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

Assoc. Prof. Dr. Amata Anantpinijwatna

Table of contents

1 Introduction

2 Methodology

3 Result

4 Conclusion

First Semester

Introduction

Background, Object and Scope

Selection of suitable substances for production processes is a long-standing problem in the petrochemical industry. Development of a model by using machine learning (ML) become more popular because they are faster and more accurate than the traditional methods

Fig 1. Chemical Experiment

Objective

1. To study the converting of SMILE structures into molecular fingerprints.
2. To model and improve the accurate of prediction model on T_b and P^{sat} of substances from molecular fingerprints using machine learning and techniques.

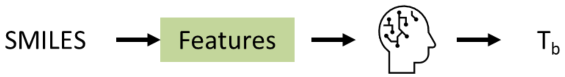
Scope

1. To study T_b and P^{sat} 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

Ref [1] : Thermodynamics in process development in the chemical industry, https://doi.org/10.1016/0378-3812(91)85029-T

Introduction

Review Previous Work



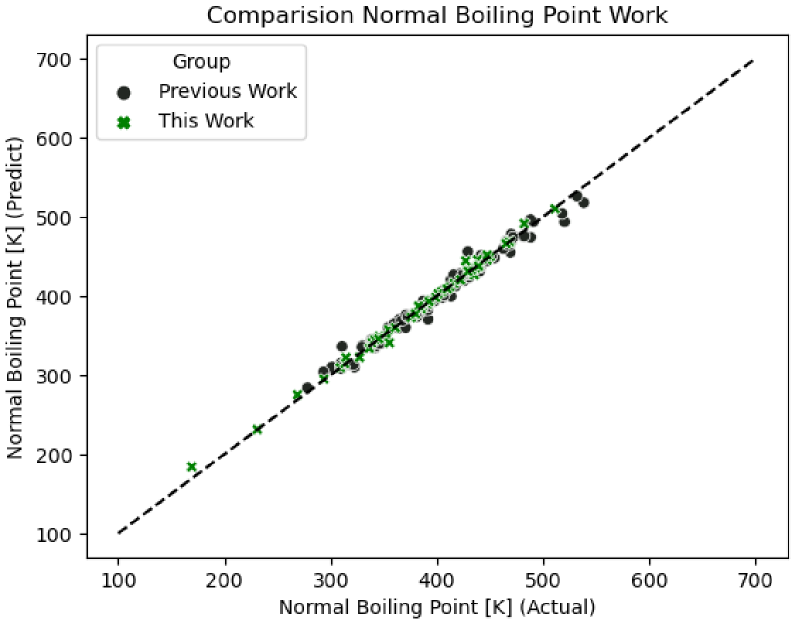


Fig 2. Comparison T_b Work on CH Scope

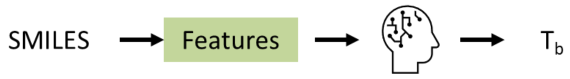
Table 1. Performance of Each Work on T_b Prediction Model within CH Scope

Error Metric	Previous Work [2]	This Work
MAE	5.86	3.20
MAPE(%)	1.47	0.84
RMSE	7.92	4.70
R^2	0.98	0.99

Ref [2] : 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 Previous Work



The Previous work (Nattasinee and colleagues) [2] use Machine Learning (ML) by convert the representation of molecule from text format (SMILES) into numbers in the following table while This work convert SMILES to “Molecular Fingerprint”

Table 2. Compare Prediction Between Each Work of Similar Structure Substance

SMILES	C	Double	Triple	Branch	Cyclic	Actual T_b	Predict T_b Previous Work [2]	Predict T_b This Work
C1CCC=CCC1	7	1	0	0	1	388.15	375.85	387.34
CC1=CCCCC1	7	1	0	0	1	383.45	375.85	377.95
CC1CCC=CC1	7	1	0	0	1	375.85	375.85	370.31

