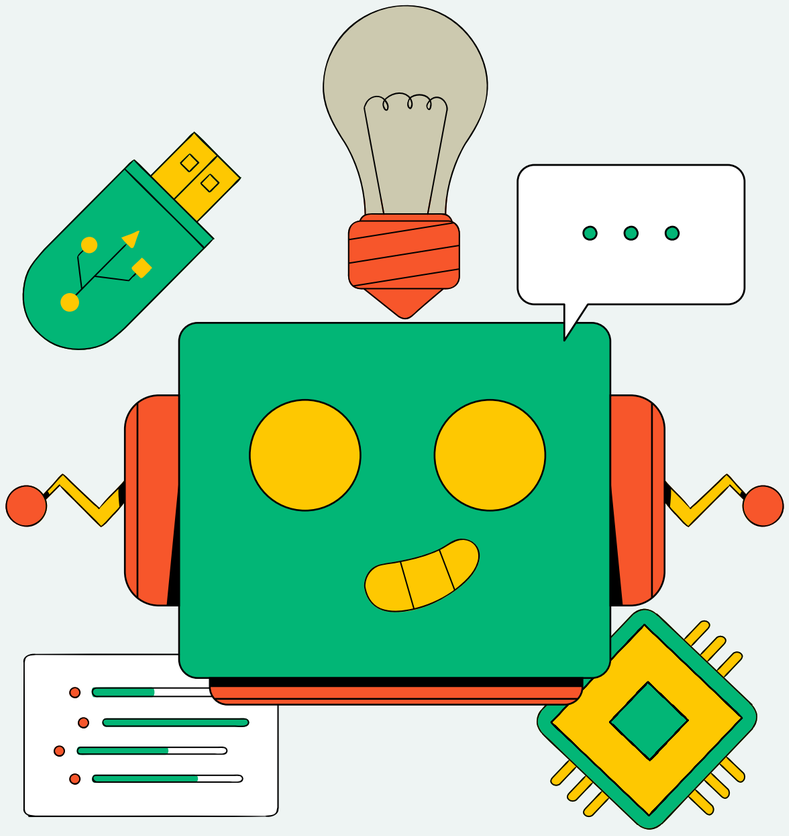
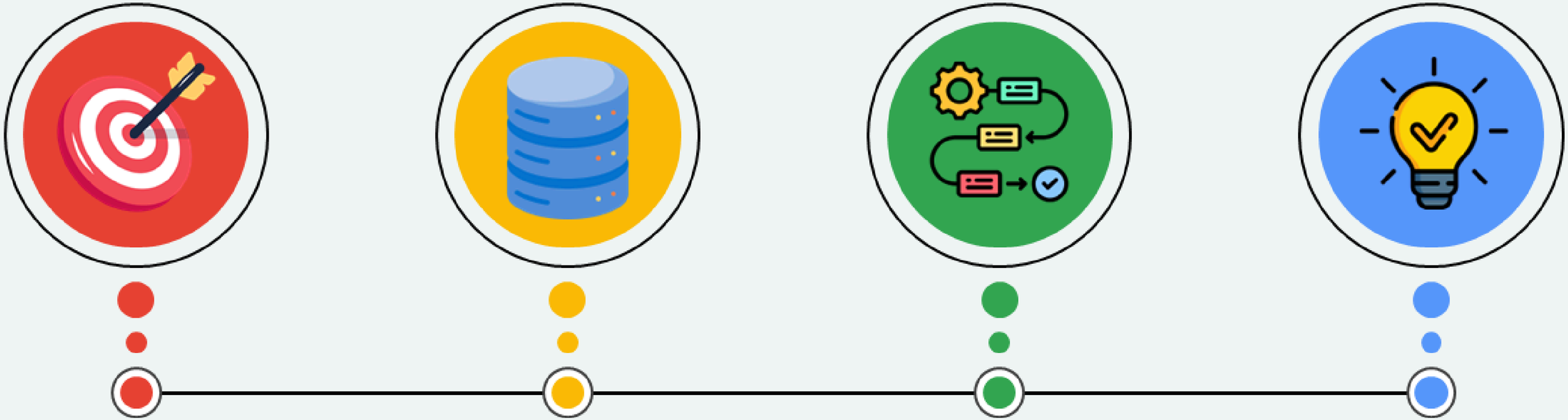


DATA MINING "MELTING POINT"

