Department of Chemical Engineering,
King monkut's Institute of Technology Ladkrabang

Fingerprint-Based Machine Learning Prediction of Chemical Properties

Presented by

Sukit Kerdsawat 63010987

Ekkaparb Parisupat 63011091

Advisor

Asst. Prof. Amata Anantpinijwatna, Ph.D.

Table of contents

Introduction

Background

Example of Chemical properties that being used in industry include Solubility of solvents for Azeotrope Distillation, Activity Coefficient for Flammability, C_p for Sizing Heat Exchanger Equipment

The traditional way to determine the properties of a substance is through experimentation. This can be expensive and time-consuming, as it requires chemicals, equipment, and labor.

↓

Properties Prediction Model
(QSPR Model)

Figure 1. Chemical Experiment

Ref [1] : Industrial Needs in Physical Properties, https://doi.org/10.1021/ie030170v
Ref [2] : Thermodynamics in process development in the chemical industry, https://doi.org/10.1016/0378-3812(91)85029-T

Introduction

QSPR & Machine Learning

QSPR (Quantitative structure–property relationship) establishes relations between a molecule's structure and its properties using mathematical or statistical methods. The QSPR model is used to predict the properties of molecules from molecular structures.

Traditional QSPR Model
Group Contribution

Similar Group Structure

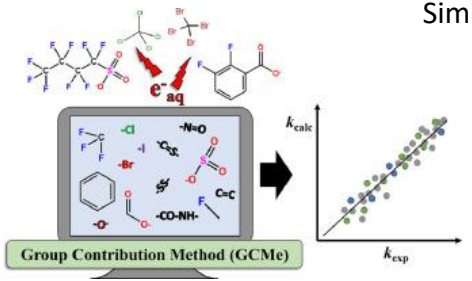


Figure 2. Group Contribution

Modern QSPR Model
Machine Learning

perform complicated calculations

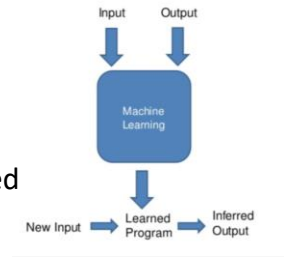
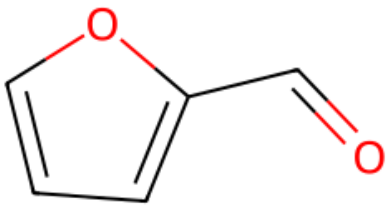


Figure 3. Machine Learning

Ref [3] : A review on machine learning algorithms for the ionic liquid chemical space, https://doi.org/10.1039/D1SC01000J
Ref [4] : New Group Contribution Method for Estimating Properties of Pure Compounds , https://doi.org/10.1002/aic.690401011

Introduction

Review Previous Work



T_b = ……. K

Molecular Structure to Property

Ref [5] : Novel method for properties prediction of pure organic compounds using machine learning https://doi.org/10.1016/B978-0-323-88506-5.50068-1

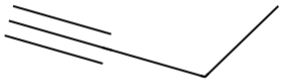
Introduction

Review SMILES

The simplified molecular-input line-entry system (SMILES) is a specification in the form of a line notation for describing the structure of chemical species using short ASCII strings.

SMILES	CCCC

Figure 4. Butane

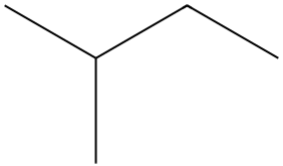


SMILES	C#CCC

Figure 6. Butyne

SMILES	C=CCC

Figure 5. Butene



SMILES	C(C)CCC

Figure 7. Isopentane

Introduction

Review SMILES

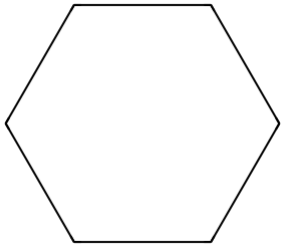


Figure 8. Cyclic

SMILES	C1CCCCC1

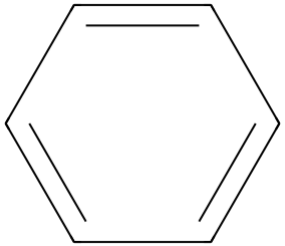


Figure 9. Aromatic

SMILES	c1ccccc1

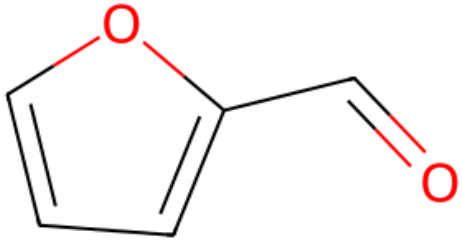
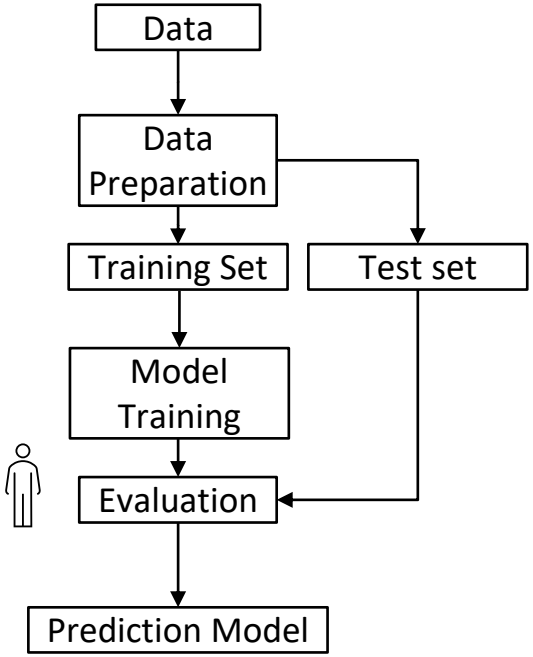