Similar SMILES

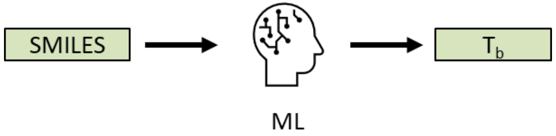
Same Features

Same Predictions

Ref [2] : 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

Result



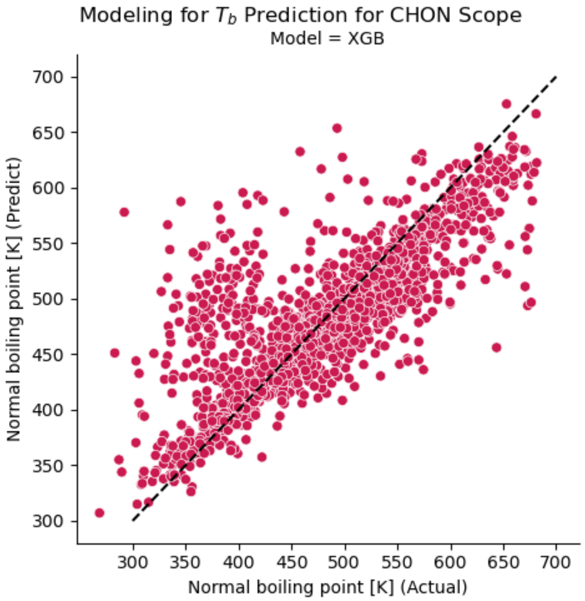


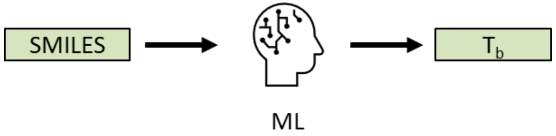
Fig 3. Best T_b model on CHON Scope

Table 3. Performance on T_b Prediction Model within CHON Scope

Dataset	MAE	MAPE (%)	RMSE	R^2
Train	20.42	4.51	30.38	0.83
Test	26.72	5.94	43.40	0.66

Introduction

Result



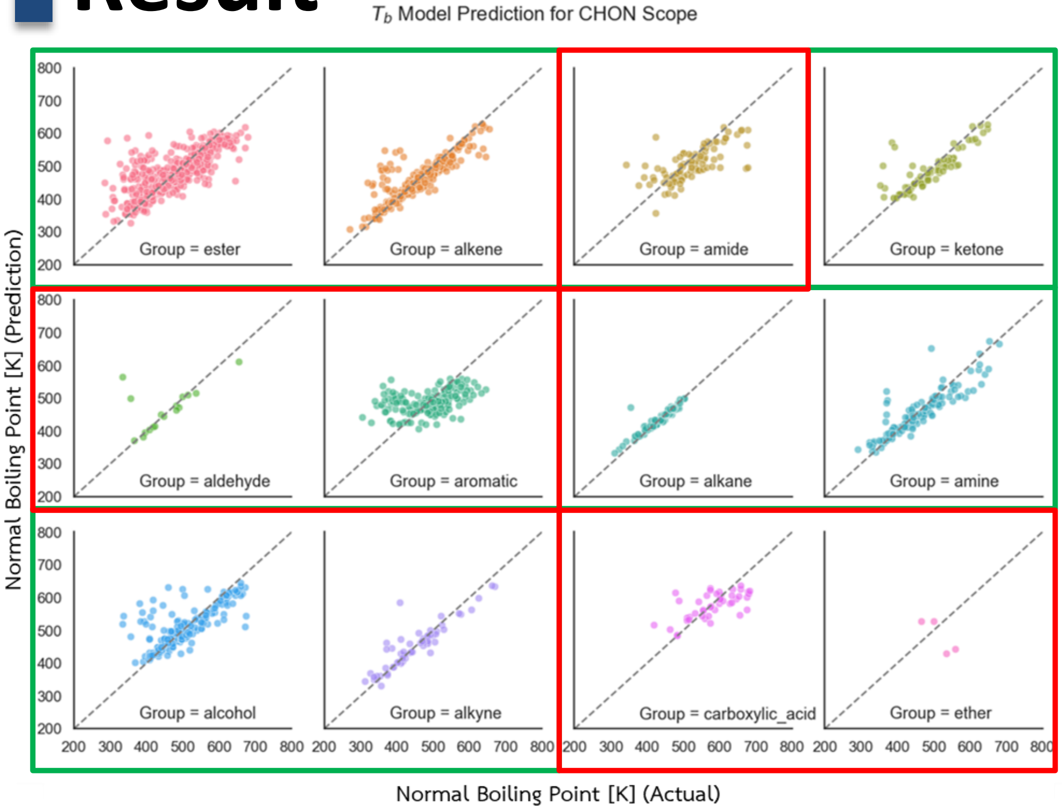


Fig 4. T_b Prediction on CHON Scope Group by Functional Group

The functional group that this T_b prediction model can predict well is that the group that have condition with $R^2 \geq 0.5$ and MAPE $\leq 10\%$

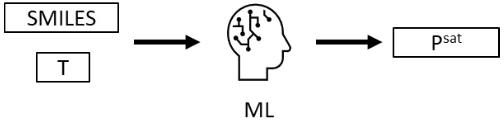
Acceptable: Ester, Alkene, Ketone, Alkane, Amine, Alcohol, Alkyne

