.67$ vs $R^2 = 0.56$) in terms of the gas-phase dataset may be because SVR exploits relatively weak, but positive linear correlation between temperature and pressure ($r \approx +0.70$) via fitting the RBF kernel to capture the smooth T–P dependence on a small dataset better than tree-based piecewise models. Nevertheless, despite showing better performance compared to tree-based models, SVR still achieves only about two-thirds of the variance in pressure values, which is far from being sufficient for making quantitative predictions.

5.5.2 Internal Validation: Gas-Phase Promoters (SVR)

Internal validation of the SVR model for the gas-phase promoter category was conducted across two analytical conditions: the effect of propane concentration in the gas phase on hydrate equilibrium pressure and a comparison of propane (C_3H_8) and ethane (C_2H_6) as gas-phase promoters. As noted in Section 5.5.1, the gas-phase category is characterized by the smallest dataset among the four promoter categories (~150 data points), and external validation was not performed for this category. The results are presented in Figures 36 and 37, respectively.

Effect of Propane Concentration in the Gas Phase

Modelled SVR predictions of the effect of propane concentration on hydrate equilibrium for propane in the gas phase at two concentrations (2 mol% and 3 mol%) for temperatures ranging from ~274 to ~302 K and pressures from ~1.0 to ~6.25 MPa are shown in Figure 51. Here, we assess the sensitivity of the model to the concentration of a guest gas promoter in its mixtures (propane in our case), which defines the occupancies of the hydrate cages and their equilibrium thermodynamics.

The concentration sensitivity is well predicted by the model because the 3 mol% propane mixtures operate at a significantly higher temperature at comparable pressure ranges, indicating stronger thermodynamic promotion of the hydrate formation in the mixtures containing larger concentrations of a guest gas promoter.

For the 2 mol% propane mixtures, which include many data points between ~274 to ~286.5 K, the model predicts hydrate formation with noticeably higher accuracy. The predicted pressure is very close to the real data values between ~274 K (~1.0 MPa) and around ~282–285 K (~2.3–3.7 MPa), where the difference between the predicted and actual values does not exceed ~0.1 MPa. For higher temperatures at ~286–286.5 K, the model underestimates the last two data points of the series, with the last value of ~5.0 MPa deviating from the prediction by about 0.6 MPa.

For the series of 3 mol% propane, which includes only five points between ~289 to ~301 K, the model demonstrates less accurate predictions with larger deviations compared to the previous case. At the lower temperature value of ~289 K, the pressure predicted by the model (of ~2.0 MPa) largely overpredicts the experimental data (of ~1.0 MPa), resulting in a deviation of ~1.0 MPa. Better agreement can be observed for intermediate temperatures with the convergence of predictions and measurements in the range of ~295–300 K (~2.3–5.25 MPa), while at the highest temperature value of ~301 K the model reproduces the experimental data reasonably accurately (of ~6.10 MPa predicted and ~6.25 MPa measured). The strong overprediction of the lowest temperatures for the 3 mol% propane series, along with the lower overall prediction accuracy when compared to the 2 mol% propane series, indicates a poorer dataset for training purposes, limiting the ability of the model to predict the correct pressure–temperature dependence of the system in this portion of the feature space.

In summary, for the gas-phase category, we see that the SVR model can capture the thermodynamics of both promoter identities and their concentrations, although with lower quantitative accuracy. This is due to the small size of the dataset used, the binary nature of the identity variable that allows limited continuous variation, and the high nonlinearity of the pressure-temperature relationship in these types of promoters. It reflects in the overall R^2 of only 0.67 for the gas-phase category, being the worst-performing promoter category out of all four.

Effect of Promoter Identity: Propane versus Ethane

Figure 52 provides the predictions generated by the SVR models using propane (C_3H_8) and ethane (C_2H_6) gas-phase hydrate promoters over a temperature range of approximately 276 to 285.5 K and a pressure range of approximately 0.14 to 3.15 MPa. In these models, the two promoters are represented using a single binary-valued predictor (0 for ethane; 1 for propane), which is the only predictor used to differentiate the two gases from one another in the absence of molecular descriptors.

The model correctly captures the qualitative separation between these two promoters: C_3H_8 always requires significantly higher equilibrium pressures compared to C_2H_6 at similar temperatures. This fact reflects the superior hydrate formation-promoting effect of the former. In general, C_3H_8 equilibrium pressures are about four to five times greater than those of C_2H_6 at any temperature.

For the latter, the predictions generated by the model are fairly accurate. For example, the exact match occurs at 278 K, where the actual and predicted values of equilibrium pressure are both equal to approximately 0.15 MPa. Further agreement continues through 284 K, where the predicted and actual equilibrium pressures are both approximately 0.54–0.58 MPa. Near the upper boundary at approximately 285.5 K, the deviation of approximately 0.04 MPa can be seen, where the prediction deviates slightly downward (actual 0.80 MPa vs predicted 0.76 MPa).

As for the former, the model gradually begins to underestimate the equilibrium pressure as temperature increases. Thus, at 276 K, where propane is supposed to require the lowest equilibrium pressure, there is a slight overprediction of the true equilibrium pressure – approximately 0.08 MPa. At 279.5 K, the agreement is quite good with the deviations between the actual values of approximately 1.27 MPa and predicted ones being around 1.30 MPa. However, from 281 K onwards, the gap between the observed data and predictions becomes steadily larger: thus, at 281 K, the predicted values of 1.55 MPa (approx.) are significantly lower than the actual ones of approximately 1.83 MPa; for 284 K, the deviations reach approximately 0.37 MPa (approx. predicted 2.05 MPa vs approximate actual value of 2.42 MPa). At the upper boundary of approximately 285.5 K, where the model predicts an equilibrium pressure of 2.38 MPa, the observed values are considerably higher at 3.15 MPa. In general, one noticeable characteristic of the predicted curves of the two promoters in this graph is the linear behavior exhibited by the former, while the real C_3H_8 hydrates exhibit clearly accelerating, non-linear behavior.

5.5.3 Why External Validation Was Not Performed

External validation was not done in the case of the gas-phase category since the dataset obtained was too small (~150 points). The split of 80% training and 20% testing provided about 120 data points for training and 30 for testing. There were no data points that could have been excluded from the dataset before training to serve as an external validation set. External validation with too few data points – especially considering the complex nature and composition variation of gas-phase hydrate systems – could lead to misleading performance results that do not indicate true generalization capability. Thus, no external validation was done for this category because of insufficient data points available in the literature for binary and ternary natural gas hydrate systems under the gas-phase promoter approach used in this thesis.

5.6 Cross-Category Comparative Analysis

5.6.1 Performance Ranking and Thermodynamic Coherence

Table 20 presents the cross-category performance ranking based on test set R^2 of the best model for each category, together with key dataset and structural characteristics that provide the physical context for interpreting the observed ranking.

Table 20. Cross-category performance ranking and structural characteristics.