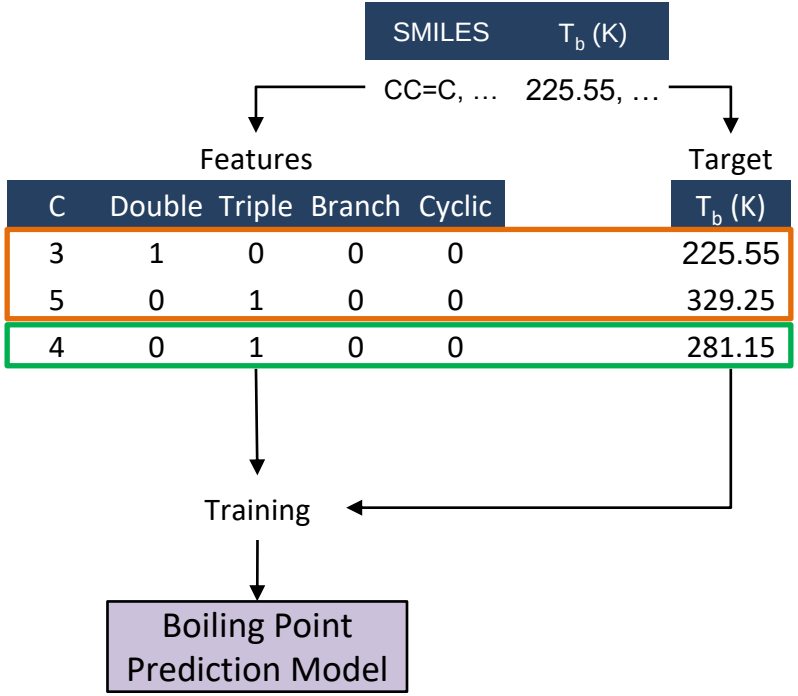
Figure 10. Furfural

SMILES	c1cc(C=O)oc1

Introduction

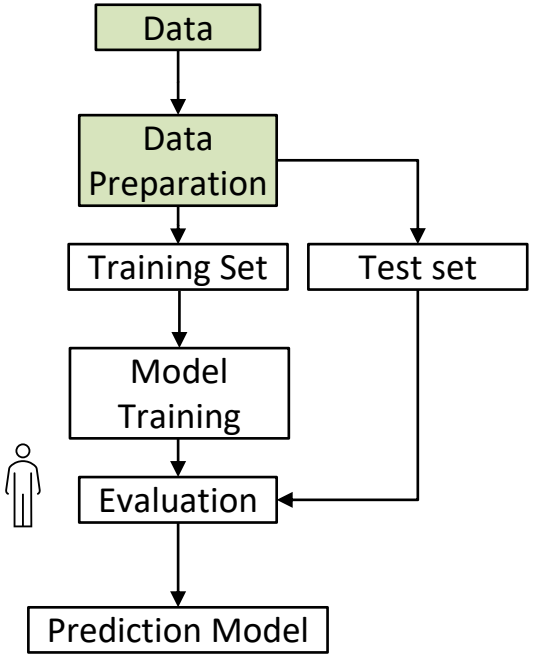
Review (Machine Learning Flowchart)

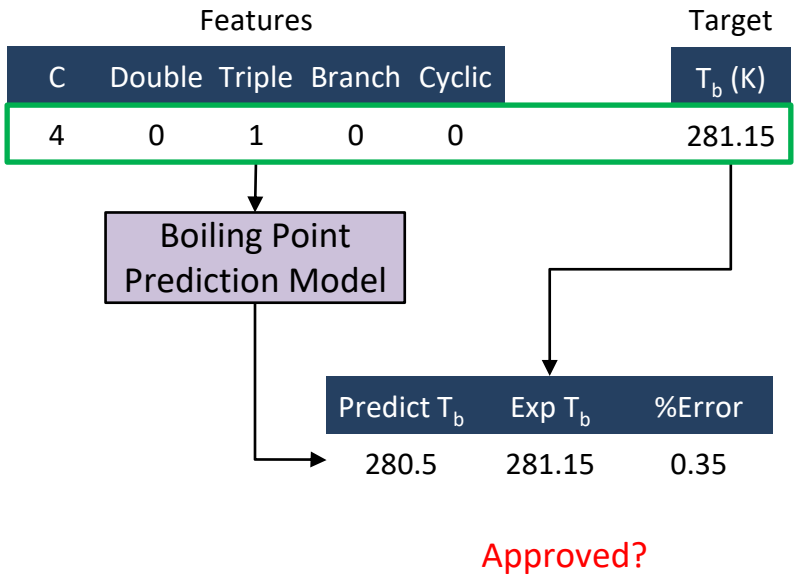




Introduction

Review (Machine Learning Flowchart)





Introduction

Review (Problem)

Table 1. Show same feature between two similar chemical substance

	Features					Target	
SMILES	C	Double	Triple	Branch	Cyclic	T_b (exp)	T_b (predict)
C1CCC=CCC1	7	1	0	0	1	388.15	375.85
CC1=CCCCC1	7	1	0	0	1	383.45	375.85
CC1CCC=CC1	7	1	0	0	1	375.85	375.85
CC1CCCC=C1	7	1	0	0	1	376.15	375.85

SMILES are different but same features

T_b predict is same value

Introduction

Objective & Scope

Objective

1. To study the converting of SMILE structures into molecular fingerprints.
2. To predict the properties of substances from molecular fingerprints using machine learning techniques.
3. To improve the accuracy of predicting the properties of substances using machine learning.

Scope

1. To study properties of pure organic compound that contain C, H, O and N atom and Number of C atom is 1-12 atom
2. To study morgan fingerprint that one of tool in RDKit Library using Python Programming Language via Spyder Program

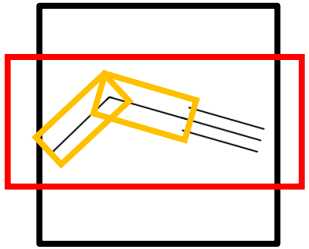
Introduction

Molecular Fingerprint

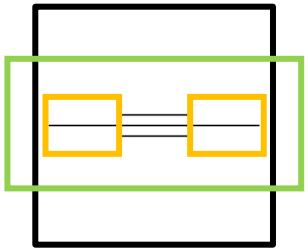
Molecular Fingerprint is a representation of structure in a set of number text which being used to store structure of molecule and each number represent a substructure in molecule.