Unacceptable: Amide, Aldehyde, Aromatic, Carboxylic Acid, Ether

Second Semester

Methodology

Data Collection

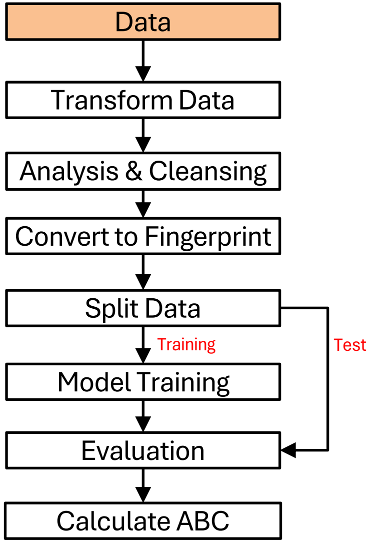


SMILES

T

ML

P^{sat}



Data

Transform Data

Analysis & Cleansing

Convert to Fingerprint

Split Data

Training

Test

Model Training

Evaluation

Calculate ABC

Table 4. ChEDL Database Collection

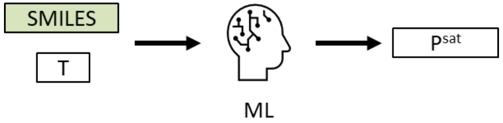
Database Source	Datapoint
Hall, K. R. Vapor Pressure and Antoine Constants for Hydrocarbons, and S, Se, Te, and Halogen Containing Organic Compounds. Springer, 1999.	6,346
Dykyj, J., and K. R. Hall. "Vapor Pressure and Antoine Constants for Oxygen Containing Organic Compounds". 2000.	
Hall, K. R. Vapor Pressure and Antoine Constants for Nitrogen Containing Organic Compounds. Springer, 2001.	

The Total amount of Substance that be included in CHON is 2,246 datapoints.

Ref [3] : Vapor Pressure (chemicals.vapor_pressure) , https://chemicals.readthedocs.io/chemicals.vapor_pressure.html?highlight=vapor%20pressure#rc9de082557a5-4

Methodology

Transform Data



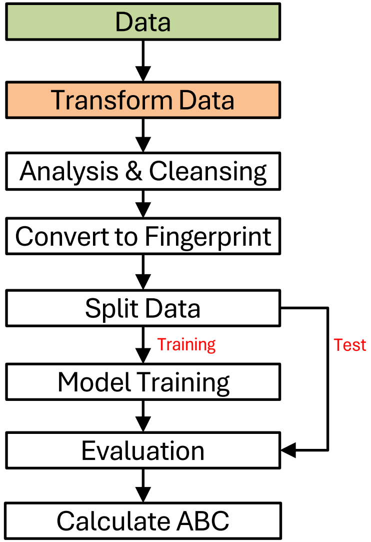


Table 5. P^{sat} Raw data

CAS	Name	A	B	C	Tmin	Tmax
50-00-0	Methanal	21.37	2204.13	-30.15	190	271
51-66-1	p-Methoxy-acetanilide	20.27	3916.19	-177.10	456	533
51-75-2	N-Methylbis(2-chloroethyl)amine	25.61	6563.29	0	273	333
541-35-5	Butyramide	22.08	4164.35	-109.32	320	398

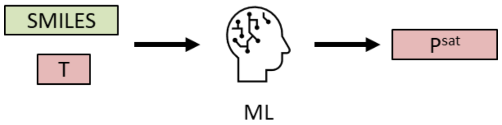
Use PubChem Database to
convert Name of compound to its SMILES

Name	SMILES	A	B	C	Tmin	Tmax
Butyramide	CCCC(=O)N	22.08	4164.35	-109.32	320	398

Ref [4] PubChemPy documentation — PubChemPy 1.0.4 documentation, https://pubchempy.readthedocs.io/en/latest/guide/gettingstarted.html/

Methodology

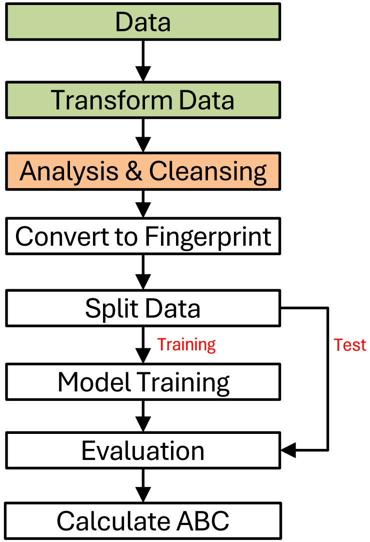
Analysis & Cleansing


SMILES

T

ML

P^{sat}


Data

Transform Data

Analysis & Cleansing

Convert to Fingerprint

Split Data

Training

Test

Model Training

Evaluation

Calculate ABC

Calculate the Vapor Pressure value using the Antoine equation, which is divided into 5 points according to Tmin and Tmax range, which can be shown in the table below.