Rank	Category	Best model	Training R^2	Training RMSE (MPa)	Dominant T-P correlation	Hydrate structure	Dataset size
1	Semi-clathrate	SVR	0.934	0.724	$r = -0.98$	Unique semi-clathrate	620
2	Water-soluble	CatBoost	0.92	0.91	$r = +0.34$	sII (uniform)	700
3	Water-insoluble	SVR	0.84	1.12	$r = -0.21$	sII + sH (mixed)	450
4	Gas-phase	SVR	0.67	1.25	$r = +0.70$	sII (sI-sII competition)	150

It should be noted that the physical interpretability of this hierarchy is obvious – each of the top three categories demonstrates high thermodynamic coherence and homogeneity of structure. The best results in terms of prediction accuracy are shown by the semi-clathrate category due to the high correlation coefficient of the T-P relationship ($r = -0.98$), which is characteristic of the studied dataset due to a very pronounced global equilibrium curve of semi-clathrate structures for all studied compounds. Next on the performance ranking list comes the water-soluble category, thanks to the homogenous structure of the sII hydrate formation mechanism for all 7 water-soluble promoters. The relatively poor results of the water-insoluble category in comparison with the water-soluble category come from the presence of two types of structural families, sII and sH, and weaker relationships between molecular features and pressure levels. And the worst-performing is the gas phase category owing to the small size of the dataset, severe problems with multicollinearity, and a lack of effective molecular descriptor-based feature representation.

This interpretation can be provided based on the theoretical concept of Chapter 3 about the ability of machine learning models to quantitatively reflect the degree of thermodynamic regularity and data quality of the studied system. Thus, under this concept, the semi-clathrate category would demonstrate the highest regularity due to the consistency and stability of the equilibrium relationship between temperature, concentration, and pressure for all 6 compounds of this category. Second place goes to the water-soluble category as a relatively homogeneous category, while the water-insoluble category has relatively low regularity but is disrupted by the presence of structural heterogeneity. And the gas phase category would show the lowest level of regularity, not because of any thermodynamic inconsistencies, but rather due to the lack of appropriate data and features.

5.6.2 SVR Dominance Across Three Categories

The superior performance of SVR over all other algorithms for three out of four categories is another noteworthy finding of this thesis. In general, ensemble methods based on trees – Random Forest (RF) and various boosting techniques – are the most cited top-performing algorithms in the hydrate machine learning literature, followed by support vector regression (SVR), which ranks lower among the best algorithms (Atta et al., 2025). Therefore, this thesis offers an interesting counterexample where SVR consistently achieves superior or comparable accuracy for predicting hydrate phase equilibria using smooth promoter-assisted systems with relatively small datasets compared to ensemble-based tree algorithms.

In terms of physical intuition, the superiority of SVR can be explained by the characteristics of the Gaussian radial basis function kernel and its applicability to hydrate phase equilibrium data. The RBF kernel measures the similarity of two data points as a smooth Gaussian function of the Euclidean distance between them, which results in a smooth regression function defined by interpolation between the training points (Smola & Schölkopf, 2004). The latter feature of the RBF kernel is consistent with the nature of hydrate phase equilibrium data, as the relationship between pressure and temperature in hydrate phase equilibria is a smooth, continuous function of temperature, concentration, and gas mixture composition, except for discontinuities occurring at structural transitions. By contrast, since tree-based models recursively divide the feature space into disjoint rectangular regions to learn a piecewise constant or linear function, the number of trees needed to approximate a smooth continuous function would be high in both cases. For smaller datasets and higher-dimensional feature spaces, this problem would be exacerbated further. Hence, due to the global smoothness of the pressure surface in hydrate phase equilibrium data, the advantage of SVR over tree-based algorithms is evident.

However, there is one exception to the rule, where CatBoost surpasses SVR for the water-soluble category. In this case, strong concentration nonlinearities in the water-soluble dataset near the stoichiometric composition of THF result in sharp pressure gradients that cannot be approximated efficiently by the globally smooth function learned by SVR. In contrast, CatBoost's ordered boosting approach allows capturing such local nonlinearities and sharp gradients by constructing decision trees in order.

5.6.3 Implications for Promoter Selection

In addition to the results of model performance testing and method development described above, quantitative analysis of cross-category model performance provides further understanding of the relative thermodynamic properties of the four promoter categories considered for practical promoter selection.

The exceptional predictability of the semi-clathrate formers ($R^2 = 0.934$) is a result of the thermodynamic predictability of TBAB and other ionic substances used as promoter molecules, which demonstrate a highly consistent promotion effect on different gases under a variety of gas concentrations and promoter concentrations. Consequently, semi-clathrate promoters allow estimation of hydrate formation equilibrium conditions for a target pressure

and temperature with very high accuracy, which is extremely valuable in processes requiring precise pressure control, such as pre-combustion CO₂ capture and hydrogen storage.

The good but slightly weaker predictability of water-soluble promoters ($R^2 = 0.92$) is associated with the consistent structure and thermodynamic effects of THF promotion of hydrate formation. The limited prediction accuracy under extreme concentrations and pressures suggests that water-soluble promoters are most reliable when applied in stoichiometric concentrations and relatively low pressures and temperatures, because the deviation of gas composition or hydrate formation conditions from the studied ranges may increase uncertainty about the prediction results.

The modest predictability of the water-insoluble category ($R^2 = 0.84$) is determined by the higher variability of the category structure compared to the others and the specific difficulty of hydrate formation at interfaces between immiscible liquids. The fact that the predictors based on the concentration data are significantly better for sII than for sH promoters (0.82 vs. 0.68 R^2 for water-insoluble promoters) indicates the lower consistency of thermodynamic effects for sH-forming molecules such as methycyclohexane and DMCH than for those promoting sII structure.

Lastly, the relatively poor predictability of gas-phase promoters ($R^2 = 0.67$) should not be considered as an indicator of the thermodynamic unpredictability of gas-phase-promoted hydrates in general. In contrast, many hydrate-forming gas mixtures such as methane-ethane and methane-ethane-propane systems, widely studied in experimental literature (Holder & Hand, 1982; Nixdorf & Oellrich, 1997), have been demonstrated to exhibit physical regularity in their equilibrium state.

5.7 Limitations of the Study

5.7.1 Dataset Size and Distribution Imbalances

The major drawback of this analysis is the skewed distribution of data points across the compounds in each class. As pointed out in Chapter 3, THF and TBAB, being benchmark compounds in their classes, are much more prevalent in literature and thus more abundant in the data sets than THP, neopentane, or isopropylamine, which are less studied promoters. The consequence of this discrepancy is that the trained models are biased toward compounds having much more data, and may fail to generalize across other compounds in the same class. In future research, this problem needs to be overcome by either collecting new data from the literature

where equilibrium constants were measured for the less studied promoters or through some sort of data balancing technique.

5.7.2 Gas-Phase Category Limitations

However, the category of gas-phase promoter faces three interrelated drawbacks: small dataset size (around 150 data points), strong multicollinearity between variables representing the composition of the gas, and a lack of molecular descriptor variables to describe the promoting gas molecules. It follows that this category cannot be effectively modeled using existing modeling techniques, which makes its results rather uncertain. As can be seen from the $R^2 = 0.67$ of the best model (SVR), no external testing has been performed; therefore, the results of gas-phase promoter modeling should be viewed very skeptically, and their further analysis cannot be provided based on the data obtained in the present study. The suggestions aimed at overcoming the identified issues in the gas-phase promoter modeling are given in Chapter 6.