Table 2. Show different feature between two similar chemical substance

Name	SMILES	Bit 1 CO	Bit 2 CCC#C	Bit 3 CC#CC	Bit 4 CC	Bit 5 CCC
1-Butyne	CCC#C	0	1	0	2	1
2-Butyne	CC#CC	0	0	1	2	0



1-Butyne

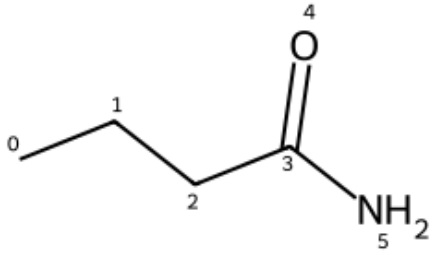


2-Butyne

Introduction

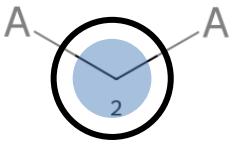
Morgan Fingerprint

Step 1 : Specify indices in molecule

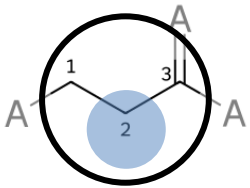


Step 2 : Specify substructure with given radius at each index

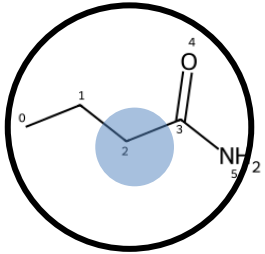
r = 0



r = 1



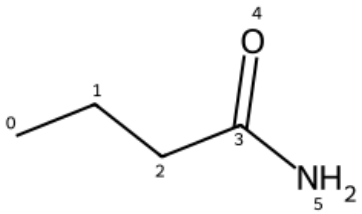
r = 2

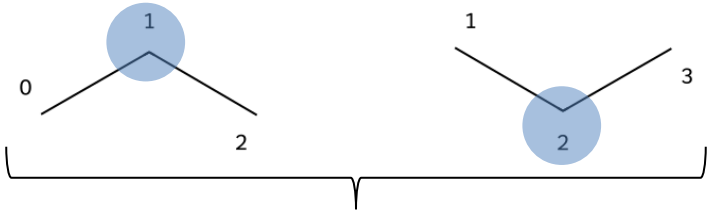


Ref [6] : Getting Started with the RDKit in Python, https://www.rdkit.org/docs/GettingStartedInPython.html#morgan-fingerprints-circular-fingerprints
Ref [7] : Extended-Connectivity Fingerprints (Morgan Algorithm), https://doi.org/10.1021/ci100050t

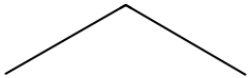
Introduction
Morgan Fingerprint

Step 3 : Remove duplicate structure and store into Bit of fingerprint

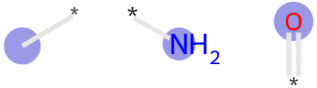


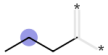


Duplicate



Result





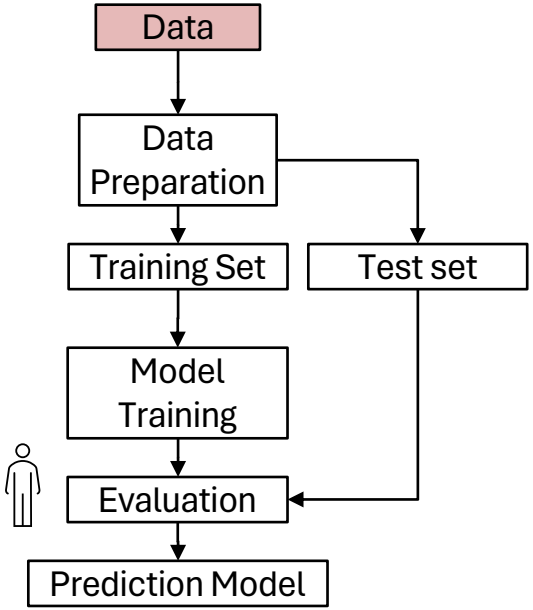
Bit 0	Bit 1	Bit 2	Bit 3	Bit 4	…	Bit 539	…	Bit 791	Bit 792	…	Bit 1022	Bit 1023
4	1	1	0	0	…	1		0	2		0	0

nBit = 1024

Ref [6] : Getting Started with the RDKit in Python, https://www.rdkit.org/docs/GettingStartedInPython.html#morgan-fingerprints-circular-fingerprints
Ref [7] : Extended-Connectivity Fingerprints (Morgan Algorithm), https://doi.org/10.1021/ci100050t

Methodology

Machine Learning Flow



Data Collecting

1. Previous work : 560 datapoint Hydrocarbon C,H
2. Chemical database of Chemical Engineering Design Library (ChEDL)

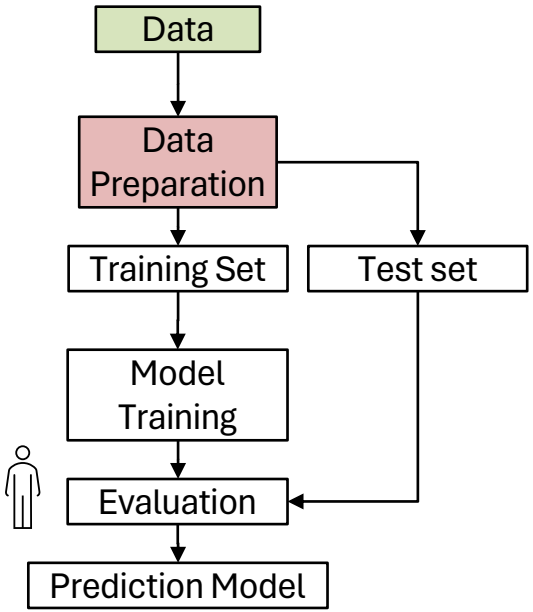
Table 3. ChEDL Database Collection

Database Source	Datapoint
CRC Handbook of Chemistry and Physics	5,542
CAS Common Chemistry	10,419
NIST Webbook	5,847
Wikidata	872
Yaws, “Thermophysical Properties of Chemicals and Hydrocarbons”	13,461
Joback, “Estimation of Pure-Component Properties from Group-Contributions”	23,068

Ref [8] : Chemicals 1.1.5 documentation, https://chemicals.readthedocs.io/chemicals.phase_change.html#boiling-point

Methodology

Machine Learning Flow



Data Preparation – Data Cleaning

Normal Distribution : z-score
Threshold : z = 2 (≈ 95%)

$$z = \frac{x - \mu}{\sigma}$$

Where
x = boiling point
$\bar{\mu}$ = mean of boiling point
σ = standard Deviation of boiling point