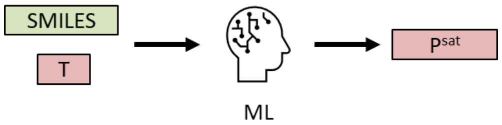
SMILES	A	B	C	T_{min}	T_{max}
CCCC(=O)N	22.08	4164.35	-109.32	320	398

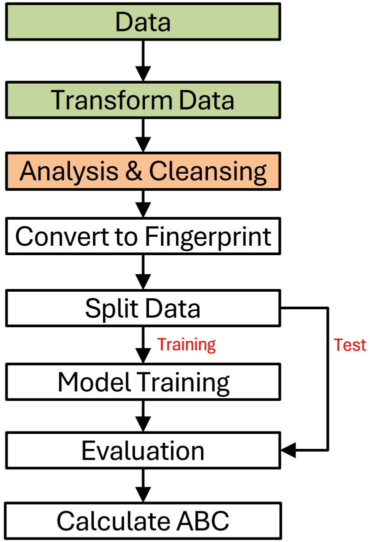
SMILES	VP1	VP2	VP3	VP4	VP5
CCCC(=O)N	2.32	3.99	5.40	6.61	7.66

Note : VP – Vapor Pressure in ln(P^{sat}), P^{sat} in Pa, T in K

Methodology

Analysis & Cleansing





After get the Vapor Pressure value of all substance, use Boxplot to see distribution of Vapor Pressure so that can we detect abnormal Vapor Pressure

Table 6. SMILES and Vapor Pressure Data

SMILES	VP1	VP2	VP3	VP4	VP5
CCCC(=O)N	2.32	3.99	5.40	6.61	7.66
CC(CO)O	9.89	10.36	10.79	11.19	11.57
.	.	.	.	.	.
.	.	.	.	.	.
COC1=CC=C(C=C1)CC=C	4.85	7.15	8.93	10.35	11.52

2,246 datapoint

Use boxplot to detect outliers

Outliers

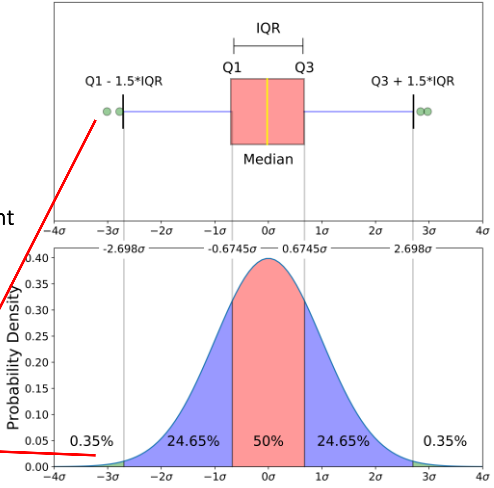
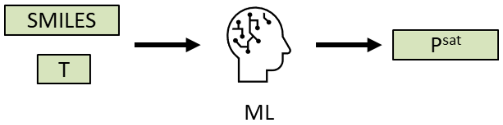
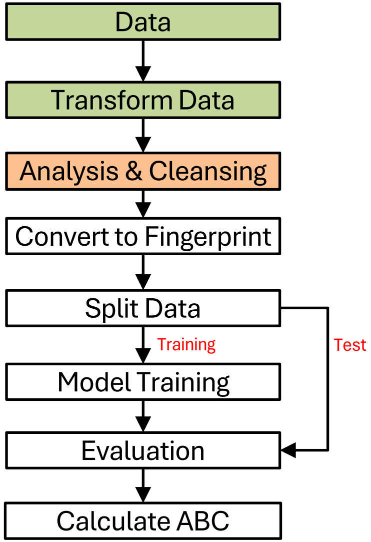


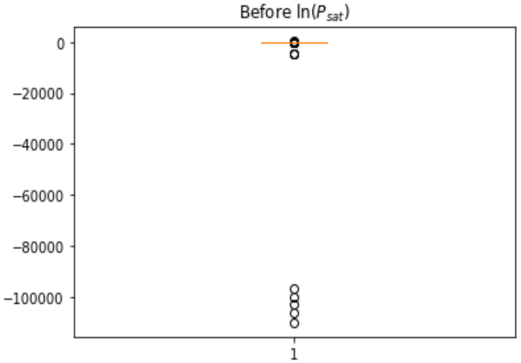
Figure 5. Boxplot and Normal Distribution