AGENDA

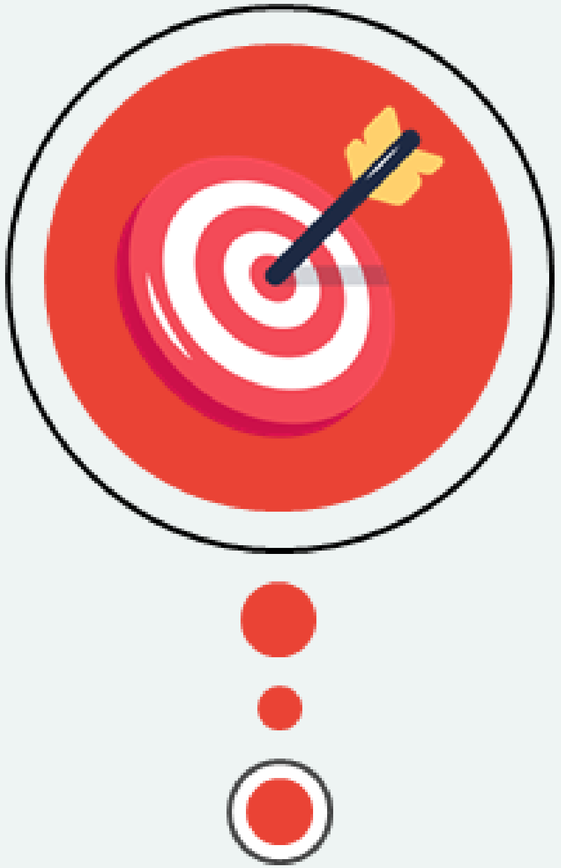


Problem
Motivation
Research Question

Data Exploration
Feature
Engineering

Methodology
Models
Results

Interpretation
Key Learnings
Future Work



Problem
Motivation
Research Question

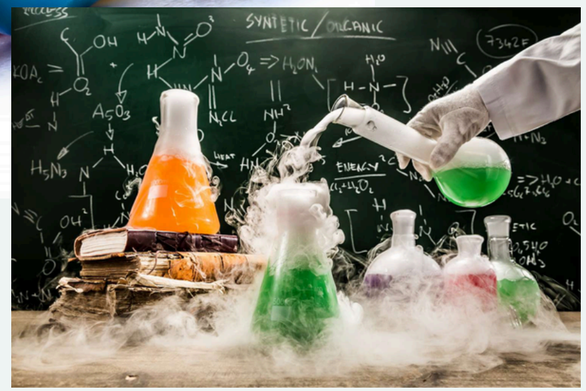
CONTEXT& RELEVANCE

JOHN HEDENGREN · COMMUNITY PREDICTION COMPETITION · 15 DAYS TO GO

Submit Prediction

Thermophysical Property: Melting Point

Your goal is to build ML models that predict melting point in temperature units of Kelvin for organic compounds given molecular descriptors.



PROBLEM

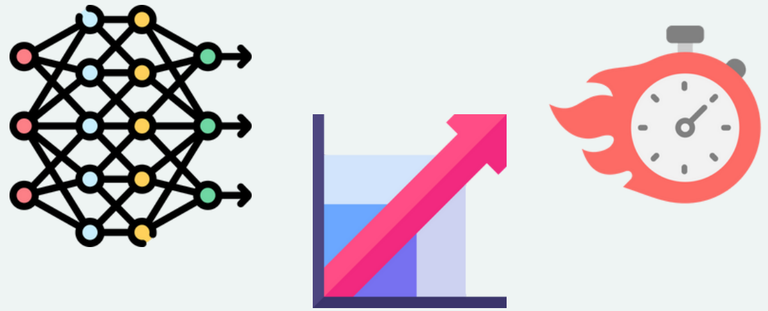
Experimental Determination

- Laboratory measurements are expensive
- Experiments require significant time
- Large chemical space cannot be fully tested

GOAL

Machine Learning Based Prediction

- Predict MT without physical experiments
- Fast and scalable evaluation
- Enables early stage screening of compounds



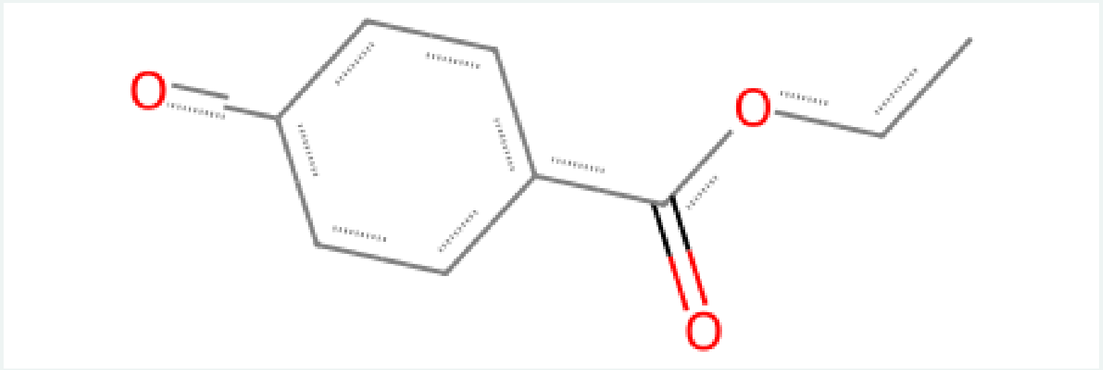
CENTRAL RESEARCH QUESTION

“To what extent can the melting point of organic molecules be predicted from molecular structure using machine learning methods?”

Data Exploration
Feature
Engineering

WHAT ARE WE TRYING TO PREDICT?