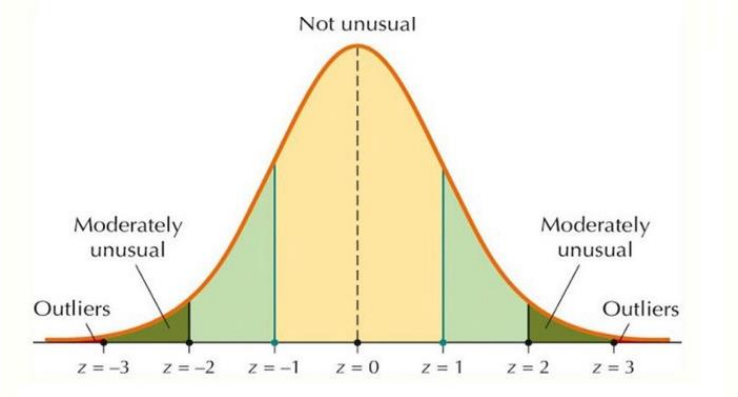
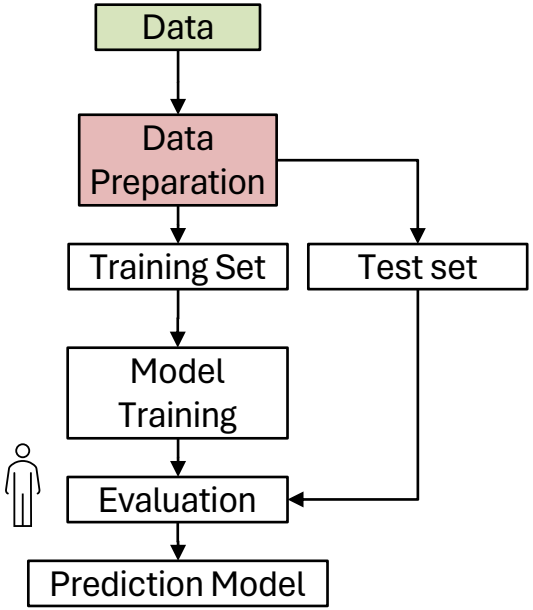


Figure 11. Detecting Outliner with z-scores

Ref [9] : Outlier Detection using PDF and z-score | Kaggle, https://www.kaggle.com/code/faressayah/outlier-detection-using-pdf-and-z-score

Methodology

Machine Learning Flow



Data Preparation – Convert SMILES to Features

Previous Work

SMILES
CCC#C (1-Butyne)

C	Double	Triple	Branch	Cyclic
4	0	1	0	0

This Work

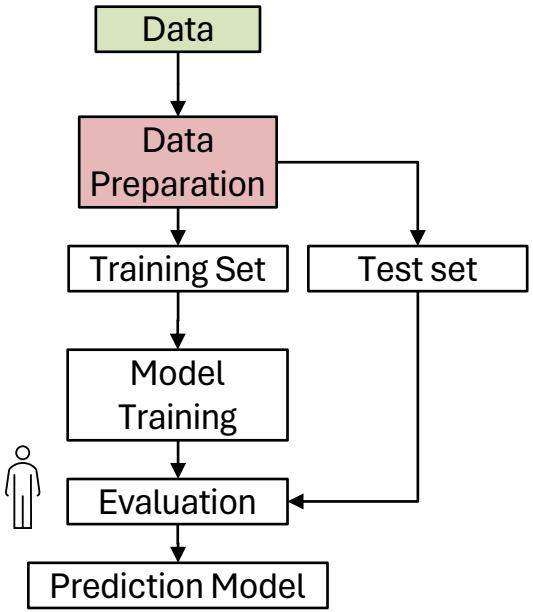
SMILES
CCC#C (1-Butyne)

Bit 0	Bit 1	Bit 2	…	Bit 4096
4	1	1	…	0

Molecular Fingerprint

Methodology

Machine Learning Flow



Data Preparation – Feature Selection (SelectKBest)

SelectKBest is a feature selection technique that is used to select the k best features from a dataset.

Table 4. Feature Score

Features (Bit)	Score
1	1270.96
2	956.88
3998	932.28
3096	931.06
32	930.64
.	.
.	.
3552	4.04

Utilizing SelectKBest, a ranking score is determined between the input (bit) and output (T_b).

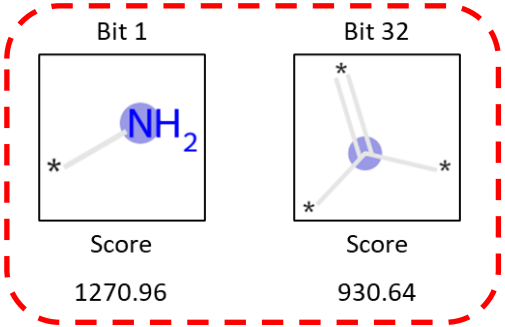
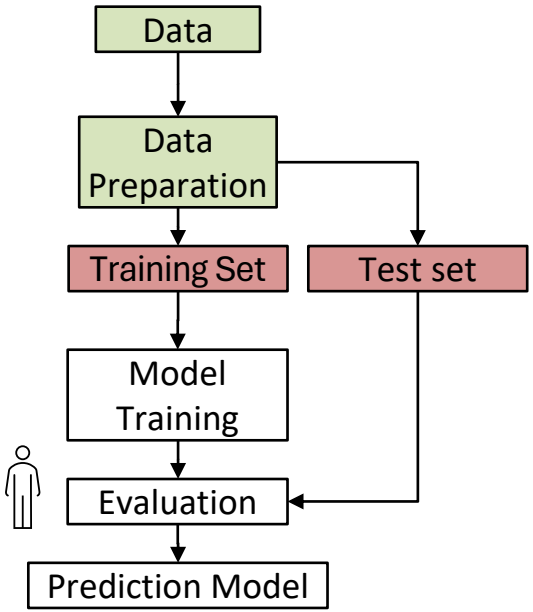


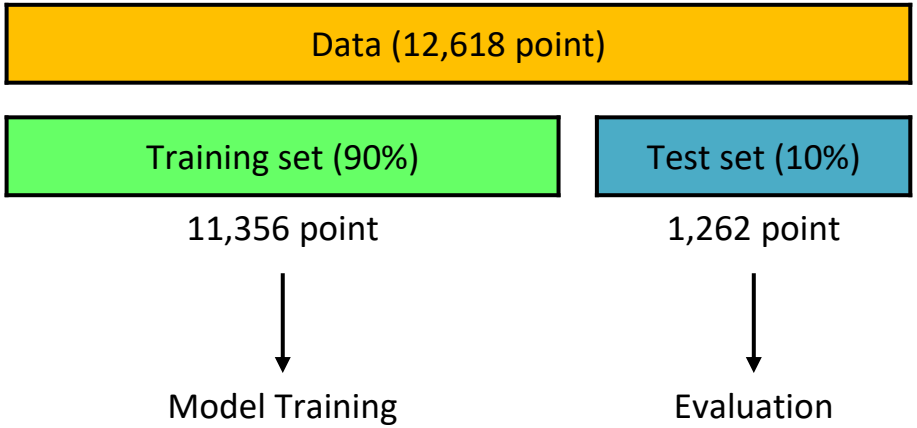
Figure 12. Show substructure of each bit