Methodology

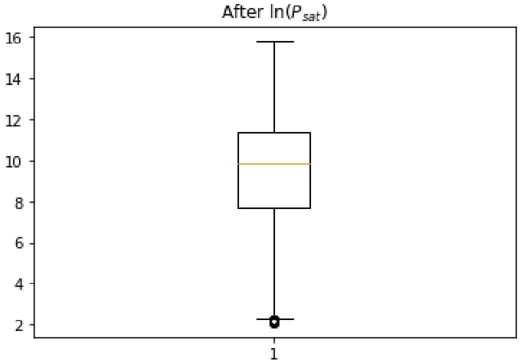
Analysis & Cleansing







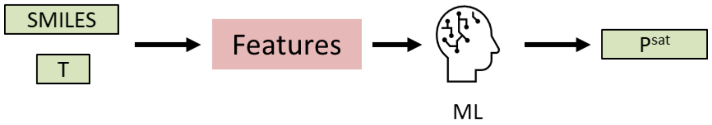
Value	ln(P^{sat})
Min	-1.10 x 10^5
Max	340.72
Point	2,246



Value	ln(P^{sat})	P^{sat} (atm)
Min	2.09	8.1 x 10^{-5}
Max	15.78	71.58
Point	1,787	

Methodology

Fingerprint Selecting



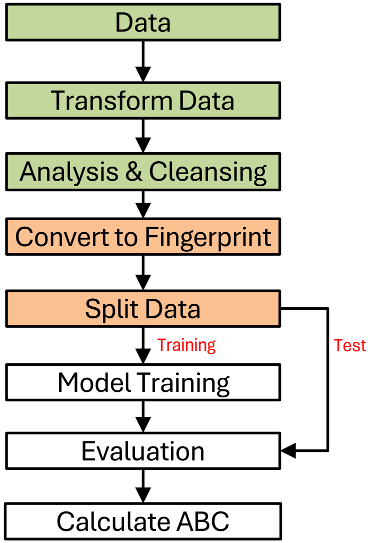
SMILES

T

Features

ML

P^{sat}



Data

Transform Data

Analysis & Cleansing

Convert to Fingerprint

Split Data

Training

Test

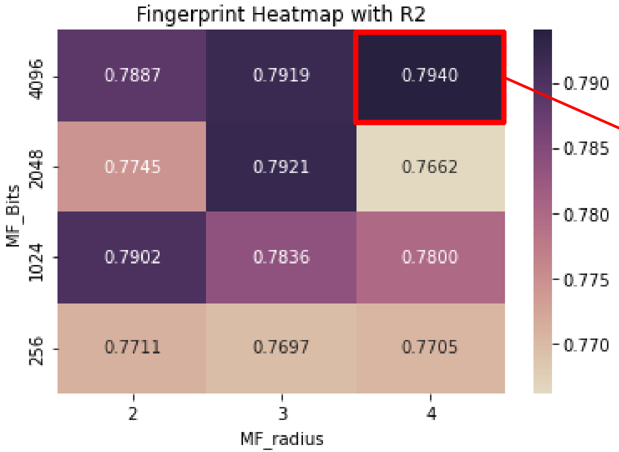
Model Training

Evaluation

Calculate ABC

All Data		
Alcohol	Alkyne	Carboxylic Acid
Aldehyde	Amide	Ester
Alkane	Amine	Ether
Alkene	Aromatic	Ketone

Training Data	Test Data
80%	20%

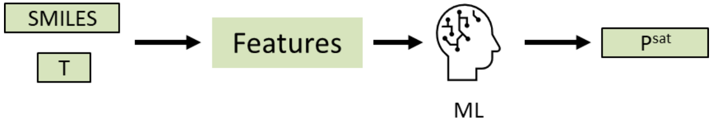


The best radius and bits for Fingerprint

Fig 6. Heatmap Fingerprint for P^{sat} modeling

Methodology

Training



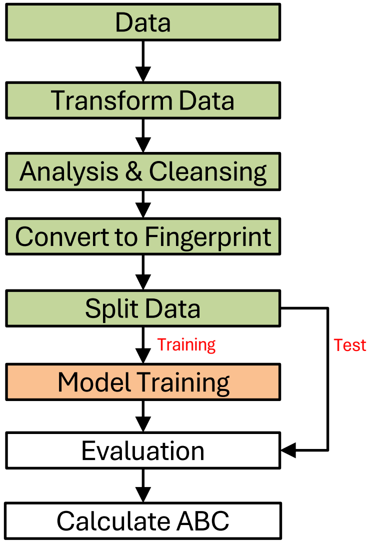
SMILES

T

Features

ML

P^{sat}



Data

Transform Data

Analysis & Cleansing

Convert to Fingerprint

Split Data

Training

Test

Model Training

Evaluation

Calculate ABC

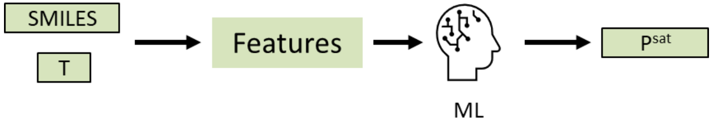
Table 7. Setting of All Learning Algorithms

