香港中文大學
The Chinese University of Hong Kong

Ron

Some papers in chemistry and machine learning

30 June 2025

Index

- Tokenization
- NovelSeek
- Tools
- ChemOCR

Tokenization

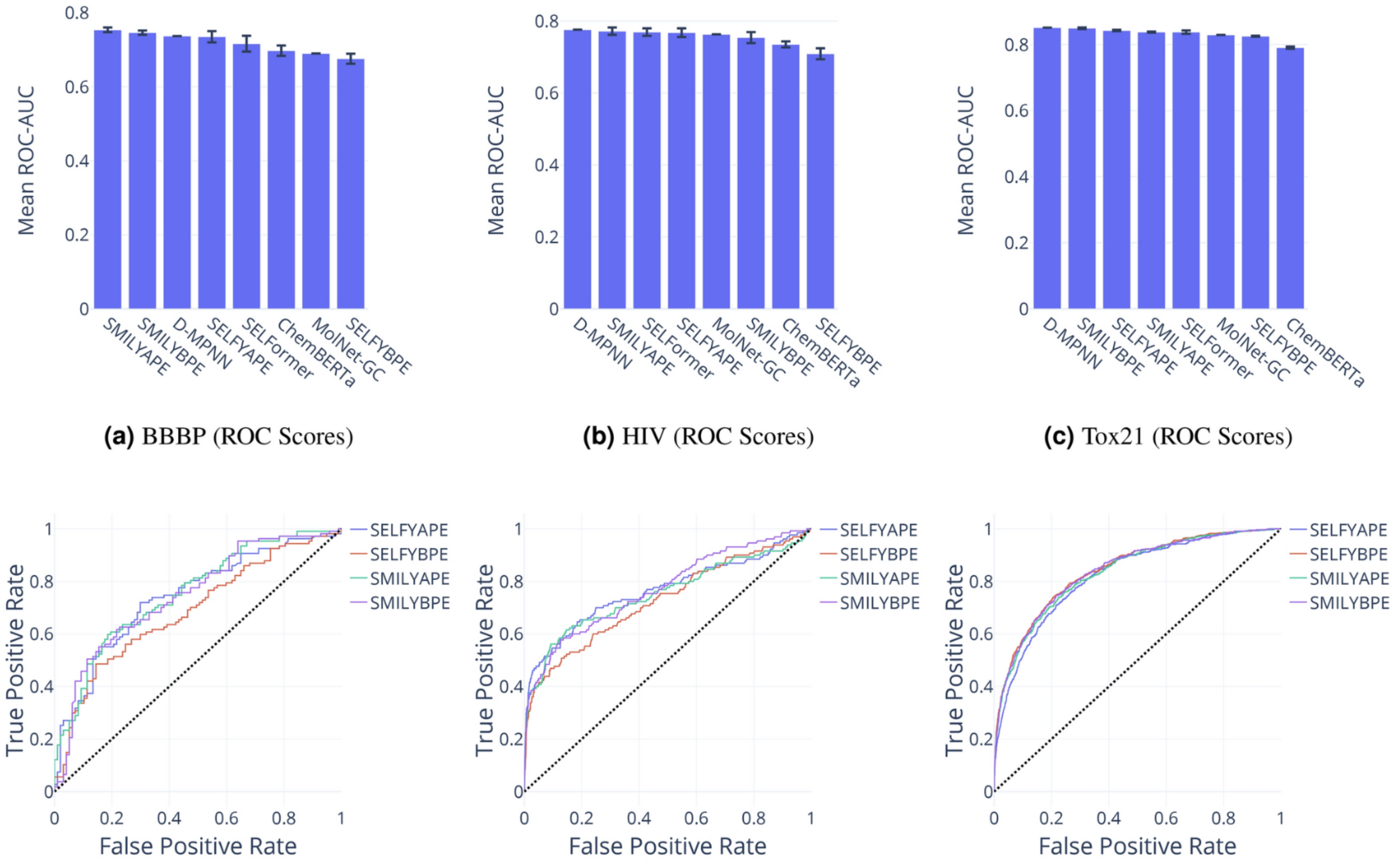
A way to feed molecules to machines

30 June 2025

Is there a difference in performance?