Methodology

Machine Learning Flow

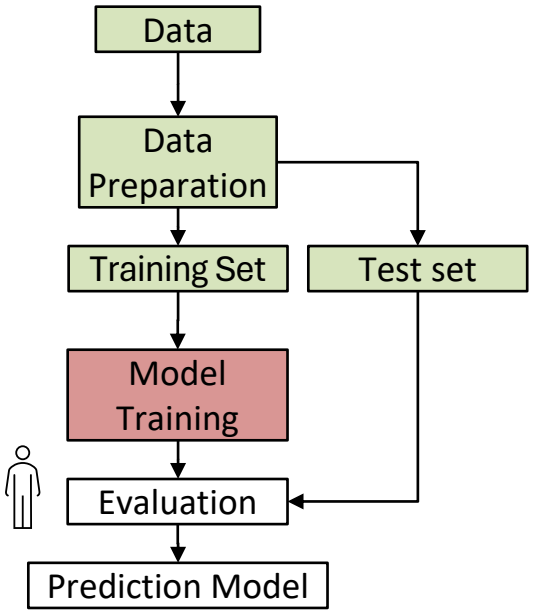


Data Splitting

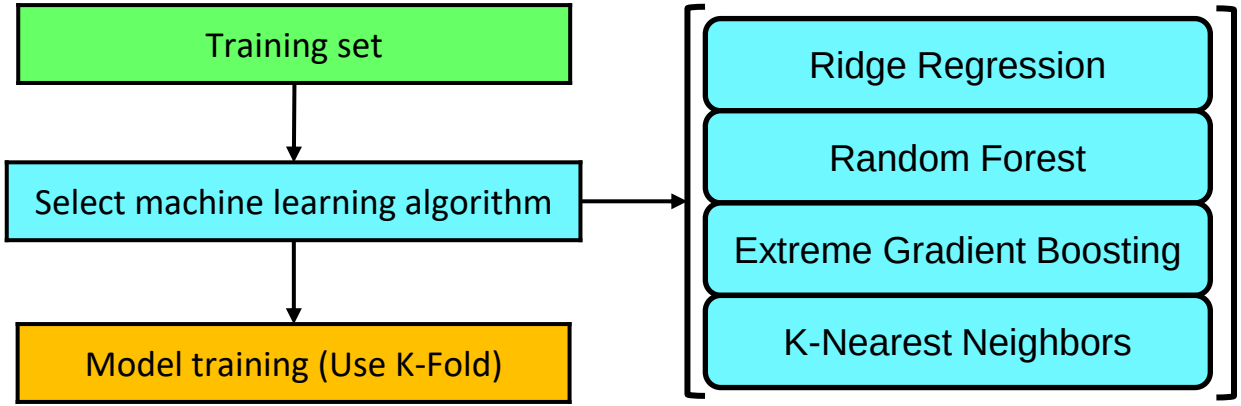


Methodology

Machine Learning Flow

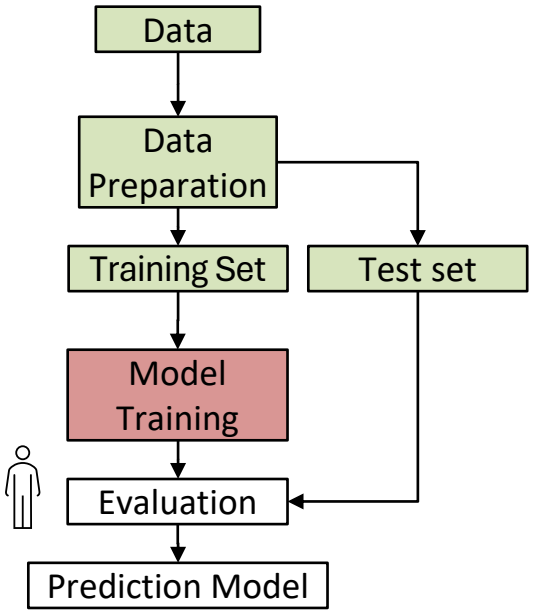


Model Training



Methodology

Machine Learning Flow



Cross Validation: K-Fold

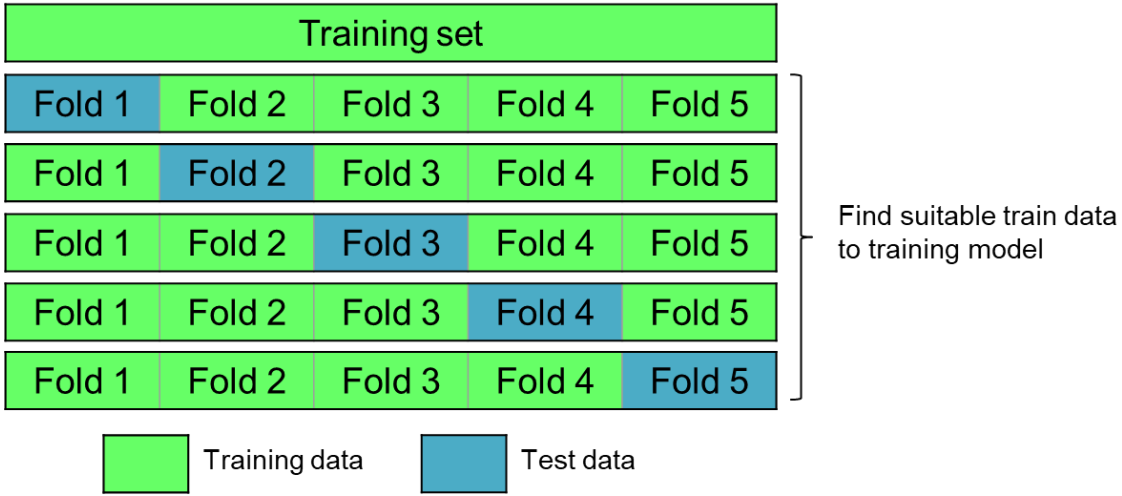
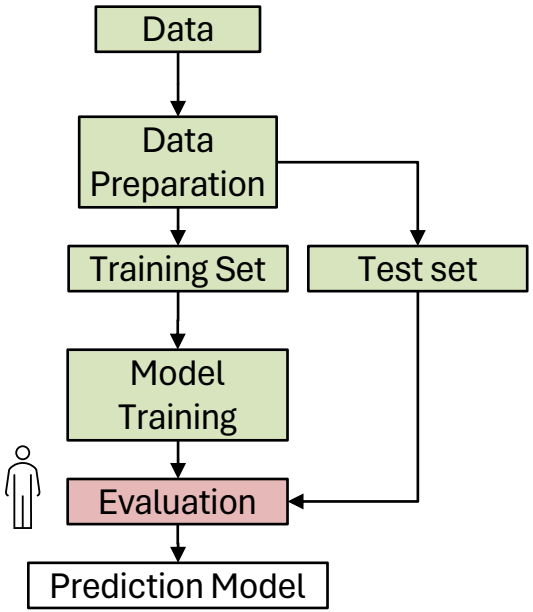