ML Algorithm	Hyperparameter
DT	max depth = None, min samples leaf = 1, min samples split = 2
RF	max depth = None, max feature = None, n estimators = 200
XGB	max depth = 5, learning rate = None, n estimators = 400
KNN	Algorithm = ball tree, n neighbors = 5, weights = distance

Note : Each ML algorithm use K-Fold with K = 5 and Grid Search

Result

Evaluation



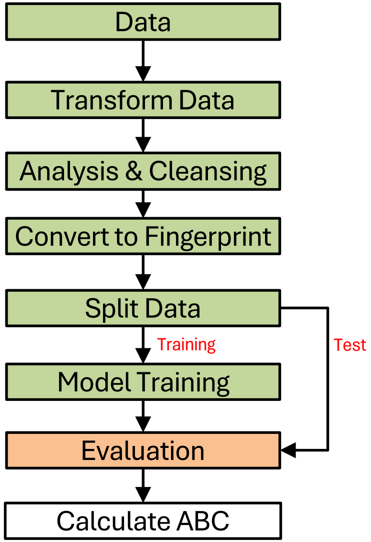
SMILES

T

Features

ML

P^{sat}



Data

Transform Data

Analysis & Cleansing

Convert to Fingerprint

Split Data

Training

Test

Model Training

Evaluation

Calculate ABC

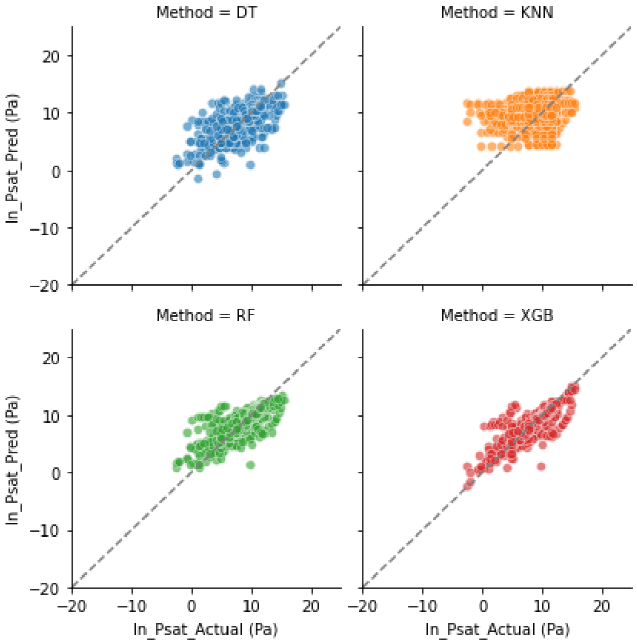


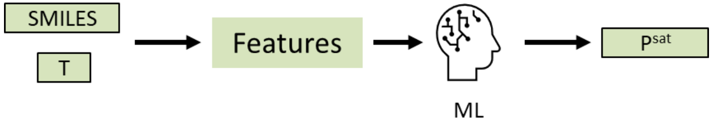
Fig 7. Result of each algorithm

Table 8. Error metrics of each algorithm

Method	MAE	MAPE (%)	RMSE	R^2
DT	0.80	15.00	1.38	0.68
KNN	1.76	22.30	2.42	-1.24
RF	0.63	15.90	1.10	0.73
XGB	0.59	13.60	1.06	0.80