(Not much)

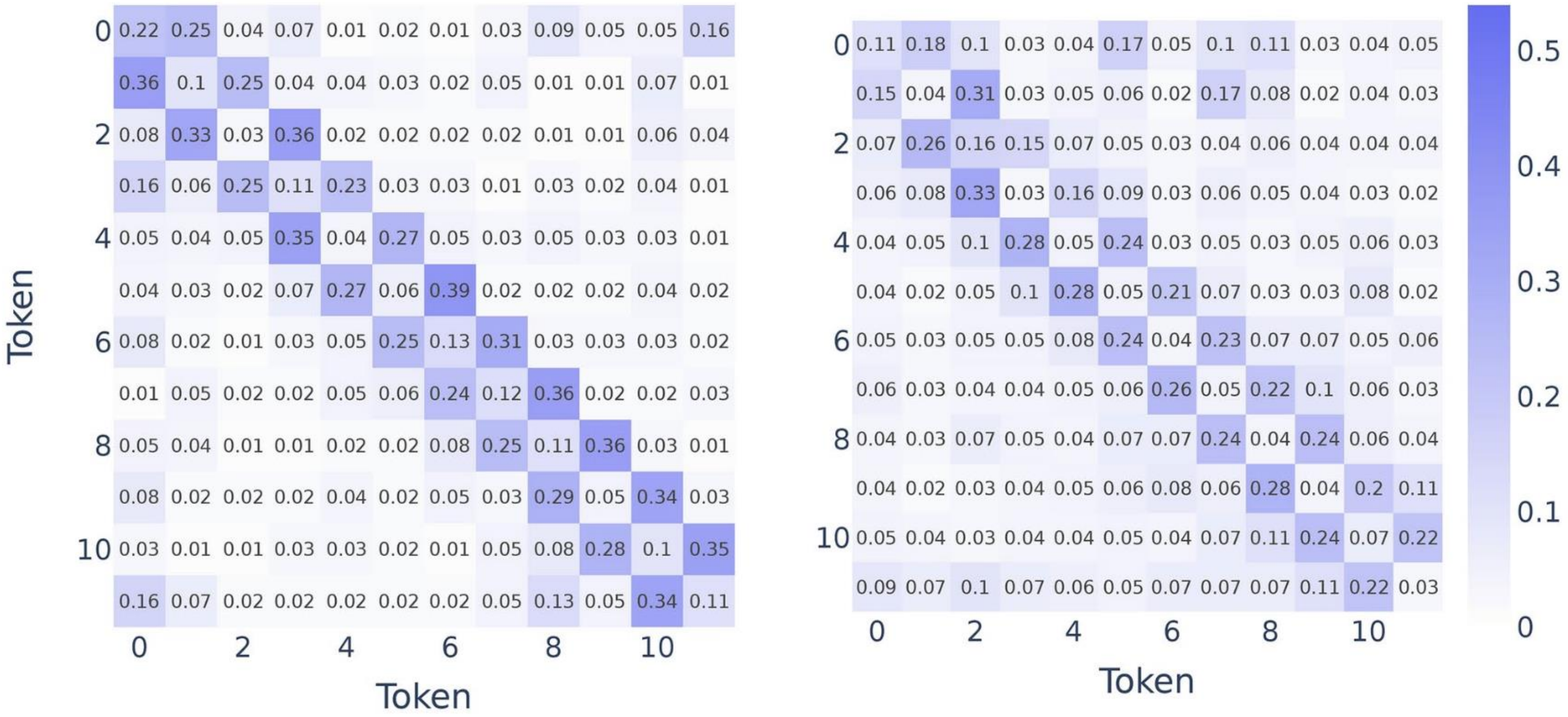
No.



(a) BBBP (ROC Scores) **(b)** HIV (ROC Scores) **(c)** Tox21 (ROC Scores)

(d) BBBP (ROC Curves) **(e)** HIV (ROC Curves) **(f)** Tox21 (ROC Curves)

Fig. 5. ROC-AUC scores and curves of selected benchmarks. Top row: (**a**) BBBP, (**b**) HIV, and (**c**) Tox21 bar charts. Bottom row: (**d**) BBBP, (**e**) HIV, and (**f**) Tox21. The Tox21 is the largest dataset, giving better results with all models.



	0		2		4		6		8		10	
0	0.22	0.25	0.04	0.07	0.01	0.02	0.01	0.03	0.09	0.05	0.05	0.16
	0.36	0.1	0.25	0.04	0.04	0.03	0.02	0.05	0.01	0.01	0.07	0.01
2	0.08	0.33	0.03	0.36	0.02	0.02	0.02	0.02	0.01	0.01	0.06	0.04
	0.16	0.06	0.25	0.11	0.23	0.03	0.03	0.01	0.03	0.02	0.04	0.01
4	0.05	0.04	0.05	0.35	0.04	0.27	0.05	0.03	0.05	0.03	0.03	0.01
	0.04	0.03	0.02	0.07	0.27	0.06	0.39	0.02	0.02	0.02	0.04	0.02
6	0.08	0.02	0.01	0.03	0.05	0.25	0.13	0.31	0.03	0.03	0.03	0.02
	0.01	0.05	0.02	0.02	0.05	0.06	0.24	0.12	0.36	0.02	0.02	0.03
8	0.05	0.04	0.01	0.01	0.02	0.02	0.08	0.25	0.11	0.36	0.03	0.01
	0.08	0.02	0.02	0.02	0.04	0.02	0.05	0.03	0.29	0.05	0.34	0.03
10	0.03	0.01	0.01	0.03	0.03	0.02	0.01	0.05	0.08	0.28	0.1	0.35
	0.16	0.07	0.02	0.02	0.02	0.02	0.02	0.05	0.13	0.05	0.34	0.11