Figure 13. Example of K-Fold

Methodology

Machine Learning Flow



Performance Evaluation

Mean Absolute Error

$$MAE = \frac{1}{N}\sum\left|y_i - \hat{y}_i\right|$$

Mean Absolute Percentage Error

$$MAPE = \frac{1}{N}\sum\left|\frac{y_i - \hat{y}_i}{y_i}\right| \times 100$$

Coefficient of determination

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

Where

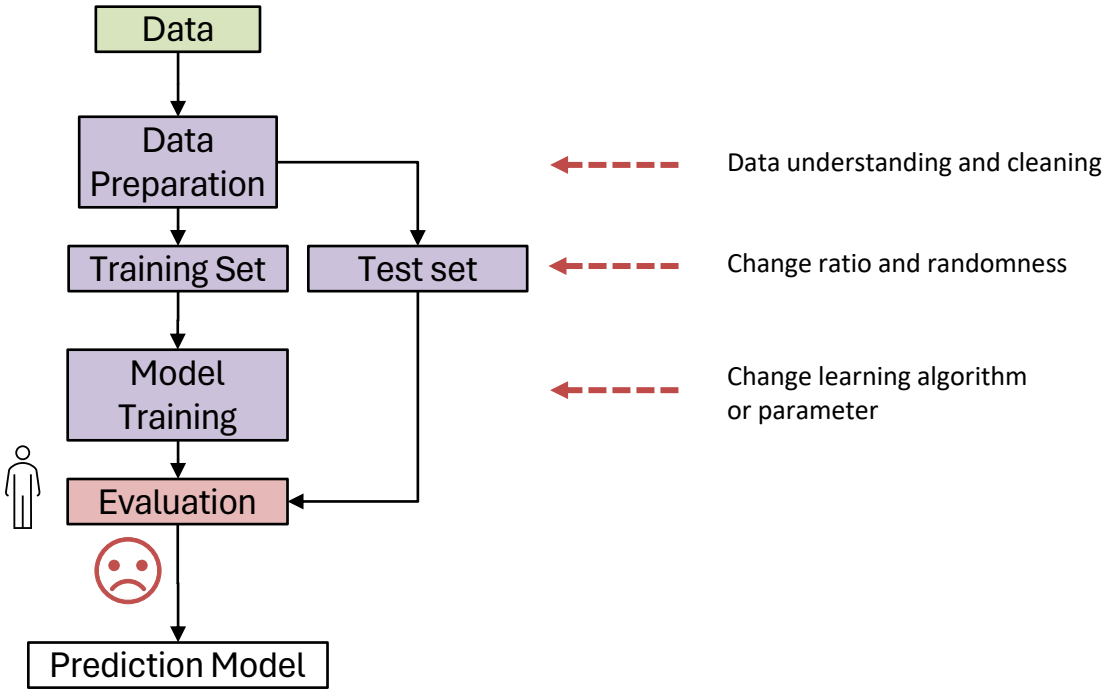
$\hat{y}_i$ = predict value of y

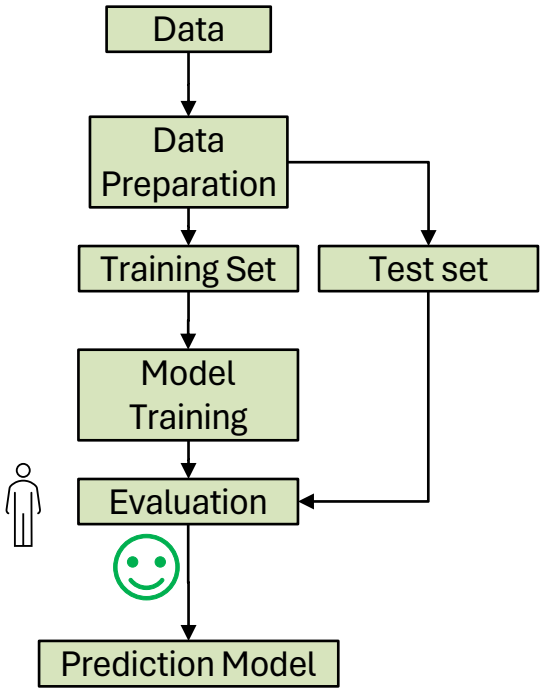
$\bar{y}$ = mean value of y

N = amount of data

Methodology

Machine Learning Flow





Results

Fingerprint with Boiling Point Model

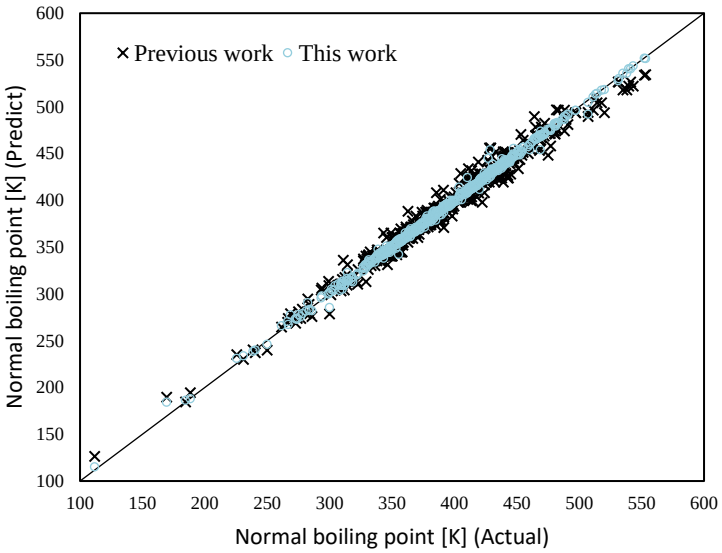


Figure 13. Comparison of the normal boiling point prediction
Scope: Hydrocarbon = C,H