5.7.3 External Validation Sample Sizes

External validation data sets used for each of the three categories, for which external validation was done, are fairly limited in size. The number of external validation instances is 9 for the water-soluble category, 4 for the water-insoluble category, and 7 for the semi-clathrate category. Even though achieving good results for such limited data is encouraging, due to their small sizes, such sets cannot yield reliable validation results concerning model generalizability. For example, the external validation for the water-insoluble category using only 4 external instances is not enough for an appropriate evaluation of the model's applicability to the whole category.

5.7.4 Feature Engineering Limitations

The molecular descriptor set used in this study, which consists of MW, dipole moment, LogP, and the semi-clathrate-specific ionic descriptors, provides a good yet inevitably simplistic approximation to the physical nature of the promoter compounds. More advanced molecular descriptors, such as cage-specific molecular diameter ratio, polar surface area, or quantum chemical parameters related to electrostatic potential, could perhaps allow one to construct even more discriminating and generalizing features, especially for the water-insoluble type of compounds, because the coexistence of both sII and sH-forming molecules in the same feature space poses certain challenges in terms of modeling. Molecular descriptor set development for hydrate promoters machine learning models is proposed as a promising area for future work.

5.7.5 Absence of Baseline Comparison

Comparison of the model's predictions with those made using classical thermodynamics models like van der Waals-Platteeuw or with commercial software tools like CSMHyd is not part of this thesis. Though not included in this study since its scope did not require it and considering the structural constraints that make such a model inappropriate for modeling semi-clathrates discussed in Chapter 2, a thermodynamic model comparison in future research will give a better understanding of the strengths of the data-driven approach and enable comparisons with Ahmadi et al. (2020)'s results where they established that ANFIS outperforms classical thermodynamic models in CO₂ promoter systems.

6. CONCLUSIONS AND RECOMMENDATIONS

6.1 Conclusions

The thesis carried out an empirical comparison of four categories of thermodynamic promoters of hydrates: water-soluble, water-insoluble, semi-clathrate, and gas phase. For this purpose, predictive machine learning models were developed, trained, and validated to estimate the hydrate equilibrium pressure based on experimental data points available in the literature, which amount to nearly 1,920 data points in total for 21 promoter molecules and 9 guest gases belonging to the four categories. Four machine learning methods were investigated for each promoter category, namely Random Forest, XGBoost, CatBoost, and Support Vector Regression. The model performance was assessed using the metrics R², RMSE, AIC, and BIC, while an independent validation was carried out for three out of four promoter categories based on additional experimental data.

6.1.1 Category-Specific Conclusions

The modeling of water-soluble promoters showed excellent predictivity, with CatBoost performing better than other methods, as evidenced by the value of the squared correlation coefficient R² being equal to 0.92 and the root mean square error being 0.91 MPa. The success of CatBoost over RF, XGBoost, and SVR in this category lies in the fact that the ordered nature of boosting used by the algorithm allows addressing the nonlinearity present around the stoichiometric point of THF content for the water-soluble category. Validation showed that the trained model accurately ranks the efficacy of promoting the reaction for THF, acetone, 1,4-dioxane, and propylene oxide despite the lack of explicit labels for categorization of promoter

types; hence, molecular descriptors like MW, dipole moment, and LogP can capture physical-chemical characteristics of promoters from the water-soluble category.

The most difficult to model was the water-insoluble promoters category, due to the presence of both sII-forming and sH-forming chemicals in one set of data, and because promoter concentration is not used as a variable for this category due to excesses. The best results were obtained by SVR ($R^2 = 0.84$, RMSE = 1.12 MPa). SVR slightly outperformed CatBoost, while the overly aggressive growth of trees by XGBoost, which resulted in overfitting on the more heterogeneous and smaller dataset, caused its inferior performance compared to others ($R^2 = 0.68$). The SVR model showed good internal validation results, successfully distinguishing between structurally similar promoters based solely on the difference in molecular descriptors (cyclopentane and fluorocyclopentane). External validation using four data points yielded the best results in terms of external validation from all categories ($R^2 = 0.97$, MAE = 0.147 MPa).

However, semi-clathrate promoters produced the most pronounced result from all experiments, as SVR produced $R^2 = 0.934$ and RMSE = 0.724 MPa, which is a significantly higher value than obtained for tree-based techniques (the best among them is XGBoost, producing $R^2 = 0.71$). The reason behind such a discrepancy in results lies in a near-deterministic inverse relation between pressure and temperature for the semi-clathrate promoter data ($r \approx -0.98$), which can be better approximated by a smooth RBF function of SVR than by the step-wise division of tree-based ensembles. According to both AIC and BIC metrics, SVR produces a superior fit in terms of significantly less complex models compared to ensemble techniques. Internal validation showed precise prediction of the TBAB concentration effect for four concentration levels, correct differentiation between TBAB and TBPB for diluted solutions, as well as successful reproduction of an extreme effect of the difference between H_2 and CH_4 gas compositions. External validation for seven data points confirmed generalization properties ($R^2 = 0.86$, MAE = 0.114 MPa – the minimal value among all categories).

The gas-phase promoters yielded the lowest accuracy scores across all four algorithms, where three of the four algorithms (RF, XGBoost, CatBoost) all converged on an R^2 score of 0.56, while the only slightly higher values were achieved by the SVR algorithm, at $R^2 = 0.67$ and $RMSE = 1.25$ MPa. The convergence of algorithms to the same level of performance suggests that there is some sort of inherent limitation of the model structure, namely its reliance on small sample size (150 observations), near-perfect multicollinearity between the gas phase composition features ($r \approx \pm 1.00$), and binary (0/1) representation of promoters, which does not provide enough continuous variability for the model to learn the curvature of the curve. Validation using external datasets was not conducted since there was no additional data available.

6.1.2 Cross-Cutting Methodological Conclusions

The semi-clathrate > water-soluble > water-insoluble > gas-phase cross-category performance ranking (R^2 values of 0.934, 0.92, 0.84, and 0.67, respectively) indicates a hierarchical order of physical sense and data quality in all four promoter categories. These results provide direct experimental support to the main methodological hypothesis proposed by this thesis: machine learning models' performance measured using both information-theoretic measures (AIC and BIC) and conventional predictive performance metrics (R^2 and RMSE) is indicative of the thermodynamic regularity of a particular promoter category as well as the appropriateness of the data compilation used for its modeling.

SVR with an RBF kernel turned out to be the most successful algorithm in three out of four studied categories, offering a unique contribution to the field of ML applied to hydrates, where tree ensembles have become the go-to methods in previous research. The current findings suggest that the ability of SVR to globally optimize smoothly is well-suited to the physical consistency and continuity of phase equilibrium hydrate data at moderate sample sizes, while CatBoost's boosting approach is preferable only if strong non-linearities are present in the target variable (e.g., the influence of stoichiometric concentration on water-soluble systems).