(a) SMILYAPE $C_{17}H_{21}NO_2$

	0		2		4		6		8		10	
0	0.11	0.18	0.1	0.03	0.04	0.17	0.05	0.1	0.11	0.03	0.04	0.05
	0.15	0.04	0.31	0.03	0.05	0.06	0.02	0.17	0.08	0.02	0.04	0.03
2	0.07	0.26	0.16	0.15	0.07	0.05	0.03	0.04	0.06	0.04	0.04	0.04
	0.06	0.08	0.33	0.03	0.16	0.09	0.03	0.06	0.05	0.04	0.03	0.02
4	0.04	0.05	0.1	0.28	0.05	0.24	0.03	0.05	0.03	0.05	0.06	0.03
	0.04	0.02	0.05	0.1	0.28	0.05	0.21	0.07	0.03	0.03	0.08	0.02
6	0.05	0.03	0.05	0.05	0.08	0.24	0.04	0.23	0.07	0.07	0.05	0.06
	0.06	0.03	0.04	0.04	0.05	0.06	0.26	0.05	0.22	0.1	0.06	0.03
8	0.04	0.03	0.07	0.05	0.04	0.07	0.07	0.24	0.04	0.24	0.06	0.04
	0.04	0.02	0.03	0.04	0.05	0.06	0.08	0.06	0.28	0.04	0.2	0.11
10	0.05	0.04	0.03	0.04	0.04	0.05	0.04	0.07	0.11	0.24	0.07	0.22
	0.09	0.07	0.1	0.07	0.06	0.05	0.07	0.07	0.07	0.11	0.22	0.03