ID	SMILES	Tm	Group 1	Group 2	Group 15	Group 30	...	Group 424
2175	CCOC(=O)c1ccc(O)cc1	213.15	0	1	0	2	...	0



Two dimensional representation of an organic molecule

Groups	Best Pattern	Correlation
Group 1	CH3	(0.915379)
Group 2	CH2	(0.923681)
Group 15	Aromatic_C	(0.897967)
Group 30	Phenol_OH	(0.985509)
Group 45	COO	(0.269556)
Group 424	Aromatic C	(0.278946)

DATA SET

Files

- train.csv : Features + target (Tm)
- test.csv : Features only, no target
- sample_submission.csv : Template with columns [id,Tm]

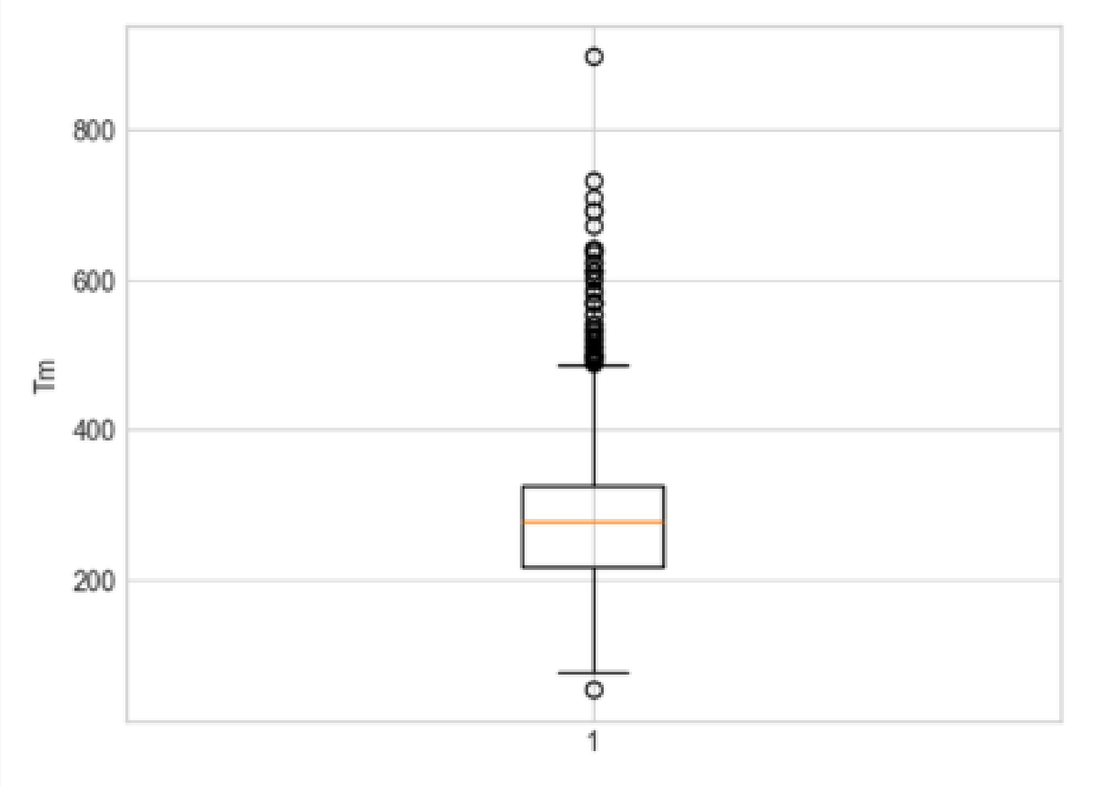
Columns

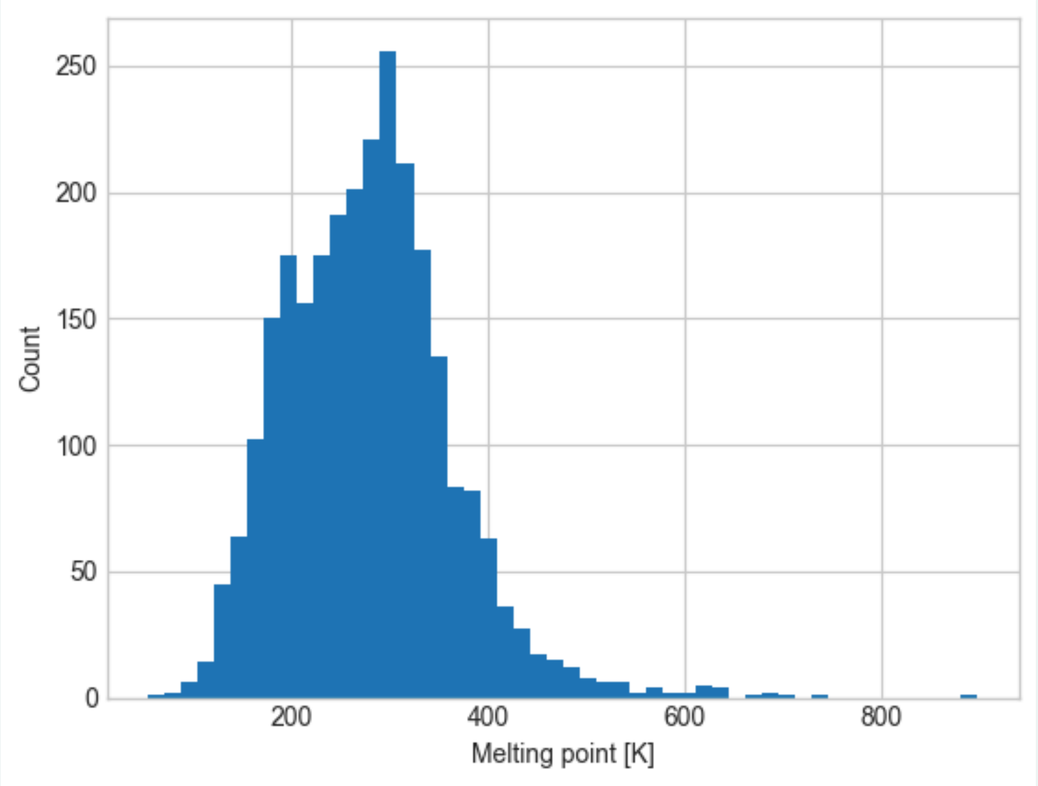
- id : unique ID
- SMILES : molecular string
- Group 1..N : descriptor features
- Tm : melting point (Kelvin) [train only]