- Previous Work : C, Double, Triple, Bracket, Cyclic
- This work : Count-based Morgan Fingerprint
  : r = 2, nBit = 1024
  : XGB Algorithm with SelectKBest, K-fold

Table 5. Model Performance Comparison

Name	MAE	MAPE (%)	R^2
Previous work	5.862	1.472	0.984
This work	3.378	0.927	0.993

Results

Fingerprint with Boiling Point Model

Previous Work Problem : Similar structure molecules, same features

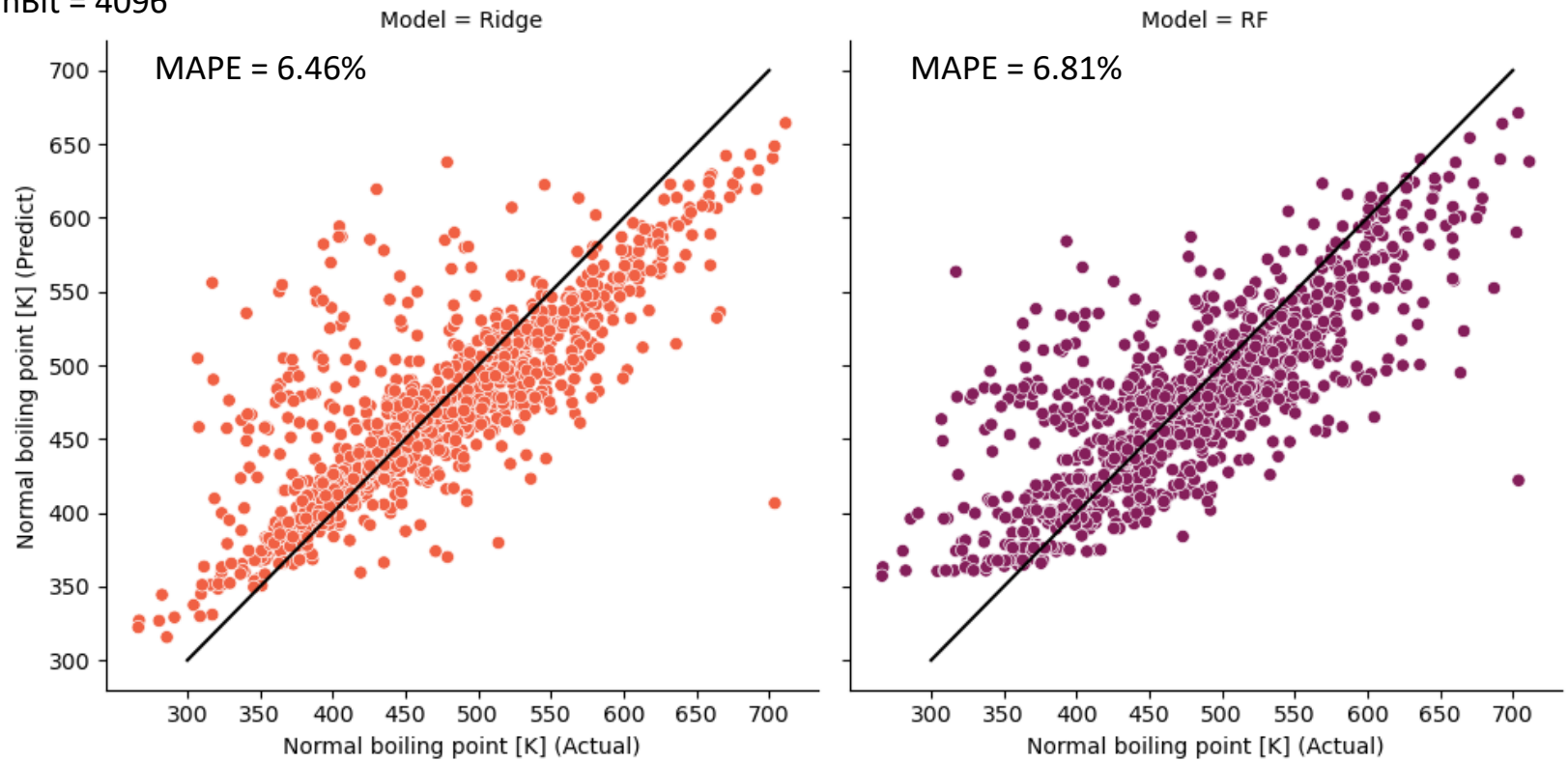
Table 6. Comparison Prediction Performance of Previous and This work

SMILES	Predict T_b, K Previous Work	Predict T_b, K This Work	Actual T_b , K
C1CCC=CCC1	375.85	389.26	388.15
CC1=CCCCC1	375.85	380.76	383.45
CC1CCC=CC1	375.85	373.72	375.85
CC1CCCC=C1	375.85	371.31	376.15