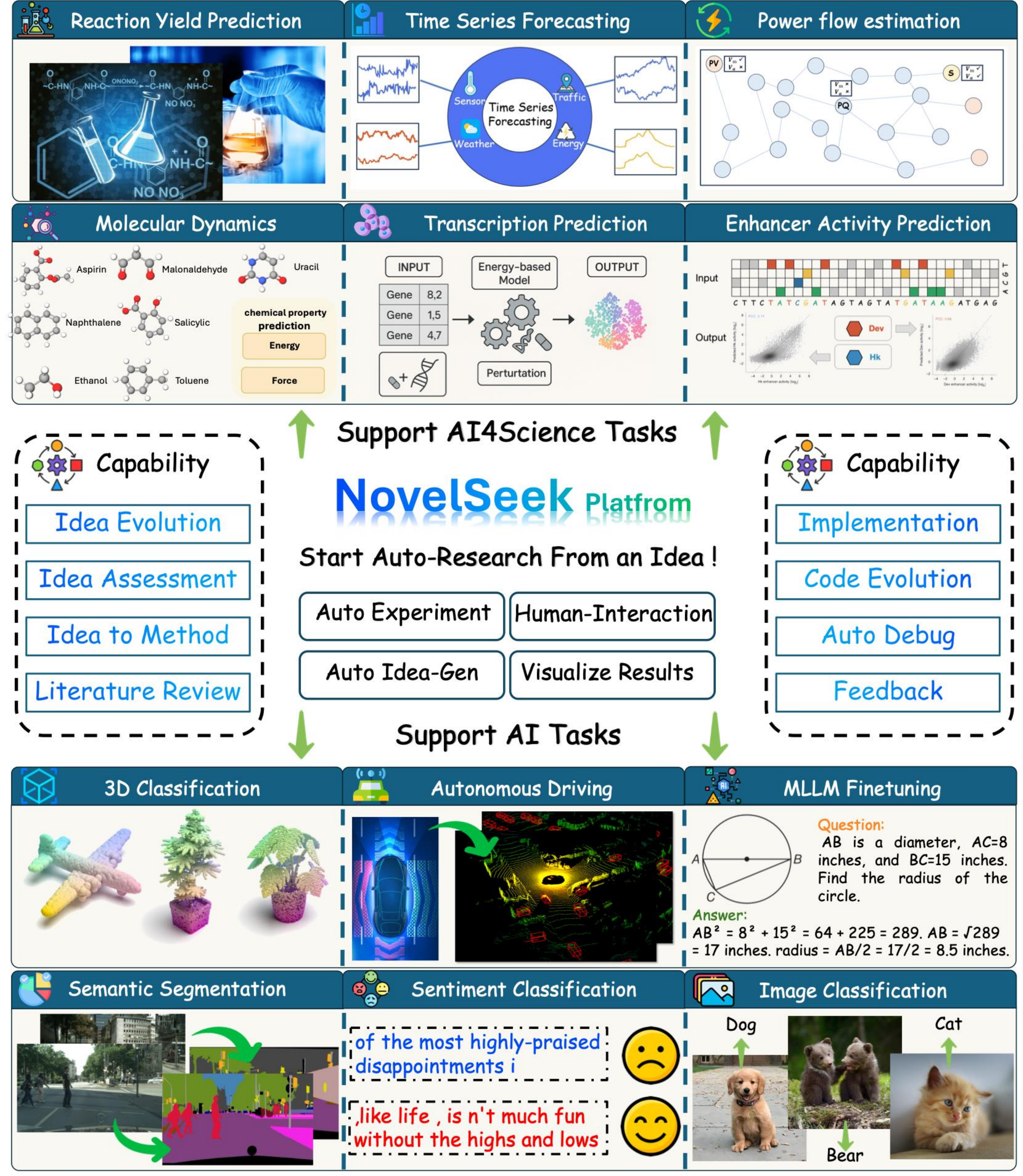
(b) SELFYAPE $C_{17}H_{21}NO_2$

Fig. 6. Attention patterns of the SMILYAPE (**a**) and SELFYAPE (**b**) models, respectively, for molecule $C_{17}H_{21}NO_2$ from the BBBP dataset. The visualization highlights a stronger focus on attention between adjacent tokens (darker squares along the diagonal). However, some attention is also directed towards distant tokens (lighter squares), suggesting the models capture both local context and broader semantic relationships within the molecule.

NovelSeek

When Agent Becomes the Scientist – Building Closed-Loop System from Hypothesis to Verification

30 June 2025

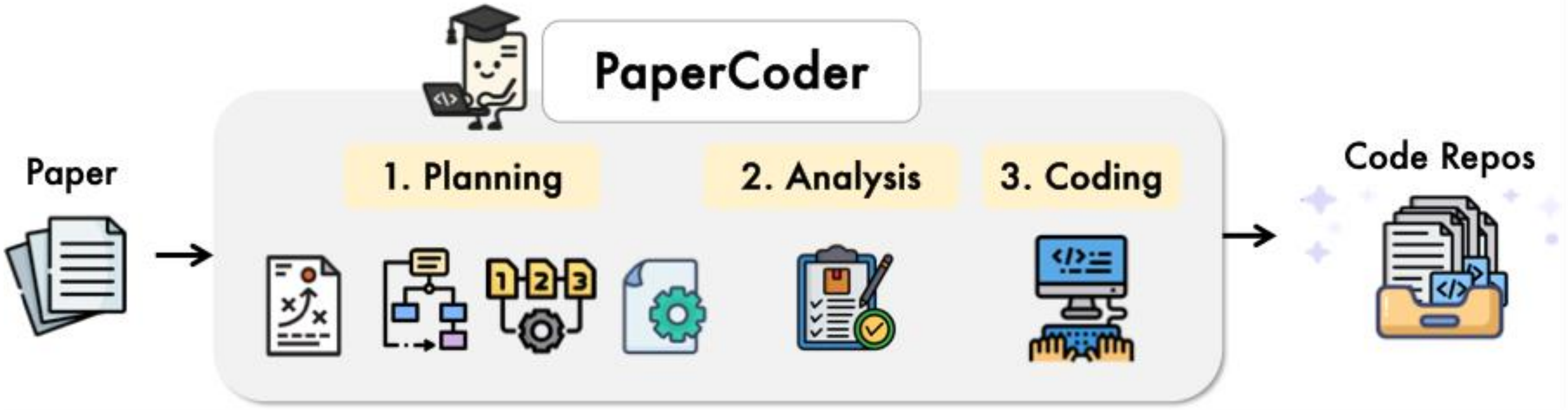

Reaction Yield Prediction
Time Series Forecasting
Sensor
Traffic
Time Series Forecasting
Weather
Energy
Power flow estimation
PV
S
PQ
Molecular Dynamics
Aspirin
Malonaldehyde
Uracil
Naphthalene
Salicylic
chemical property prediction
Energy
Force
Ethanol
Toluene
Transcription Prediction
INPUT
Energy-based Model
OUTPUT
Gene 8,2
Gene 1,5
Gene 4,7
Perturtation
Enhancer Activity Prediction
Input
Output
Dev
Hk
Support AI4Science Tasks
Capability
Idea Evolution
Idea Assessment
Idea to Method
Literature Review
NovelSeek Platfrom
Start Auto-Research From an Idea !
Auto Experiment
Human-Interaction
Auto Idea-Gen
Visualize Results
Capability
Implementation
Code Evolution
Auto Debug
Feedback
Support AI Tasks
3D Classification
Autonomous Driving
MLLM Finetuning
A
B
C
Question:
AB is a diameter, AC=8 inches, and BC=15 inches. Find the radius of the circle.
Answer:
AB² = 8² + 15² = 64 + 225 = 289. AB = √289 = 17 inches. radius = AB/2 = 17/2 = 8.5 inches.
Semantic Segmentation
Sentiment Classification
of the most highly-praised disappointments i
,like life , is n't much fun without the highs and lows
Image Classification
Dog
Cat
Bear