	id	SMILES	Tm	Group 1	Group 2	Group 3	Group 4	Group 5	Group 6	Group 7	...	Group 415	Group 416	Group 417	Group 418	Group 419	Group 420	Group 421	Group 422	Group 423	Group 424
0	2175	FC1=C(F)C(F)(F)C1(F)F	213.15	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
1	1222	c1ccc2c(c1)ccc3Nc4ccccc4c23	407.15	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
2	2994	CCN1C(C)=Nc2ccccc12	324.15	2	1	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
3	1704	CC#CC(=O)O	351.15	1	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
4	2526	CCCCC(S)C	126.15	2	3	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
2657	707	ClCCBr	256.45	0	1	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
2658	1573	N#CC(Cl)(Cl)Cl	231.15	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
2659	1634	Cc1ccc2c(C)cccc2c1	256.25	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
2660	3022	CCC(=O)c1ccc2ccccc2c1	333.15	1	1	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
2661	1673	Brc1ccc(cc1)N(C)C	328.15	2	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0

DATA EXPLORATION

Distribution of the Target Variable Tm



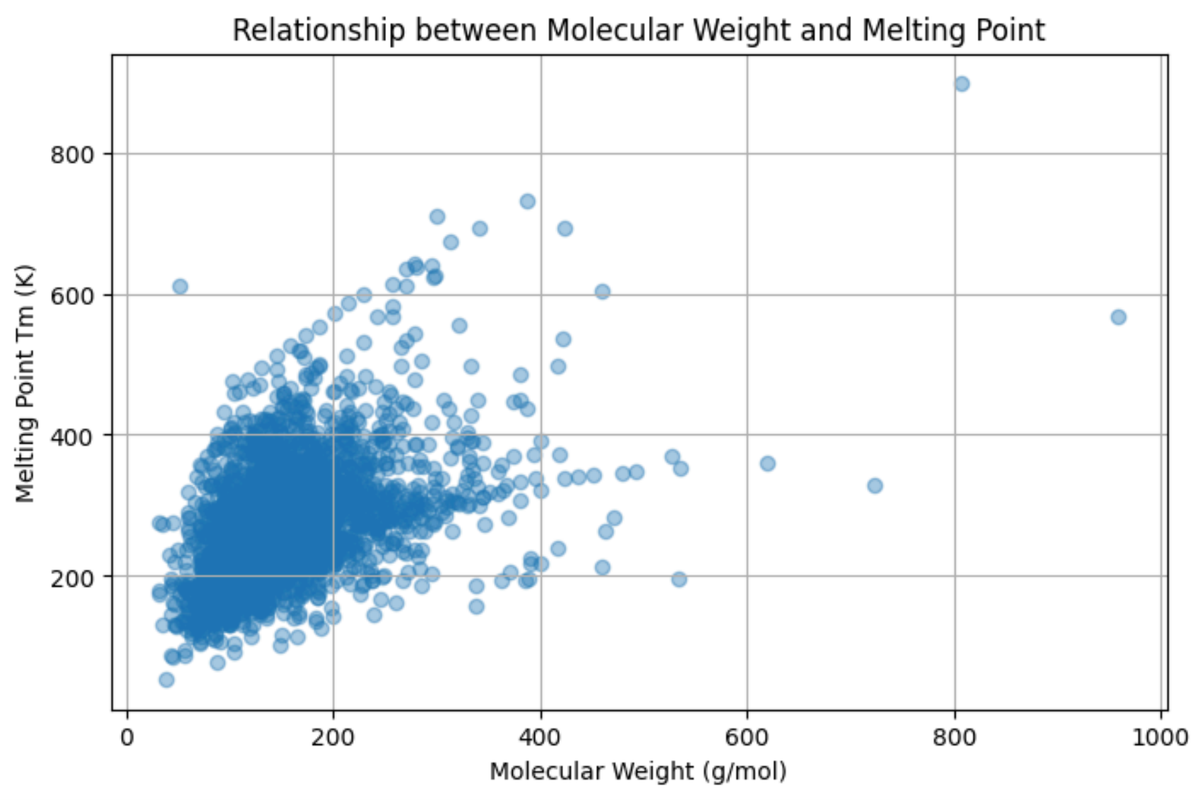


- Max Tm = 897.15K
- Min Tm = 53.54K

- right skewed distribution
- The mean Tm ≈ 278 K

FEATURE ENGINEERING: WEIGHT

We used the SMILES to derive the mol mass of the molecules and added this as a new feature