Result

Evaluation



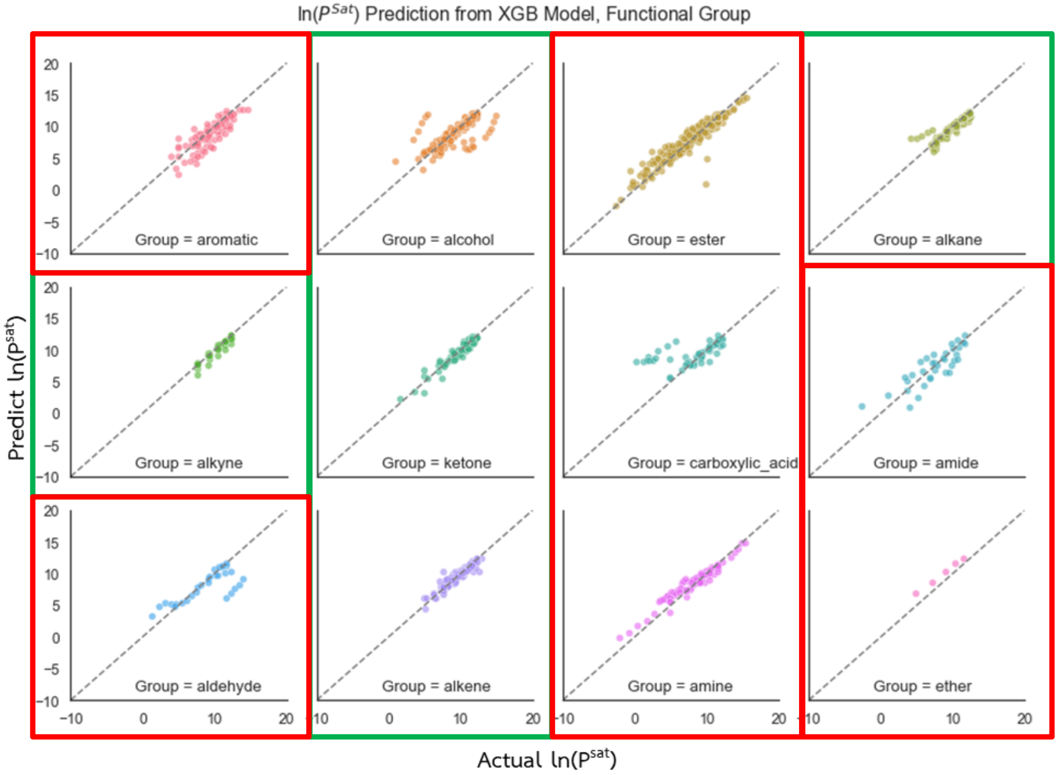


Fig 8. P^{sat} Prediction on CHON Scope Group by Functional Group

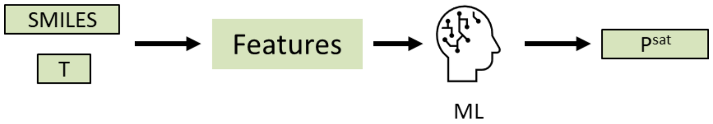
The functional group that this ln(P^{sat}) prediction model can predict well is that the group that have condition with $R^2 \geq 0.5$ and MAPE ≤ 10%

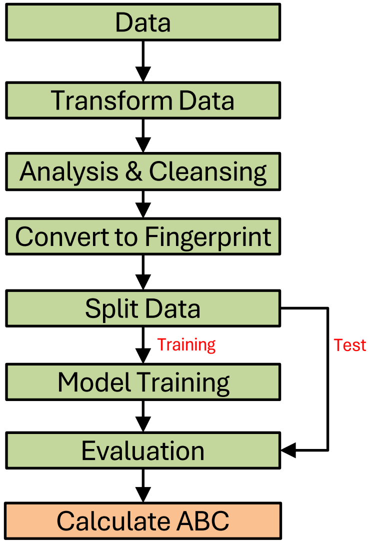
Acceptable: Alcohol, Alkane, Alkyne, Alkene, Ketone

Unacceptable: Aromatic, Ester, Carboxylic Acid, Amide, Aldehyde, Amine, Ether

Result

Calculation





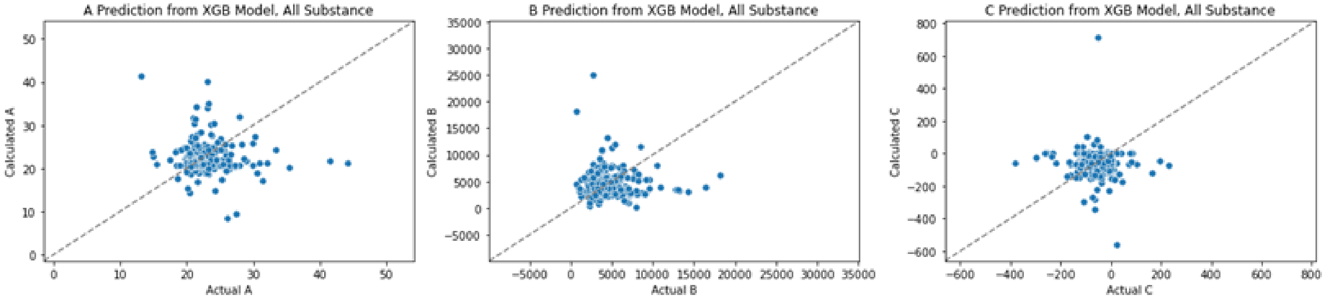
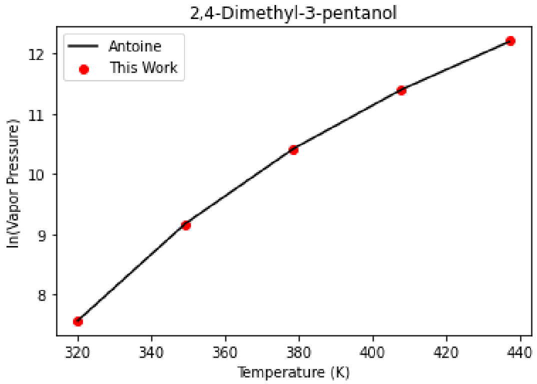