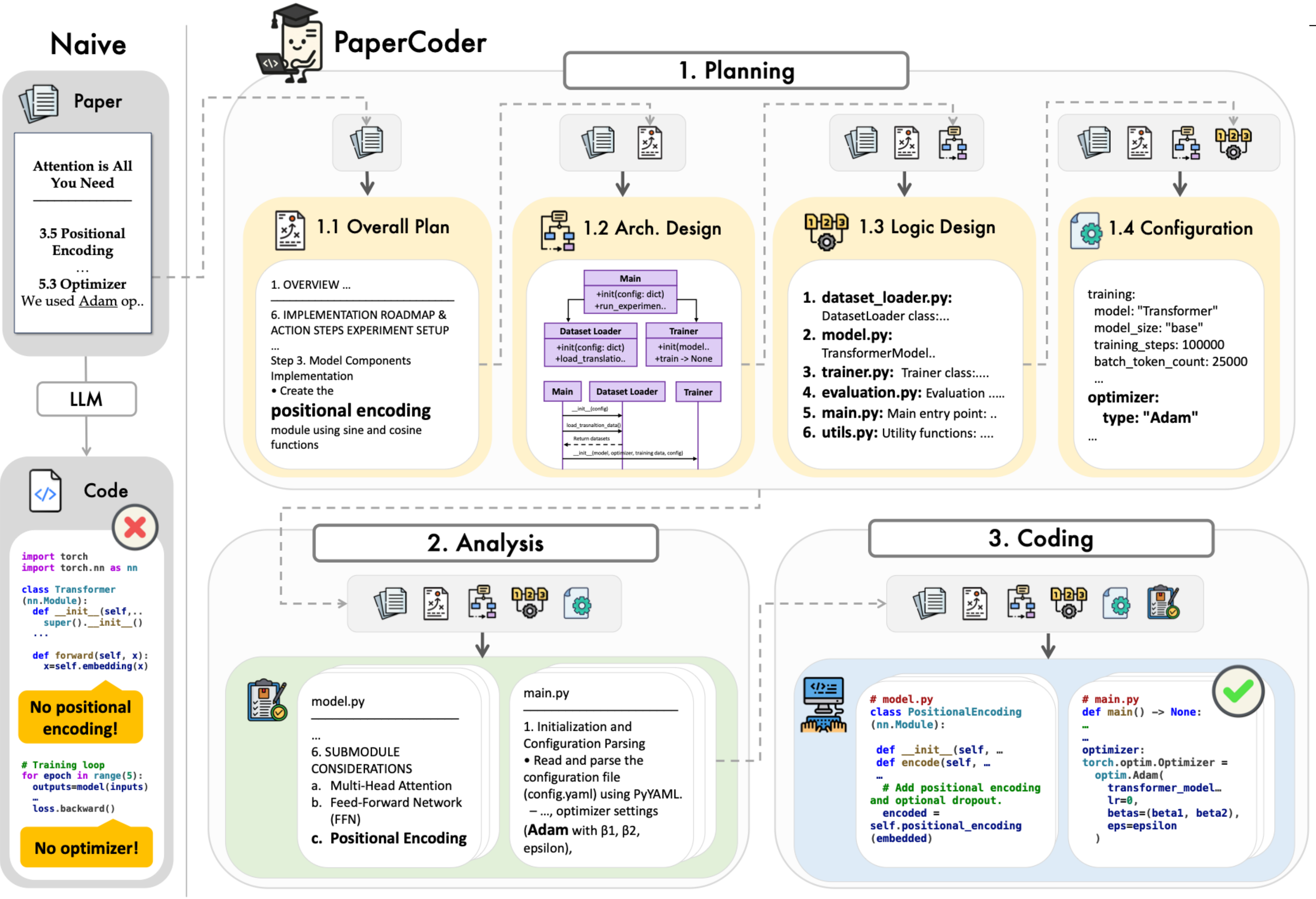
Some useful tools

(For you?)

Paper2Code

Replicates code from a(n ML) paper autonomously.





Attention is All You Need

3.5 Positional Encoding

…

5.3 Optimizer

We used Adam op..

```
import torch
import torch.nn as nn

class Transformer
(nn.Module):
  def __init__(self,..
    super().__init__()
  ...

  def forward(self, x):
    x=self.embedding(x)
```

No positional encoding!

```
# Training loop
for epoch in range(5):
  outputs=model(inputs)
  …
  loss.backward()
```

No optimizer!