The molecular descriptor feature engineering technique chosen for this study was corroborated by the success of the internal validation performed for all three categories, with successful modeling, meaning that the descriptors chosen contain enough structural information to be able to make generalizations about other molecules in each promoter category. In turn, the ionic descriptor features, which were specific to the semi-clathrate category, turned out to

be good for distinguishing between quaternary ammonium and phosphonium salt hydrate promoters.

6.2 Recommendations

6.2.1 Data Expansion and Curation

The most important improvement for future research is the comprehensive increase in the size of datasets in each of the four categories, especially the gas-phase category. The current gas-phase dataset (with 150 points) is too small to develop machine learning models that can account for the heterogeneity of the composition and structure of binary/ternary natural gas hydrate systems. Future studies will benefit from gathering more phase equilibrium information on methane-ethane and methane-ethane-propane systems, especially the information that focuses on the intermediate composition and pressure ranges currently lacking in the dataset. Concerning the water-insoluble category and semi-clathrate category, it is essential to gather more information related to the underrepresented compounds, such as neopentane, isopropylamine, and methylcyclopentane to ensure a more uniform data distribution, as mentioned in Section 5.7.1. Additionally, enlarging the external validation sets for each of the four categories beyond the current 4-9 points would make the generalizability test much more statistically significant.

6.2.2 Feature Engineering Enhancements

The molecular descriptor set used in this thesis should be viewed as being a relatively conservative choice of features. Further research could examine the effect of introducing additional molecular descriptors based on more physical and structural properties. In the case of the water-insoluble and gas phase categories, it might become necessary as the present descriptor set failed to provide enough information for achieving satisfactory accuracy for those two sub-sets. For the water-insoluble category, a descriptor reflecting the presence or absence of the sII or sH structure will directly resolve the issue highlighted in Section 5.7.4 about the lack of structure homogeneity information, thus allowing differentiation between the formation of sII and sH structures regardless of MW or LogP difference between molecules. In the case of the gas phase category, a replacement of binary 0/1 promoter encoding with continuous molecular descriptors describing the structure of the promoting gas will allow the inclusion of physicochemical information that the current binary encoding does not contain. As expected, it will significantly increase the ability of the model to describe the nonlinear pressure and temperature relationship seen in propane-promoted systems.

6.2.3 Algorithmic and Modeling Framework Improvements

Further research might consider the use of more complex machine learning algorithms than the ones explored in the present study. For instance, deep learning models, specifically physics-informed neural network (PINN) models that encode the thermodynamic relations from the van der Waals-Platteeuw framework into regularization terms, can be an interesting direction to enhance generalization capabilities on data-scarce classes like the gas-phase promoters class, in which the lack of training data makes ML algorithms hard to use. The use of transfer learning methodologies, whereby a previously trained model using the abundant datasets of the water-soluble and semi-clathrate classes would be fine-tuned using the smaller datasets of the water-insoluble and gas-phase classes, could also contribute to mitigating the data imbalance among classes. Moreover, the addition of a classical thermodynamic model (CSMHyd or vdWP model) as a comparison method to the proposed ML model can serve to quantify the benefits of the latter.

6.2.4 Industrial Application Directions

The machine learning models developed and validated externally during this thesis can be readily used for process screening and initial hydrate process design. In particular, the CatBoost model for water-soluble promoters and the SVR model for semi-clathrates appear especially suitable for use in screening and pre-combustion CO₂ capture and hydrogen storage processes. The latter model, specifically (external $R^2 = 0.86$, MAE = 0.114 MPa), seems highly appropriate in such cases as the potential use of TBAB or TBPB semi-clathrate systems, which are technologically viable options but require rapid estimates of equilibrium pressure in varied conditions to optimize processes. Finally, the CatBoost model for water-solubles (external $R^2 = 0.88$, MAE = 0.237 MPa) represents an effective method of screening THF and similar cyclic ethers for natural gas storage and CO₂ sequestration purposes. Going forward, researchers should aim to incorporate these models into process simulation, specifically using equation-of-state substitutions in popular commercial flow assurance software.

7. REFERENCES

1. Ahmadi, M., Chen, Z., Clarke, M., & Fedutenko, E. (2020). Comparison of kriging, machine learning algorithms and classical thermodynamics for correlating the formation conditions for CO₂ gas hydrates and semi-clathrates. *Journal of Natural Gas Science and Engineering*, 84, 103659.
<https://doi.org/10.1016/j.jngse.2020.103659>
2. Arjmandi, M., Chapoy, A., & Tohidi, B. (2007). Equilibrium Data of Hydrogen, Methane, Nitrogen, Carbon Dioxide, and Natural Gas in Semi-Clathrate Hydrates of Tetrabutyl Ammonium Bromide. *Journal of Chemical & Engineering Data*, 52(6), 2153–2158. <https://doi.org/10.1021/jc700144p>
3. Atta, M. R., Al-Mahmodi, A. F., Lal, B., Abdulrab, H., & Khor, S. F. (2025). Artificial Intelligence and Machine Learning in Thermodynamic Gas Hydrate Studies: A Review. *Energy & Fuels*, 39(38), 18287–18310.
<https://doi.org/10.1021/acs.energyfuels.5c02863>
4. Babu, P., Yao, M., Datta, S., Kumar, R., & Linga, P. (2014). Thermodynamic and Kinetic Verification of Tetra-*n*-butyl Ammonium Nitrate (TBANO₃) as a Promoter for the Clathrate Process Applicable to Precombustion Carbon Dioxide Capture. *Environmental Science & Technology*, 48(6), 3550–3558.
<https://doi.org/10.1021/es4044819>
5. Bentéjac, C., Csörgő, A., & Martínez-Muñoz, G. (2019). *A Comparative Analysis of XGBoost*.
6. Breiman, L. (2001). Random Forests. *Machine Learning*, 45(1), 5–32.
7. Chen, Z.-Y., Li, Q.-P., Yan, Z.-Y., Yan, K.-F., Zeng, Z.-Y., & Li, X.-S. (2010). Phase Equilibrium and Dissociation Enthalpies for Cyclopentane + Methane Hydrates in