Fig 9. Graph Distribution of Antoine Coefficients

Calculate A, B, C for Predicted ln(P^{sat}) from the best model

Result

Calculation



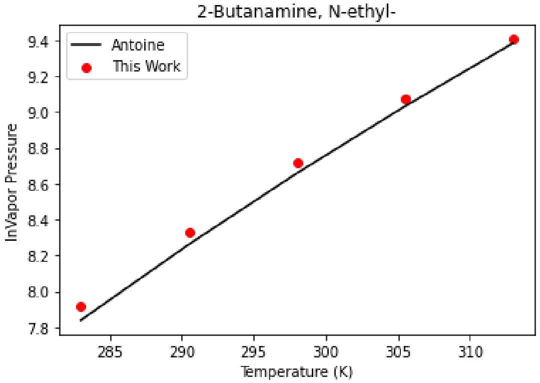


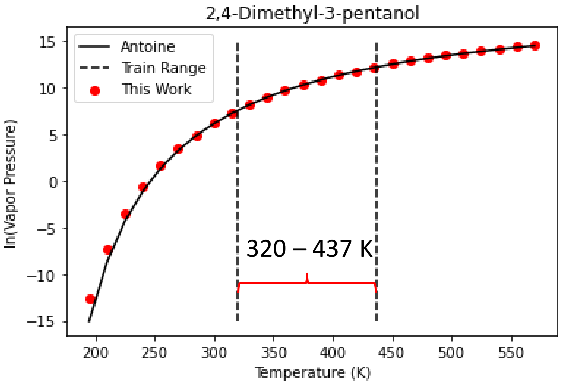
Table 8. Antoine Coefficients in this work and Antoine

SMILES	Reference			Calculation			Train Temp Range (K)
	A	B	C	A	B	C	
2,4-Dimethyl-3-pentanol	19.85	2370.74	-127.05	20.35	2615.59	-115.52	320 – 440
2-Butanamine, N-ethyl-	21.24	3083.56	-52.86	18.67	2006.16	-96.46	283 – 313

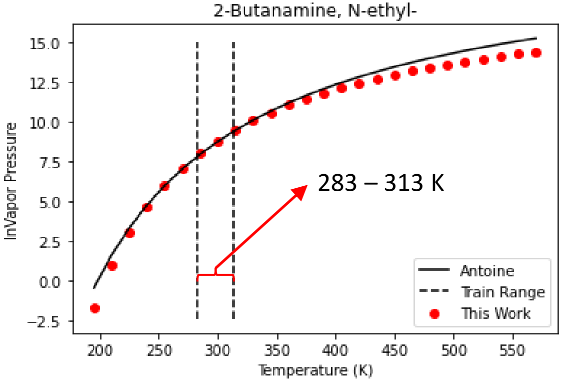
Antoine Coefficients that yield similar results to the reference source, even if slightly different, can still be used reliably.

Result

Calculation


2,4-Dimethyl-3-pentanol

320 – 437 K


2-Butanamine, N-ethyl-

283 – 313 K

When comparing temperatures at different range, it is observed that greater temperature differences range lead to more accurate predictions compared to the reference, especially when considering the lowest and highest temperatures from all references for all substances.

Table 9. Antoine Coefficients in this work and Antoine

SMILES	Reference			Calculation		
	A	B	C	A	B	C
2,4-Dimethyl-3-pentanol	19.85	2370.74	-127.05	20.35	2615.59	-115.52
2-Butanamine, N-ethyl-	21.24	3083.56	-52.86	18.67	2006.16	-96.46

Note: Temperature Range in Dataset (195 - 574 K)

Conclusion

1. Use Morgan Fingerprint to resolve Previous Work Problem “Similar Structures, get same Features”
2. Use Morgan Fingerprint to establish Properties Prediction Model (Use for Organic Compound C, H, O and N) : T_b - CHON Scope with MAPE = 5.94% and $R^2 = 0.66$

   : $\ln(P^{sat})$ - CHON Scope with MAPE = 13.6% and $R^2 = 0.80$
3. Learn about Machine Learning Algorithm, Cross Validation and Hyperparameter Tuning in each Algorithm and Neural Network Implementation for Improving Properties Prediction model

Suggestions

1. Increasing the number of bits in a molecular fingerprint can help separate substructures, but some cases cannot be separated regardless of bit addition.
2. Deep Learning for modeling is more complex than Machine Learning, making it challenging to predict properties.
3. While modeling is faster, finding the right model takes time, so planning information preparation is crucial.
4. To improve model accuracy, consider creating a model that separates specific boundaries, focusing on specific functional groups.

Q&A