Results

Model with Fingerprint and CHON Scope

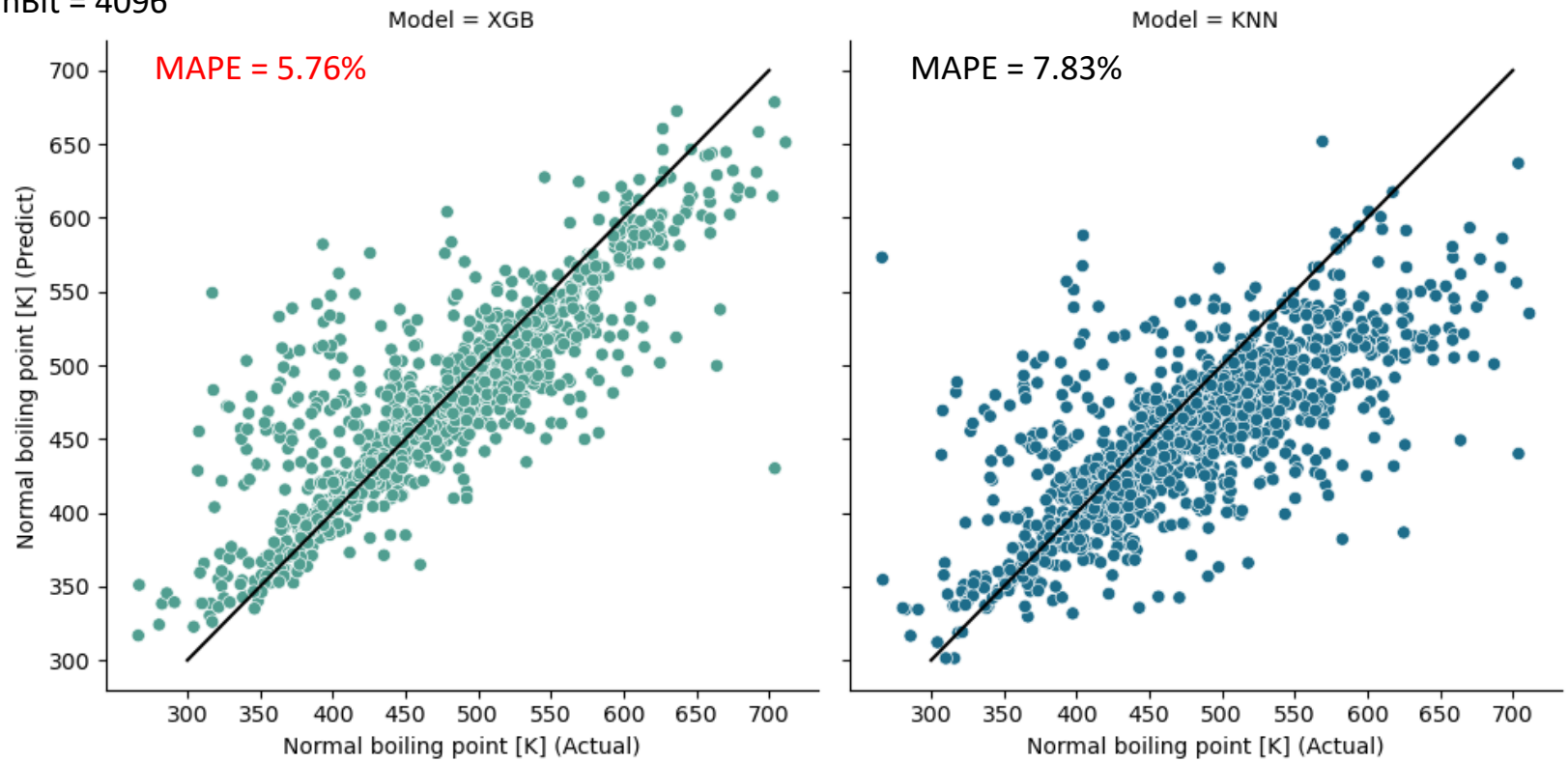
r = 3, nBit = 4096



Results

Model with Fingerprint and CHON Scope

r = 3, nBit = 4096



Results

Model with Fingerprint and CHON Scope

Count-based Morgan Fingerprint r = 3, nBit = 4096 ML Algorithm with SelectKBest, K-fold

Table 7. Model performance analysis results

Algorithm	Ridge		RF		XGB		KNN	
	Train	Test	Train	Test	Train	Test	Train	Test
R^2	0.672	0.666	0.736	0.656	0.830	0.719	0.987	0.478
MAE	28.43	29.38	27.45	30.77	21.01	26.08	1.296	38.03
%MAPE	6.288	6.459	6.121	6.813	4.665	5.755	0.297	7.828

*Note: Ridge = Ridge Regression, RF = Random Forest, XGB = Extreme Gradient Boosting, KNN = K-Nearest Neighbors

Conclusion

Compare with Previous Work

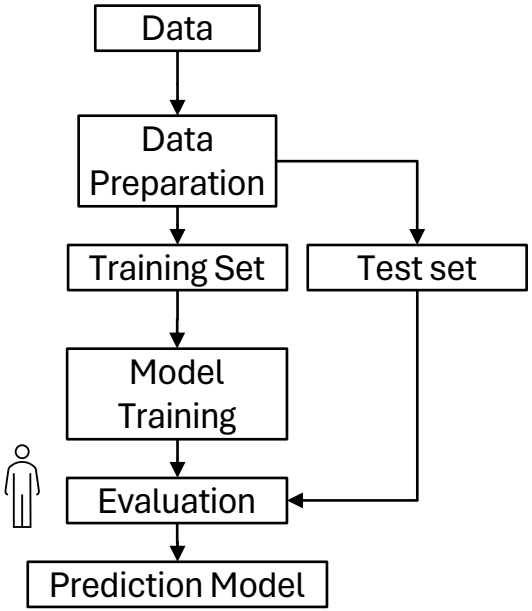


Table 8. Machine learning modeling comparison

Machine learning step	Previous work	This work
Data	560 datapoint Hydrocarbon = C,H	12,618 datapoint Organic compound with C,H,O and N atom
Data Preparation	C, Double, Triple, Brach, Cyclic	Z-score Morgan Fingerprint SelectKBest
Data Splitting	460 : 100	Ratio = 90 : 10 (11,356 : 1,262 point)
Model Training	Developed Regression	K-Fold with XGB Ridge, RF, KNN
Evaluation	MAPE, RMSE, R^2	MAE, MAPE, R^2

*Note: Ridge = Ridge Regression, RF = Random Forest, XGB = Extreme Gradient Boosting, KNN = K-Nearest Neighbors

Conclusion

Project Work

1. Use Morgan Fingerprint to establish Boiling Point Prediction Model
2. Use Morgan Fingerprint to resolve Previous Work Problem “Similar Structures, get same Features”
3. Use Morgan Fingerprint to expand scope of Boiling Point Prediction Model From C,H to C,H,O,N

Conclusion

Time Schedule

Take action Plan

1st Semester	Jul				Aug				Sep				Oct				Nov			
1. Study Previous Work																				
2. Literature Review & Study Fingerprint																				
3. Data Collecting																				
4. Data Preparation																				
5. Modeling, Evaluation and Model Improvement																				
6. Report and Presentation																				
2nd Semester	**Dec**				**Jan**				**Feb**				**Mar**				**Apr**			
1. Literature Review & Study Properties																				
2. Data Collecting																				
3. Data Preparation																				
4. Modeling, Evaluation and Model Improvement																				
5. Report and Presentation																				

Thank You

Q&A