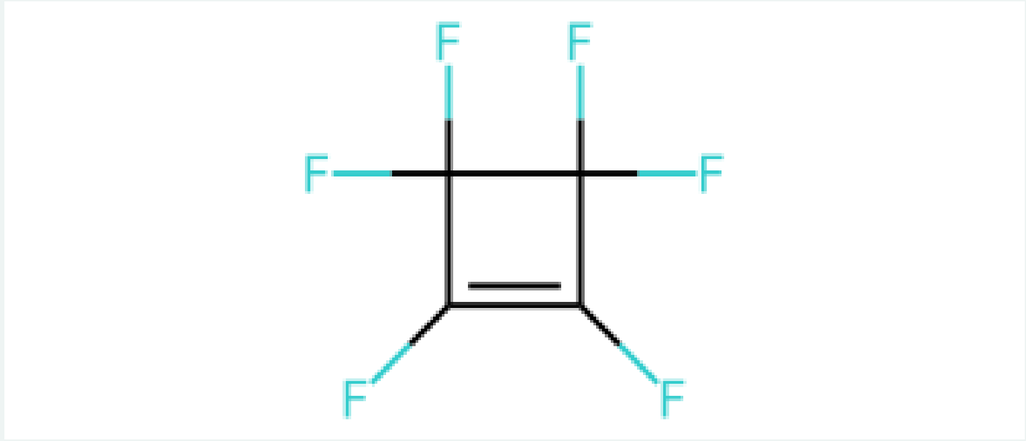


index	SMILES	MolWeight	Tm
0	FC1=C(F)C(F)(F)C1(F)F	162.032	213.15
1	c1ccc2c(c1)ccc3Nc4ccccc4c23	217.271	407.15
2	CCN1C(C)=Nc2ccccc12	160.220	324.15
3	CC#CC(=O)O	84.074	351.15
4	CCCCC(S)C	118.245	126.15

FEATURE ENGINEERING: STRUCTURE (1)

Example with one Tuple

SMILE: FC1=C(F)C(F)(F)C1(F)F



A chemist would recognize:

- Symmetry
- No H-Bridges
- negative electric charge
- ...

Expectation: High melting point

The Groups would look like this:

Group	Value
Aromatic ring	1
Halogen (F)	6
CH3	0
OH	0
C	4

Problem:

We know what we have but there is none relations to other groups.

FEATURE ENGINEERING: STRUCTURE (2)

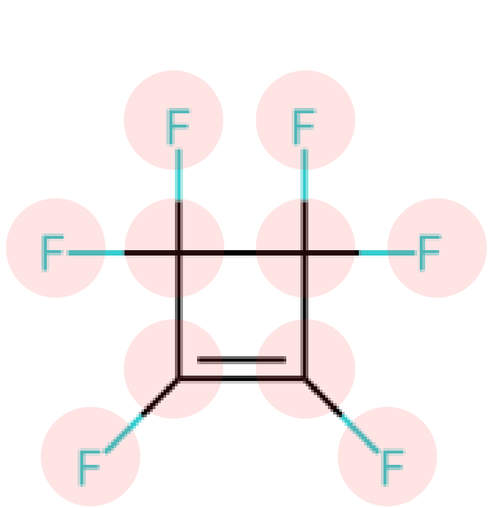
We look at every Atom as a center. It’s connected to other atoms with a bond

- A Feature which looks at the atom itself **(r=0)**
- A Feature which looks at the atom and the nearest neighbor **(r=1)**
- A Feature which looks at itself, it’s neighbor and the direct neighbor of the neighbor **(r=2)**

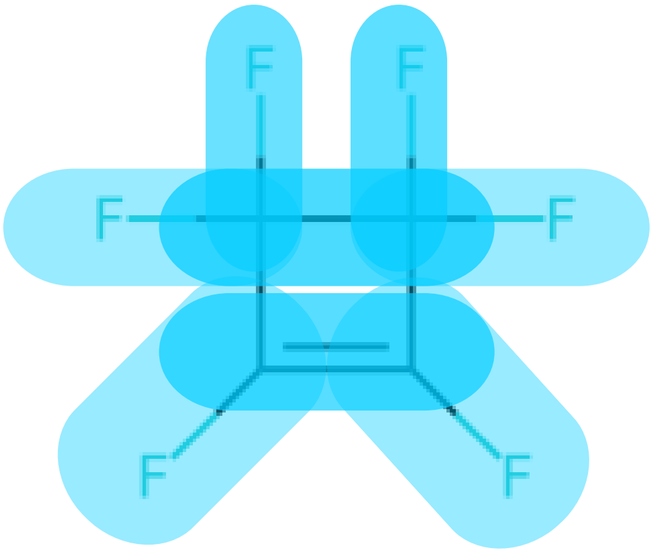
Two Examples:

Feature “C” (r=0): 4
→ because there are 4 “C” which are sp2 hybridized and part of a ring C (sp2, ring): 4