1. OVERVIEW …

6. IMPLEMENTATION ROADMAP & ACTION STEPS EXPERIMENT SETUP

…

Step 3. Model Components Implementation

- Create the **positional encoding** module using sine and cosine functions

Main
+init(config: dict) +run_experimen..

Dataset Loader
+init(config: dict) +load_translatio..

Trainer
+init(model.. +train -> None

1. **dataset_loader.py:** DatasetLoader class:...
2. **model.py:** TransformerModel..
3. **trainer.py:** Trainer class:....
4. **evaluation.py:** Evaluation .....
5. **main.py:** Main entry point: ..
6. **utils.py:** Utility functions: ....

```
training:
  model: "Transformer"
  model_size: "base"
  training_steps: 100000
  batch_token_count: 25000
  ...
optimizer:
  type: "Adam"
...
```

model.py

…

6. SUBMODULE CONSIDERATIONS
   a. Multi-Head Attention
   b. Feed-Forward Network (FFN)
   c. **Positional Encoding**

main.py

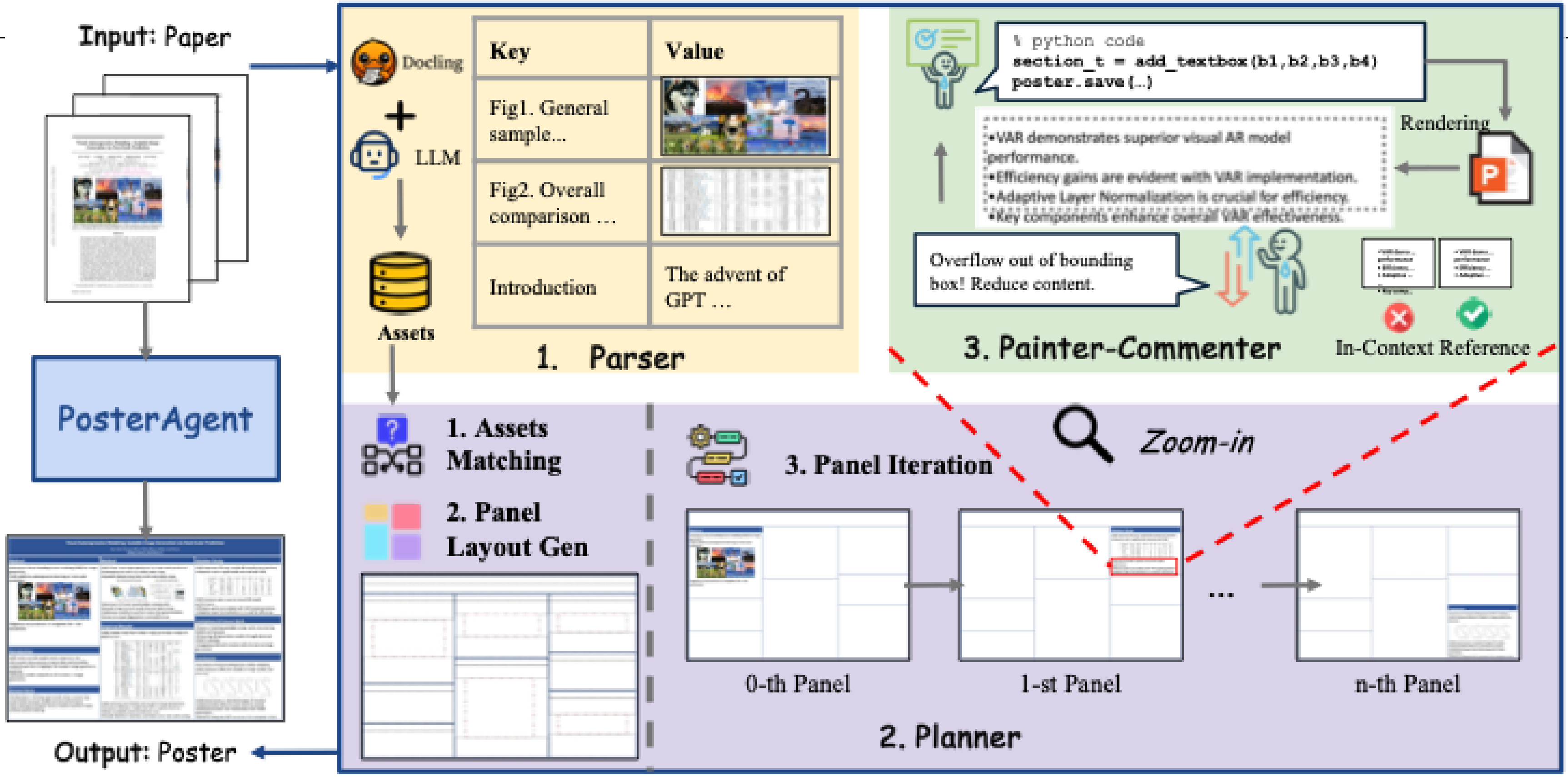
1. Initialization and Configuration Parsing