- NaCl Aqueous Solutions. *Journal of Chemical & Engineering Data*, 55(10), 4444–4449. <https://doi.org/10.1021/je100597e>
8. de Deugd, R. M., Jager, M. D., & de Swaan Arons, J. (2001). Mixed hydrates of methane and water-soluble hydrocarbons modeling of empirical results. *AIChE Journal*, 47(3), 693–704. <https://doi.org/10.1002/aic.690470316>
 9. Delahaye, A., Fournaison, L., Marinhas, S., Chatti, I., Petitet, J.-P., Dalmazzone, D., & Fürst, W. (2005). Effect of THF on Equilibrium Pressure and Dissociation Enthalpy of CO₂ Hydrates Applied to Secondary Refrigeration. *Industrial & Engineering Chemistry Research*, 45(1), 391–397. <https://doi.org/10.1021/ie050356p>
 10. Elsevier, Tohidi, B., Danesh, A., Todd, A., Burgass, R., & Ostergaard, K. (1997). HIHDPHASE EOUIUIIIA Equilibrium data and thermodynamic modelling of cyclopentane and neopentane hydrates. *Fluid Phase Equilibria*, 138, 241–250.
 11. Eslamimanesh, A., Gharagheizi, F., Mohammadi, A. H., Richon, D., Illbeigi, M., Fazlali, A., Forghani, A. A., & Yazdizadeh, M. (2011). Phase Equilibrium Modeling of Structure H Clathrate Hydrates of Methane + Water “Insoluble” Hydrocarbon Promoter Using Group Contribution-Support Vector Machine Technique. *Industrial & Engineering Chemistry Research*, 50(22), 12807–12814. <https://doi.org/10.1021/ie2011164>
 12. Fan, S., Li, Q., Nie, J., Lang, X., Wen, Y., & Wang, Y. (2013). Semiclathrate Hydrate Phase Equilibrium for CO₂/CH₄ Gas Mixtures in the Presence of Tetrabutylammonium Halide (Bromide, Chloride, or Fluoride). *Journal of Chemical & Engineering Data*, 58(11), 3137–3141. <https://doi.org/10.1021/je4005933>
 13. Fan, S., Li, S., Wang, J., Lang, X., & Wang, Y. (2009). Efficient Capture of CO₂ from Simulated Flue Gas by Formation of TBAB or TBAF Semiclathrate Hydrates. *Energy & Fuels*, 23(8), 4202–4208. <https://doi.org/10.1021/ef9003329>

14. Ghiasi, M. M., Yarveicy, H., Arabloo, M., Mohammadi, A. H., & Behbahani, R. M. (2016). Modeling of stability conditions of natural gas clathrate hydrates using least squares support vector machine approach. *Journal of Molecular Liquids*, 223, 1081–1092. <https://doi.org/10.1016/j.molliq.2016.09.009>
15. 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>
16. Hashimoto, S., Sugahara, T., Sato, H., & Ohgaki, K. (2007). Thermodynamic Stability of H₂ + Tetrahydrofuran Mixed Gas Hydrate in Nonstoichiometric Aqueous Solutions. *Journal of Chemical & Engineering Data*, 52(2), 517–520. <https://doi.org/10.1021/je060436f>
17. Hassan, H., Zebardast, S., & Gandomkar, M. (2021). A new highly stable methane hydrate from the synergic mixture of common promoters: Experimental and modeling surveys. *Journal of Natural Gas Science and Engineering*, 93, 104056. <https://doi.org/10.1016/j.jngse.2021.104056>
18. Herslund, P. J., Daraboina, N., Thomsen, K., Abildskov, J., & von Solms, N. (2014a). Measuring and modelling of the combined thermodynamic promoting effect of tetrahydrofuran and cyclopentane on carbon dioxide hydrates. *Fluid Phase Equilibria*, 381, 20–27. <https://doi.org/10.1016/j.fluid.2014.08.015>
19. Herslund, P. J., Daraboina, N., Thomsen, K., Abildskov, J., & von Solms, N. (2014b). Measuring and modelling of the combined thermodynamic promoting effect of tetrahydrofuran and cyclopentane on carbon dioxide hydrates. *Fluid Phase Equilibria*, 381, 20–27. <https://doi.org/10.1016/j.fluid.2014.08.015>

20. Ho-Van, S., Bouillot, B., Douzet, J., Babakhani, S. M., & Herri, J. M. (2019). Cyclopentane hydrates – A candidate for desalination? *Journal of Environmental Chemical Engineering*, 7(5), 103359. <https://doi.org/10.1016/j.jece.2019.103359>
21. Hughes, T. J. (2008, September 6). *Plug Formation and Dissociation of Mixed Gas Hydrates and Methane Semi-Clathrate Hydrate Stability*.
https://www.researchgate.net/publication/29488237_Plug_Formation_and_Dissociation_of_Mixed_Gas_Hydrates_and_Methane_Semi-Clathrate_Hydrate_Stability
22. Imai, S., Okutani, K., Ohmura, R., & Mori, Y. H. (2005). Phase Equilibrium for Clathrate Hydrates Formed with Difluoromethane + either Cyclopentane or Tetra-*n*-butylammonium Bromide. *Journal of Chemical & Engineering Data*, 50(5), 1783–1786. <https://doi.org/10.1021/je050212h>
23. Joon Shin, H., Lee, Y.-J., Im, J.-H., Won Han, K., Lee, J.-W., Lee, Y., Dong Lee, J., Jang, W.-Y., & Yoon, J.-H. (2009). Thermodynamic stability, spectroscopic identification and cage occupation of binary CO₂ clathrate hydrates. *Chemical Engineering Science*, 64(24), 5125–5130. <https://doi.org/10.1016/j.ces.2009.08.019>
24. 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(1-2), 101–109. [https://doi.org/10.1016/s0378-3812\(01\)00460-5](https://doi.org/10.1016/s0378-3812(01)00460-5)
25. Kim, S., Baek, I.-H., You, J.-K., & Seo, Y. (2015). Guest gas enclathration in tetra-*n*-butyl ammonium chloride (TBAC) semiclathrates: Potential application to natural gas storage and CO₂ capture. *Applied Energy*, 140, 107–112.
<https://doi.org/10.1016/j.apenergy.2014.11.076>
26. Kim, S., Choi, S.-D., & Seo, Y. (2017). CO₂ capture from flue gas using clathrate formation in the presence of thermodynamic promoters. *Energy*, 118, 950–956.
<https://doi.org/10.1016/j.energy.2016.10.122>

27. Komatsu, H., Yoshioka, H., Ota, M., Sato, Y., Watanabe, M., Smith, R. L., & Peters, C. J. (2010). Phase Equilibrium Measurements of Hydrogen–Tetrahydrofuran and Hydrogen–Cyclopentane Binary Clathrate Hydrate Systems. *Journal of Chemical & Engineering Data*, 55(6), 2214–2218. <https://doi.org/10.1021/jc900767h>
28. Kumar, A., Vedula, S. S., Kumar, R., & Linga, P. (2018). Hydrate phase equilibrium data of mixed methane-tetrahydrofuran hydrates in saline water. *The Journal of Chemical Thermodynamics*, 117, 2–8. <https://doi.org/10.1016/j.jct.2017.05.014>
29. 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(6), 2729–2733. <https://doi.org/10.1016/j.energy.2009.05.026>
30. Lee, J., Jin, Y. K., & Seo, Y. (2018). Characterization of cyclopentane clathrates with gaseous guests for gas storage and separation. *Chemical Engineering Journal*, 338, 572–578. <https://doi.org/10.1016/j.cej.2018.01.054>
31. Lee, S., Lee, Y., Park, S., Kim, Y., Cha, I., & Seo, Y. (2013). Stability conditions and guest distribution of the methane + ethane + propane hydrates or semiclathrates in the presence of tetrahydrofuran or quaternary ammonium salts. *The Journal of Chemical Thermodynamics*, 65, 113–119. <https://doi.org/10.1016/j.jct.2013.05.042>
32. Lee, S., Lee, Y., Park, S., & Seo, Y. (2012). Structural Transformation of Isopropylamine Semiclathrate Hydrates in the Presence of Methane as a Coguest. *The Journal of Physical Chemistry B*, 116(45), 13476–13480. <https://doi.org/10.1021/jp308640m>
33. Lee, Y.-J., Kawamura, T., Yamamoto, Y., & Yoon, J.-H. (2012). Phase Equilibrium Studies of Tetrahydrofuran (THF) + CH₄, THF + CO₂, CH₄ + CO₂, and THF + CO₂ + CH₄ Hydrates. *Journal of Chemical & Engineering Data*, 57(12), 3543–3548. <https://doi.org/10.1021/jc300850q>