For r=1 we look at the bonds: F-C (sp2, ring): 6 and also C-2*F(ring): 2

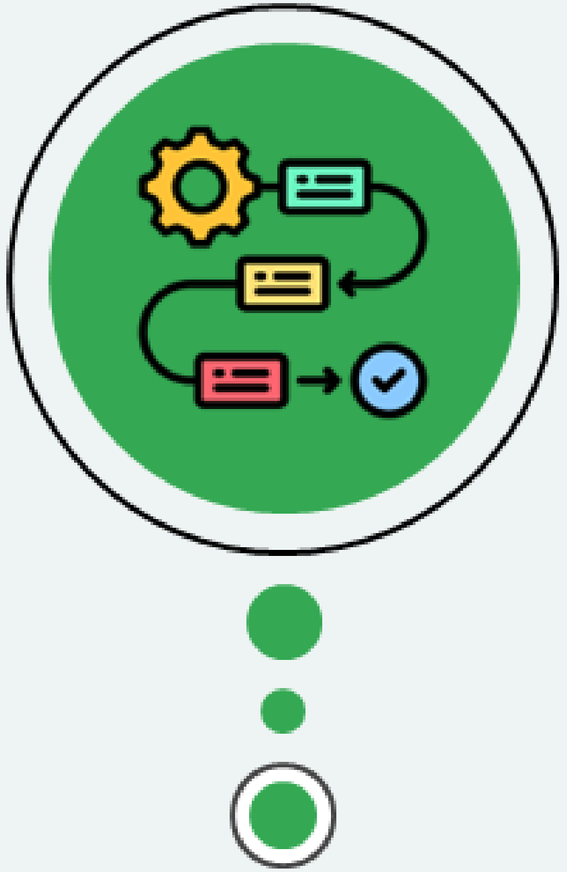


r=0



r=1

→ We got 2432 new Features!