- Read and parse the configuration file (config.yaml) using PyYAML.
  – …, optimizer settings (**Adam** with β1, β2, epsilon),

```
# model.py
class PositionalEncoding
(nn.Module):

 def __init__(self, …
 def encode(self, …
 …
  # Add positional encoding
and optional dropout.
  encoded =
self.positional_encoding
(embedded)
```

```
# main.py
def main() -> None:
…
…
optimizer:
torch.optim.Optimizer =
  optim.Adam(
    transformer_model…
    lr=0,
    betas=(beta1, beta2),
    eps=epsilon
  )
```



```
% python code
section_t = add_textbox(b1,b2,b3,b4)
poster.save(...)
```

•VAR demonstrates superior visual AR model performance.
•Efficiency gains are evident with VAR implementation.
•Adaptive Layer Normalization is crucial for efficiency.
•Key components enhance overall VAR effectiveness.

Overflow out of bounding box! Reduce content.

Paper2Poster

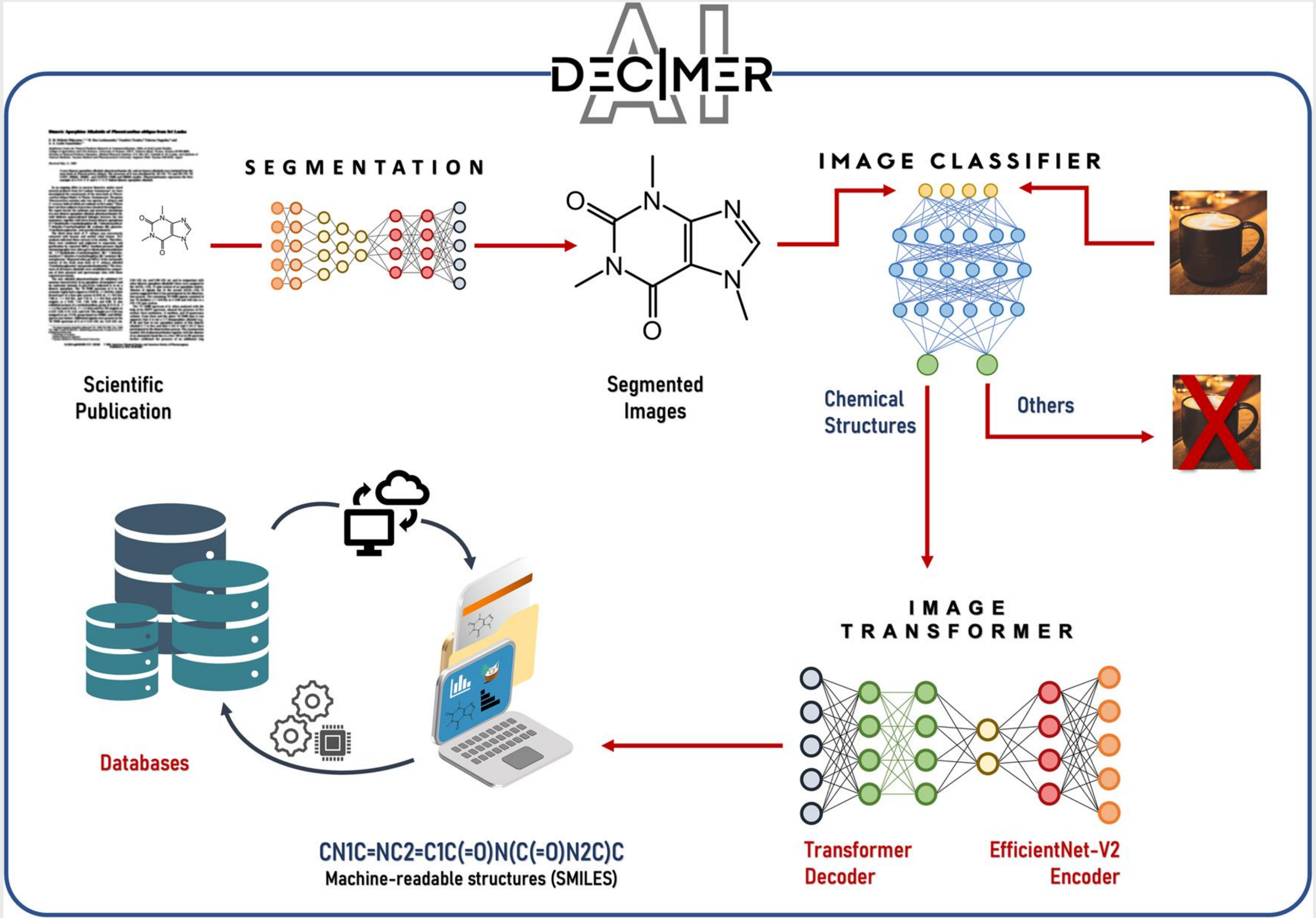
Converts papers to poster autonomously (and efficiently).