34. Li, D.-L., Du, J.-W., Fan, S.-S., Liang, D.-Q., Li, X.-S., & Huang, N.-S. (2007).
Clathrate Dissociation Conditions for Methane + Tetra-*n*-butyl Ammonium Bromide
(TBAB) + Water. *Journal of Chemical & Engineering Data*, 52(5), 1916–1918.
<https://doi.org/10.1021/jc700229e>
35. LI, S., FAN, S., WANG, J., LANG, X., & WANG, Y. (2010). Clathrate Hydrate
Capture of CO₂ from Simulated Flue Gas with Cyclopentane/Water Emulsion.
Chinese Journal of Chemical Engineering, 18(2), 202–206.
[https://doi.org/10.1016/s1004-9541\(08\)60343-2](https://doi.org/10.1016/s1004-9541(08)60343-2)
36. Li, Z., Zhong, D.-L., Lu, Y.-Y., & Yan, J. (2017). Efficient CO₂ Capture from a
Simulated Shale Gas using Tetra-*n*-butylphosphonium Bromide Semiclathrate
Hydrate. *Energy Procedia*, 105, 4904–4908.
<https://doi.org/10.1016/j.egypro.2017.03.977>
37. Li, Z., Zhong, D.-L., Lu, Y.-Y., Yan, J., & Zou, Z.-L. (2017). Preferential
enclathration of CO₂ into tetra-*n*-butyl phosphonium bromide semiclathrate hydrate
in moderate operating conditions: Application for CO₂ capture from shale gas.
Applied Energy, 199, 370–381. <https://doi.org/10.1016/j.apenergy.2017.05.043>
38. Li, Z., Zhong, D.-L., Zheng, W.-Y., Yan, J., Lu, Y.-Y., & Yi, D.-T. (2019).
Morphology and kinetic investigation of TBAB/TBPB semiclathrate hydrates formed
with a CO₂ + CH₄ gas mixture. *Journal of Crystal Growth*, 511, 79–88.
<https://doi.org/10.1016/j.jcrysgro.2019.01.042>
39. Lin, W., Delahaye, A., & Fournaison, L. (2008). Phase equilibrium and dissociation
enthalpy for semi-clathrate hydrate of CO₂+TBAB. *Fluid Phase Equilibria*, 264(1-2),
220–227. <https://doi.org/10.1016/j.fluid.2007.11.020>
40. Liudmila Ostroumova Prokhorenkova, Gleb Gusev, Aleksandr Vorobev, Anna
Veronika Dorogush, & Andrey Gulin. (2017). CatBoost: unbiased boosting with

- categorical features. *ArXiv (Cornell University)*.
<https://doi.org/10.48550/arxiv.1706.09516>
41. Lv, Q., Li, L., Li, X., & Chen, Z. (2016). Clathrate hydrate dissociation conditions and structure of the methane + cyclopentane + trimethylene sulfide hydrate in NaCl aqueous solution. *Fluid Phase Equilibria*, 425, 305–311.
<https://doi.org/10.1016/j.fluid.2016.06.020>
 42. Lv, Q.-N., Li, X.-S., Chen, Z.-Y., & Feng, J.-C. (2013). Phase Equilibrium and Dissociation Enthalpies for Hydrates of Various Water-Insoluble Organic Promoters with Methane. *Journal of Chemical & Engineering Data*, 58(11), 3249–3253.
<https://doi.org/10.1021/jc4007025>
 43. Majid, A. A. A., Worley, J., & Koh, C. A. (2021). Thermodynamic and Kinetic Promoters for Gas Hydrate Technological Applications. *Energy & Fuels*, 35(23), 19288–19301. <https://doi.org/10.1021/acs.energyfuels.1c02786>
 44. Mardomi, D., Ardjmand, M., Pahlavanzadeh, H., & Nosratinia, F. (2025). Ethane Hydrate Phase Equilibria in WATER, MEG, and TMAB Semiclathrate System. *Journal of Chemical & Engineering Data*, 70(9), 3851–3859.
<https://doi.org/10.1021/acs.jced.5c00180>
 45. Matsumoto, Y., Makino, T., Sugahara, T., & Ohgaki, K. (2014). Phase equilibrium relations for binary mixed hydrate systems composed of carbon dioxide and cyclopentane derivatives. *Fluid Phase Equilibria*, 362, 379–382.
<https://doi.org/10.1016/j.fluid.2013.10.057>
 46. Mech, D., & Sangwai, J. S. (2014). Phase Stability of Hydrates of Methane in Tetrahydrofuran Aqueous Solution and the Effect of Salt. *Journal of Chemical & Engineering Data*, 59(11), 3932–3937. <https://doi.org/10.1021/jc500841b>

47. Mesbah, M., Soroush, E., Roham, M., & Shahsavari, S. (2017). Phase equilibrium modeling of semi-clathrate hydrates of the CO₂+H₂/CH₄/N₂+TBAB aqueous solution system. *Petroleum Science and Technology*, 35(15), 1588–1594.
<https://doi.org/10.1080/10916466.2017.1322977>
48. Mohammadi, A. H., & Richon, D. (2009a). Phase equilibria of clathrate hydrates of methyl cyclopentane, methyl cyclohexane, cyclopentane or cyclohexane+carbon dioxide. *Chemical Engineering Science*, 64(24), 5319–5322.
<https://doi.org/10.1016/j.ces.2009.09.048>
49. Mohammadi, A. H., & Richon, D. (2009b). Phase Equilibria of Clathrate Hydrates of Tetrahydrofuran + Hydrogen Sulfide and Tetrahydrofuran + Methane. *Industrial & Engineering Chemistry Research*, 48(16), 7838–7841.
<https://doi.org/10.1021/ie900774v>
50. Mohammadi, A. H., & Richon, D. (2011). Phase equilibria of binary clathrate hydrates of nitrogen+cyclopentane/cyclohexane/methyl cyclohexane and ethane+cyclopentane/cyclohexane/methyl cyclohexane. *Chemical Engineering Science*, 66(20), 4936–4940. <https://doi.org/10.1016/j.ces.2011.06.014>
51. Mooijer-van den Heuvel, M. M., Peters, C. J., & de Swaan Arons, J. (2000). Influence of water-insoluble organic components on the gas hydrate equilibrium conditions of methane. *Fluid Phase Equilibria*, 172(1), 73–91. [https://doi.org/10.1016/s0378-3812\(00\)00367-8](https://doi.org/10.1016/s0378-3812(00)00367-8)
52. Muromachi, S., Takeya, S., Suzuki, K., & Tenma, N. (2025). Phase Equilibrium Data and Phase Behavior in Formation Systems of Semiclathrate Hydrates Formed with Partly Asymmetric Quaternary Ammonium Salts under Methane, Carbon Dioxide, and Nitrogen Gas Pressure. *Journal of Chemical & Engineering Data*, 70(7), 2787–2796. <https://doi.org/10.1021/acs.jced.5c00019>