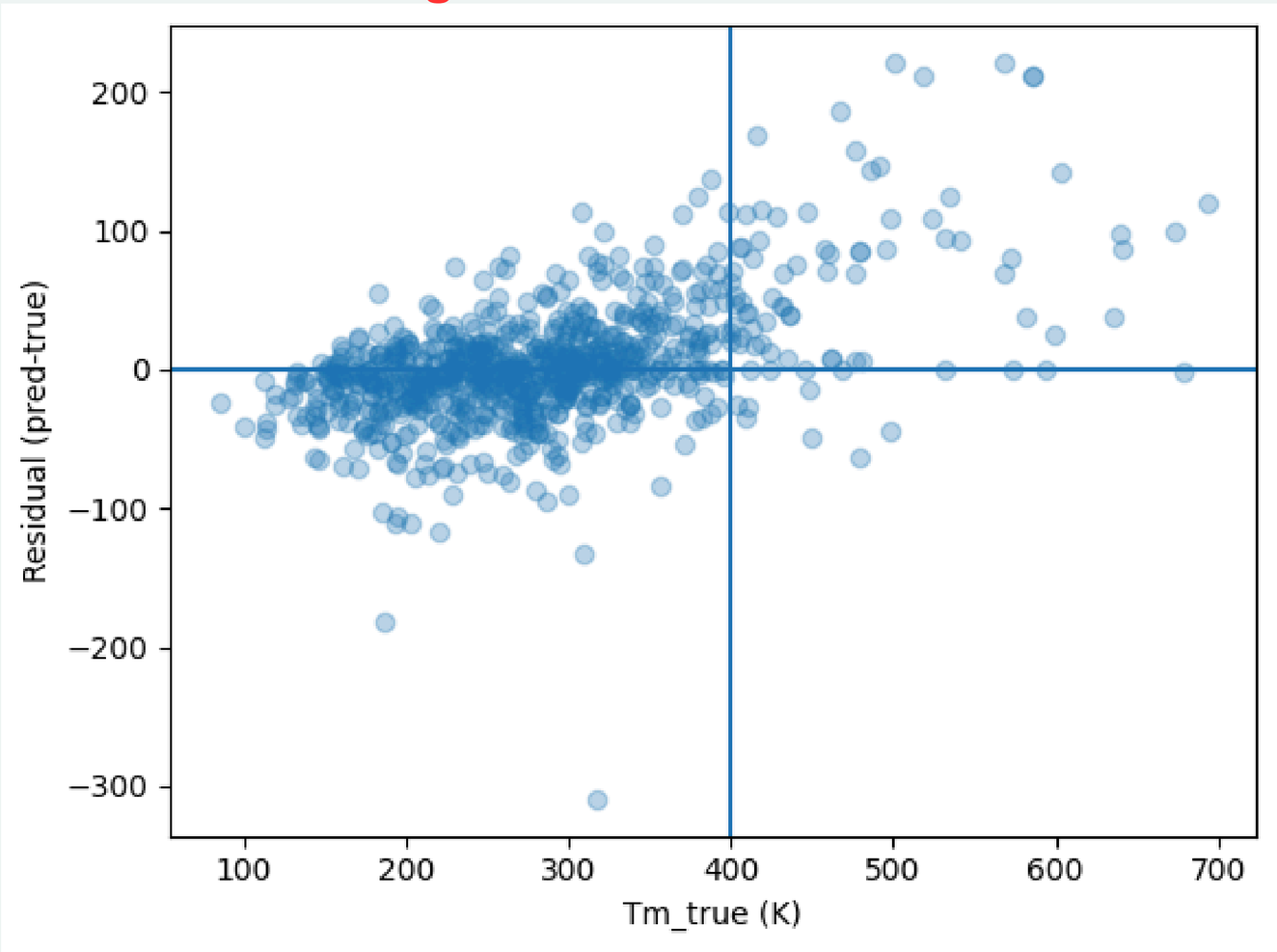
Methodology
Models
Results

MODEL APPROACH

- **Gradient Boosted Decision Trees (XGBoost)**
- Suitable for high-dimensional, non-linear tabular data
- Input: SMILES-derived features (descriptors + fingerprints)
- Evaluation metric: Mean Absolute Error (MAE)

→ We recognized that our data has many outliers as soon as the Tm gets higher!

This gave us an MAE of about 25K



TWO-STAGE (RESIDUAL) MODEL (1)

→ Stage 1 – Base Model

- XGBoost regression on SMILES-derived features
- Sample weighting to emphasize high melting points (Tm > 450 K)
- Produces initial melting point prediction

→ Stage 2 – Residual Model

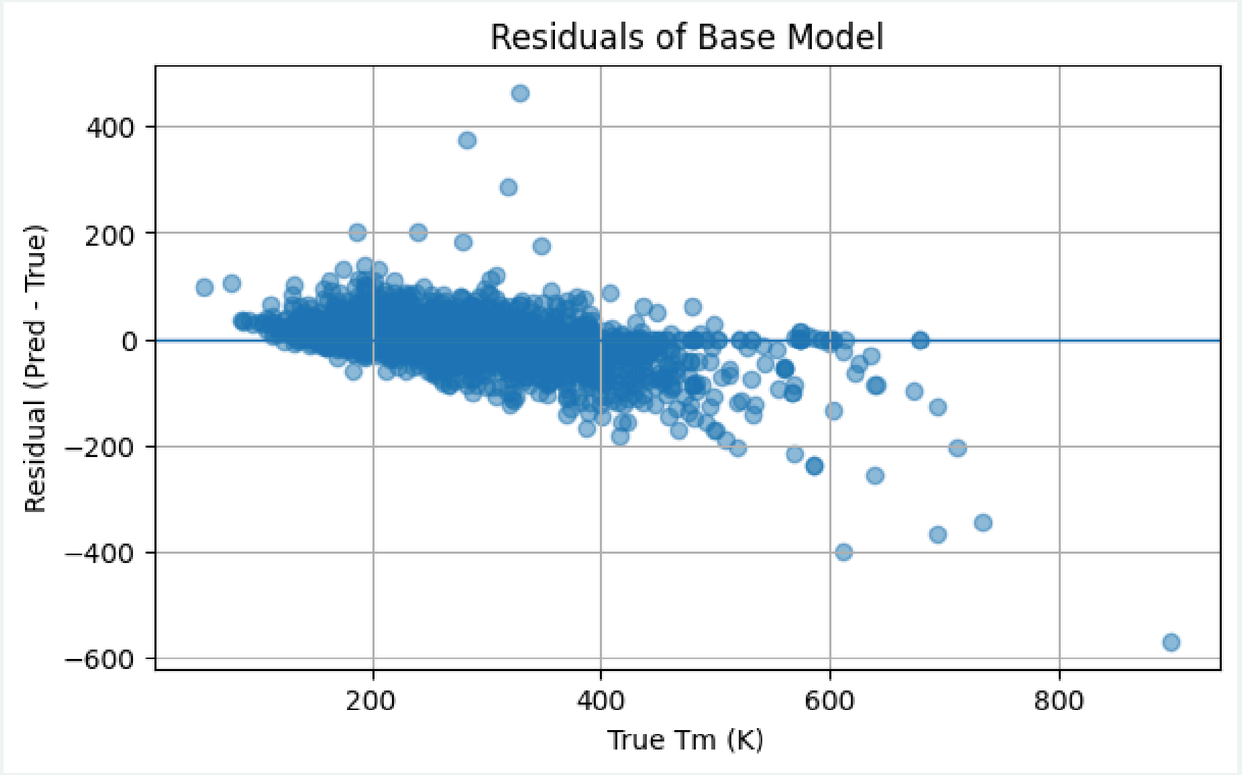
- Second XGBoost model trained on prediction residuals
- Learns systematic errors not captured by the base model