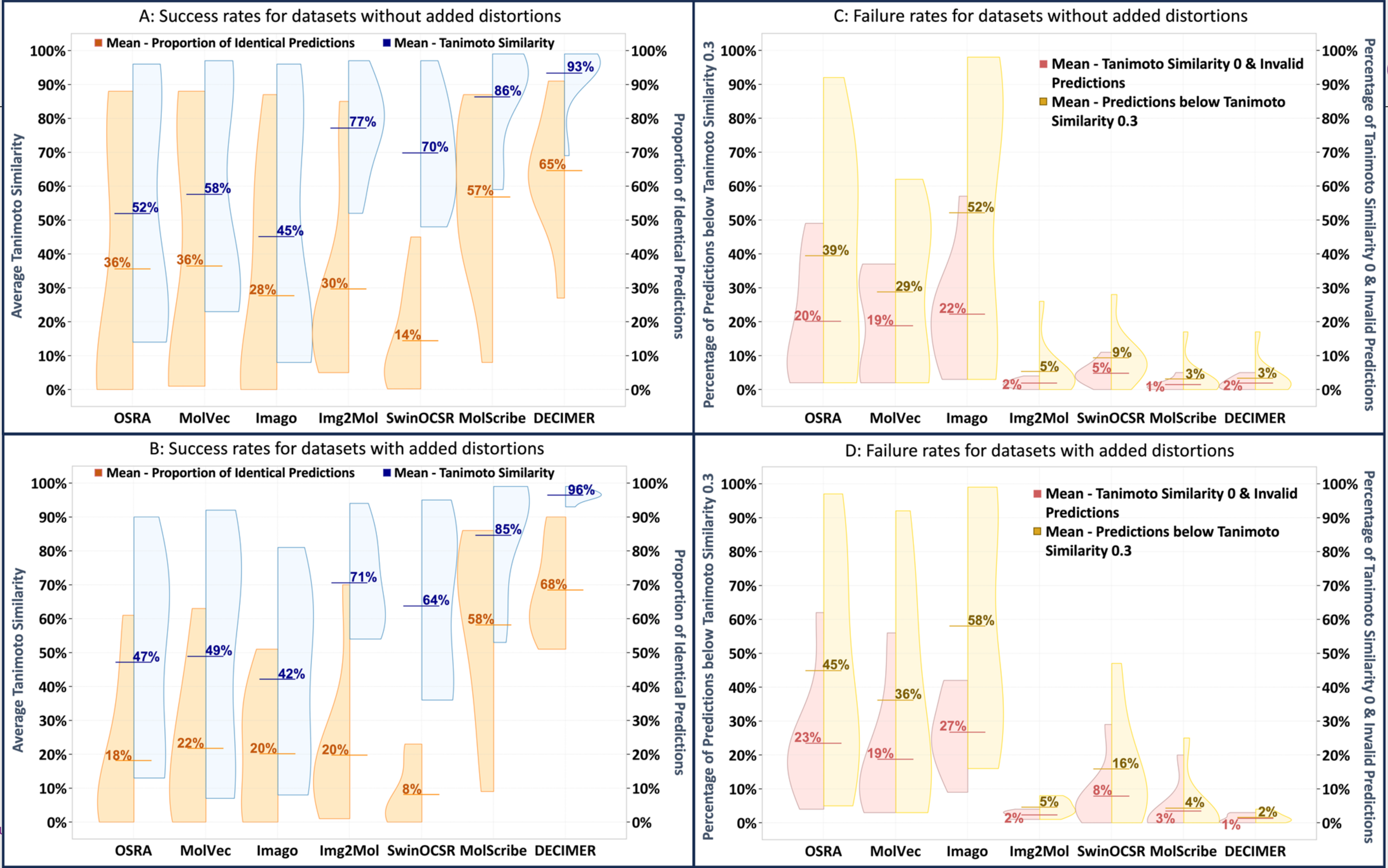
Identifying molecules from images

Motivation

- As mentioned in a previous journal club presentation, a large amount of chemical knowledge is documented on paper or scanned images of them.
- Efficiently digitizing them would help models absorb their knowledge.

Decimer One out of many such software





A: Success rates for datasets without added distortions

B: Success rates for datasets with added distortions

C: Failure rates for datasets without added distortions

D: Failure rates for datasets with added distortions

Thank You

30 June 2025