53. Nakamura, T., Makino, T., Sugahara, T., & Ohgaki, K. (2003). Stability boundaries of gas hydrates helped by methane—structure-H hydrates of methylcyclohexane and cis-1,2-dimethylcyclohexane. *Chemical Engineering Science*, 58(2), 269–273.
[https://doi.org/10.1016/s0009-2509\(02\)00518-3](https://doi.org/10.1016/s0009-2509(02)00518-3)
54. Ohmura, R., Uchida, T., Takeya, S., Nagao, J., Minagawa, H., Ebinuma, T., & Narita, H. (2003). Clathrate hydrate formation in (methane + water + methylcyclohexanone) systems: the first phase equilibrium data. *The Journal of Chemical Thermodynamics*, 35(12), 2045–2054. <https://doi.org/10.1016/j.jct.2003.08.010>
55. Park, S., Lee, S., Lee, Y., Lee, Y., & Seo, Y. (2013). Hydrate-based pre-combustion capture of carbon dioxide in the presence of a thermodynamic promoter and porous silica gels. *International Journal of Greenhouse Gas Control*, 14, 193–199.
<https://doi.org/10.1016/j.ijggc.2013.01.026>
56. 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>
57. Seo, D., Lee, S., Moon, S., Lee, Y., & Park, Y. (2023). Investigating two synthetic routes for gas hydrate formation to control the trapping of methane from natural gas. *Chemical Engineering Journal*, 467, 143512.
<https://doi.org/10.1016/j.cej.2023.143512>
58. Seo, Y., Kang, S.-P., Lee, S., & Lee, H. (2008). Experimental Measurements of Hydrate Phase Equilibria for Carbon Dioxide in the Presence of THF, Propylene Oxide, and 1,4-Dioxane. *Journal of Chemical & Engineering Data*, 53(12), 2833–2837. <https://doi.org/10.1021/je800566y>
59. Seo, Y.-T. ., Kang, S.-P. ., & Lee, H. (2001). Experimental determination and thermodynamic modeling of methane and nitrogen hydrates in the presence of THF,

- propylene oxide, 1,4-dioxane and acetone. *Fluid Phase Equilibria*, 189(1-2), 99–110.
[https://doi.org/10.1016/s0378-3812\(01\)00580-5](https://doi.org/10.1016/s0378-3812(01)00580-5)
60. Shi, L., & Liang, D. (2019). Semiclathrate hydrate phase behaviour and structure for CH₄ in the presence of tetrabutylammonium fluoride (TBAF). *The Journal of Chemical Thermodynamics*, 135, 252–259. <https://doi.org/10.1016/j.jct.2019.04.004>
 61. 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>
 62. Strobel, T. A., Koh, C. A., & Sloan, E. D. (2009). Thermodynamic predictions of various tetrahydrofuran and hydrogen clathrate hydrates. *Fluid Phase Equilibria*, 280(1-2), 61–67. <https://doi.org/10.1016/j.fluid.2009.02.012>
 63. Sun, C.-Y., Chen, G.-J., & Zhang, L.-W. (2010). Hydrate phase equilibrium and structure for (methane + ethane + tetrahydrofuran + water) system. *The Journal of Chemical Thermodynamics*, 42(9), 1173–1179.
<https://doi.org/10.1016/j.jct.2010.04.021>
 64. Sun, Z.-G., Fan, S.-S., Guo, K.-H., Shi, L., Guo, Y.-K., & Wang, R.-Z. (2002). Gas Hydrate Phase Equilibrium Data of Cyclohexane and Cyclopentane. *Journal of Chemical & Engineering Data*, 47(2), 313–315. <https://doi.org/10.1021/jc0102199>
 65. Sun, Z.-G., Fan, S.-S., Guo, K.-H., Shi, L., & Wang, R.-Z. (2002). Equilibrium hydrate formation conditions for methylcyclohexane with methane and a ternary gas mixture. *Fluid Phase Equilibria*, 198(2), 293–298. [https://doi.org/10.1016/s0378-3812\(01\)00806-8](https://doi.org/10.1016/s0378-3812(01)00806-8)
 66. Sun, Z.-G., & Sun, L. (2010). Equilibrium Conditions of Semi-Clathrate Hydrate Dissociation for Methane + Tetra-*n*-butyl Ammonium Bromide. *Journal of Chemical & Engineering Data*, 55(9), 3538–3541. <https://doi.org/10.1021/jc100183s>

67. Takeya, S., & Ohmura, R. (2006). Phase Equilibrium for Structure II Hydrates Formed with Krypton Co-existing with Cyclopentane, Cyclopentene, or Tetrahydropyran. *Journal of Chemical & Engineering Data*, 51(5), 1880–1883.
<https://doi.org/10.1021/jc060233r>
68. Takeya, S., & Ohmura, R. (2007). Phase Equilibrium for Structure I and Structure H Hydrates Formed with Methylfluoride and Methylcyclohexane. *Journal of Chemical & Engineering Data*, 52(2), 635–638. <https://doi.org/10.1021/jc6005554>
69. Takeya, S., Yasuda, K., & Ohmura, R. (2007). Phase Equilibrium for Structure II Hydrates Formed with Methylfluoride Coexisting with Cyclopentane, Fluorocyclopentane, Cyclopentene, or Tetrahydropyran. *Journal of Chemical & Engineering Data*, 53(2), 531–534. <https://doi.org/10.1021/jc700624q>
70. Talaei Khoei, T., & Kaabouch, N. (2023). Machine Learning: Models, Challenges, and Research Directions. *Future Internet*, 15(10), 332.
<https://doi.org/10.3390/fi15100332>
71. Torré, J.-P. ., Haillot, D., Rigal, S., de Souza Lima, R., C., D., & J.-P., B. (2015). 1,3 Dioxolane versus tetrahydrofuran as promoters for CO₂-hydrate formation: Thermodynamics properties, and kinetics in presence of sodium dodecyl sulfate. *Chemical Engineering Science*, 126, 688–697.
<https://doi.org/10.1016/j.ces.2015.01.018>
72. Veluswamy, H. P., Bhattacharjee, G., Liao, J., & Linga, P. (2020). Macroscopic Kinetic Investigations on Mixed Natural Gas Hydrate Formation for Gas Storage Application. *Energy & Fuels*, 34(12), 15257–15269.
<https://doi.org/10.1021/acs.energyfuels.0c01862>
73. Wang, M., Sun, Z.-G., Li, C.-H., Zhang, A.-J., Li, J., Li, C.-M., & Huang, H.-F. (2016). Equilibrium Hydrate Dissociation Conditions of CO₂ + HCFC141b or