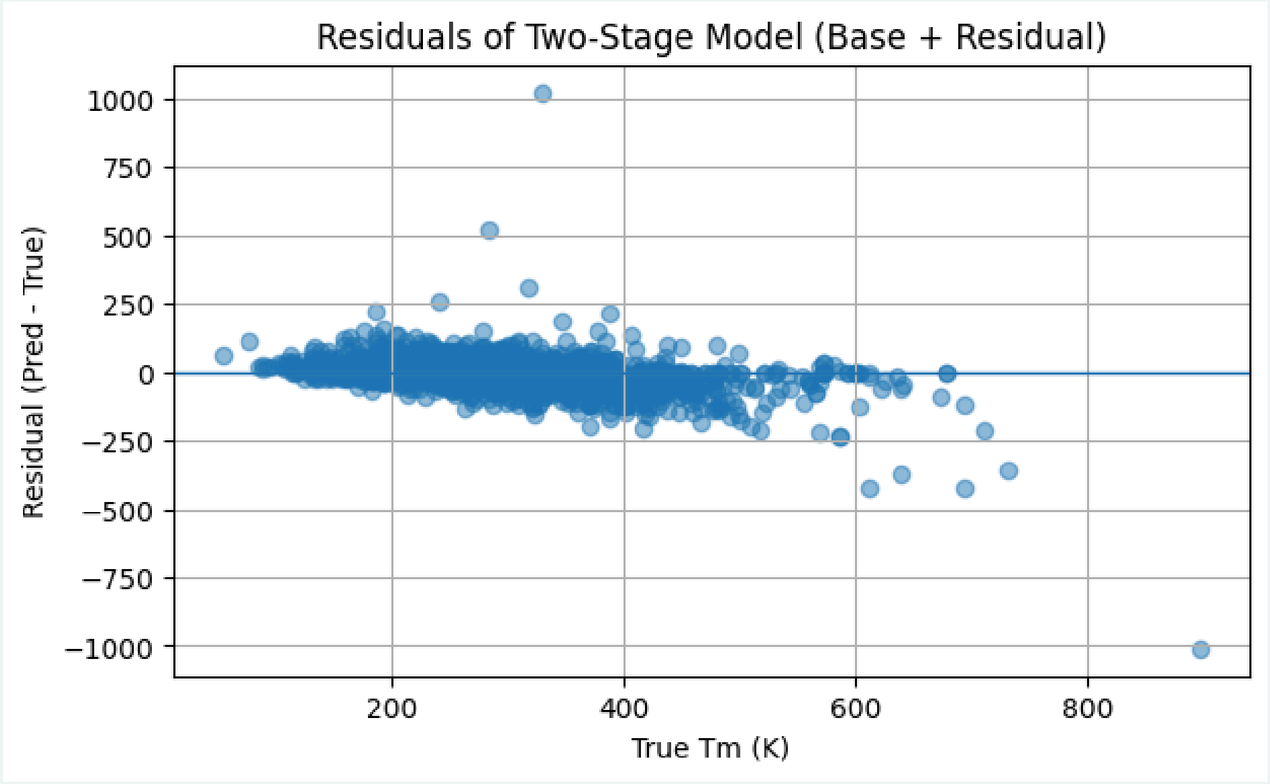
→ Final Prediction

Final Tm = Base prediction + Residual correction

With this approach we managed to achieve an MAE of 23.78K

TWO-STAGE (RESIDUAL) MODEL (2)



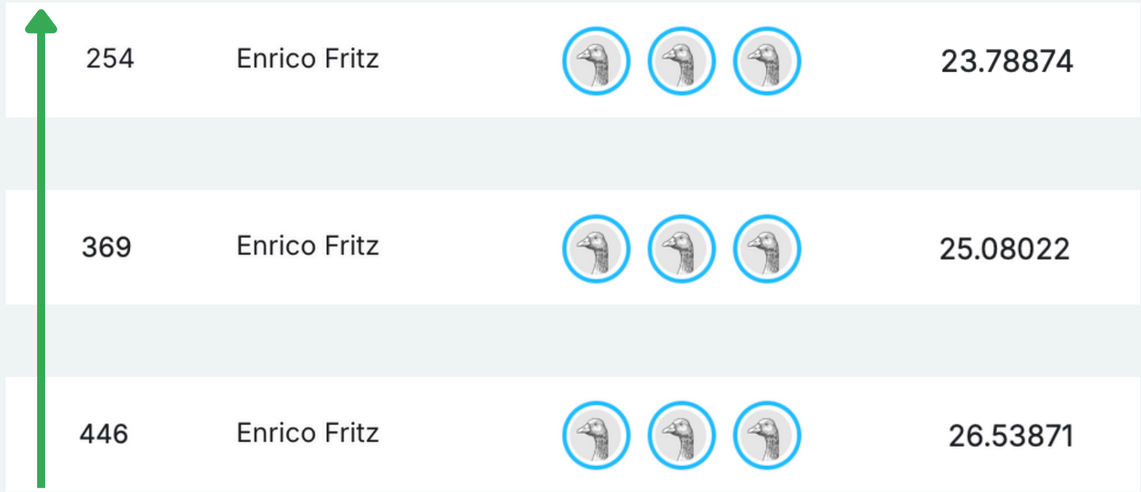


Interpretation
Key Learnings
Future Work

RESULTS OVERVIEW

Model	MAE (K)
Neural Network	~31
XGBoost Baseline	~25.4
Two-Stage XGBoost	**~23.8**

current Ranking: 254 / 1033



ANSWERING THE RESEARCH QUESTION

“To what extent can the melting point of organic molecules be predicted from molecular structure using machine learning methods?”

Answer:

- **Melting points can be predicted** to a large extent
- Reliable predictions for the majority of compounds
- **Strong performance in the mid-range** (200–400 K)
- Reduced accuracy for **rare and extreme molecular structures**

FUTURE OUTLOOK

Scale the model on an AWS server

Thank you!

Any questions?