- Cyclopentane. *Journal of Chemical & Engineering Data*, 61(9), 3250–3253.
<https://doi.org/10.1021/acs.jced.6b00333>
74. Wang, M., Sun, Z.-G., Qiu, X.-H., Zhu, M.-G., Li, C.-H., Zhang, A.-J., Li, J., Li, C.-M., & Huang, H.-F. (2017). Hydrate Dissociation Equilibrium Conditions for Carbon Dioxide + Tetrahydrofuran. *Journal of Chemical & Engineering Data*, 62(2), 812–815. <https://doi.org/10.1021/acs.jced.6b00848>
75. Wang, Y., Chen, S., Li, X., Lang, X., Li, G., & Fan, S. (2025). Hydrate phase equilibrium of hydrogen with THP, DCM, TBAB+THF and sulfur hexafluoride +TBAB aqueous solution systems. *Fluid Phase Equilibria*, 590, 114282.
<https://doi.org/10.1016/j.fluid.2024.114282>
76. Wei, Y., Worley, J., Zerpa, L. E., Chien, Y.-C. (Alice), Dunn-Rankin, D., Kezirian, M. T., & Koh, C. A. (2025). Natural gas storage in hydrates in the presence of thermodynamic hydrate promoters: Review and experimental investigation. *Fluid Phase Equilibria*, 591, 114286. <https://doi.org/10.1016/j.fluid.2024.114286>
77. Wu, W., Hao, B., Guo, Y., Yang, J., Du, M., Zheng, Q., & Bai, Z. (2023). Application of monocyclic compounds as natural gas hydrate promoters: A review. *Chemical Engineering Research and Design*, 190, 66–90.
<https://doi.org/10.1016/j.cherd.2022.11.039>
78. Xu, H., Xing, X., Chen, S., Mao, N., & He, T. (2025). Enhanced CO₂ capture from flue gas using TBAF semi-clathrate hydrates: A comprehensive thermodynamic and kinetic study. *Chemical Engineering Journal*, 508, 160955.
<https://doi.org/10.1016/j.cej.2025.160955>
79. Yang, H., Fan, S., Lang, X., & Wang, Y. (2011). Phase Equilibria of Mixed Gas Hydrates of Oxygen + Tetrahydrofuran, Nitrogen + Tetrahydrofuran, and Air +

- Tetrahydrofuran. *Journal of Chemical & Engineering Data*, 56(11), 4152–4156.
<https://doi.org/10.1021/jc200676k>
80. Yang, H., Fan, S., Lang, X., Wang, Y., & Sun, X. (2012a). Hydrate Dissociation Conditions for Mixtures of Air + Tetrahydrofuran, Air + Cyclopentane, and Air + Tetra-*n*-butyl Ammonium Bromide. *Journal of Chemical & Engineering Data*, 57(4), 1226–1230. <https://doi.org/10.1021/jc201330j>
81. Yang, H., Fan, S., Lang, X., Wang, Y., & Sun, X. (2012b). Hydrate Dissociation Conditions for Mixtures of Air + Tetrahydrofuran, Air + Cyclopentane, and Air + Tetra-*n*-butyl Ammonium Bromide. *Journal of Chemical & Engineering Data*, 57(4), 1226–1230. <https://doi.org/10.1021/jc201330j>
82. Yang, M., Jing, W., Wang, P., Jiang, L., & Song, Y. (2015). Effects of an additive mixture (THF+TBAB) on CO₂ hydrate phase equilibrium. *Fluid Phase Equilibria*, 401, 27–33. <https://doi.org/10.1016/j.fluid.2015.05.007>
83. Ye, N., & Zhang, P. (2014). Phase Equilibrium Conditions and Carbon Dioxide Separation Efficiency of Tetra-*n*-butylphosphonium Bromide Hydrate. *Journal of Chemical & Engineering Data*, 59(9), 2920–2926. <https://doi.org/10.1021/jc500630r>
84. Yu, Y.-S., Xu, C.-G., & Li, X.-S. (2018). Crystal morphology-based kinetic study of carbon dioxide-hydrogen-tetra-*n*-butyl ammonium bromide hydrates formation in a static system. *Energy*, 143, 546–553. <https://doi.org/10.1016/j.energy.2017.10.117>
85. Zang, X., & Liang, D. (2017). Phase Equilibrium Data for Semiclathrate Hydrate of Synthesized Binary CO₂/CH₄ Gas Mixture in Tetra-*n*-butylammonium Bromide Aqueous Solution. *Journal of Chemical & Engineering Data*, 62(2), 851–856.
<https://doi.org/10.1021/acs.jced.6b00876>

86. Zhang, J. S., & Lee, J. W. (2008). Equilibrium of Hydrogen + Cyclopentane and Carbon Dioxide + Cyclopentane Binary Hydrates. *Journal of Chemical & Engineering Data*, 54(2), 659–661. <https://doi.org/10.1021/jc800219k>
87. Zhang, S., Yang, K., & Zhang, P. (2025). Experimental Study on the Formation of CO₂-TBAB Hydrate with Low-Concentration Small-Molecule Alcohols. *Industrial & Engineering Chemistry Research*, 64(9), 5057–5067. <https://doi.org/10.1021/acs.iecr.4c03775>
88. Zheng, J., Yang, M., Liu, Y., Wang, D., & Song, Y. (2017). Effects of cyclopentane on CO₂ hydrate formation and dissociation as a co-guest molecule for desalination. *The Journal of Chemical Thermodynamics*, 104, 9–15. <https://doi.org/10.1016/j.jct.2016.09.006>
89. Zhong, D., Ding, K., Yang, C., Bian, Y., & Ji, J. (2012). Phase Equilibria of Clathrate Hydrates Formed with CH₄ + N₂ + O₂ in the Presence of Cyclopentane or Cyclohexane. *Journal of Chemical & Engineering Data*, 57(12), 3751–3755. <https://doi.org/10.1021/jc301024v>
90. Zhong, D.-L., Daraboina, N., & Englezos, P. (2013). Recovery of CH₄ from coal mine model gas mixture (CH₄/N₂) by hydrate crystallization in the presence of cyclopentane. *Fuel*, 106, 425–430. <https://doi.org/10.1016/j.fuel.2013.01.029>
91. Zhong, D.-L., Li, Z., Lu, Y.-Y., & Sun, D.-J. (2014). Phase Equilibrium Data of Gas Hydrates Formed from a CO₂ + CH₄ Gas Mixture in the Presence of Tetrahydrofuran. *Journal of Chemical & Engineering Data*, 59(12), 4110–4117. <https://doi.org/10.1021/jc5007482>
92. Zhou, X., Long, Z., He, Y., Shen, X., & Liang, D. (2018). Phase Equilibria and the Crystallographic Properties of TBAB–CO₂ Semiclathrate Hydrates. *Journal of*

Chemical & Engineering Data, 63(5), 1249–1255.

<https://doi.org/10.1021/acs.jced.7b00884>