

Large Language Models for enzymatic reaction prediction and characterization

Lorenzo Di Fruscia* and Jana M. Weber

Dept. of Intelligent Systems, Technical University of Delft, Delft, The Netherlands

*E-mail: l.difruscia@tudelft.nl

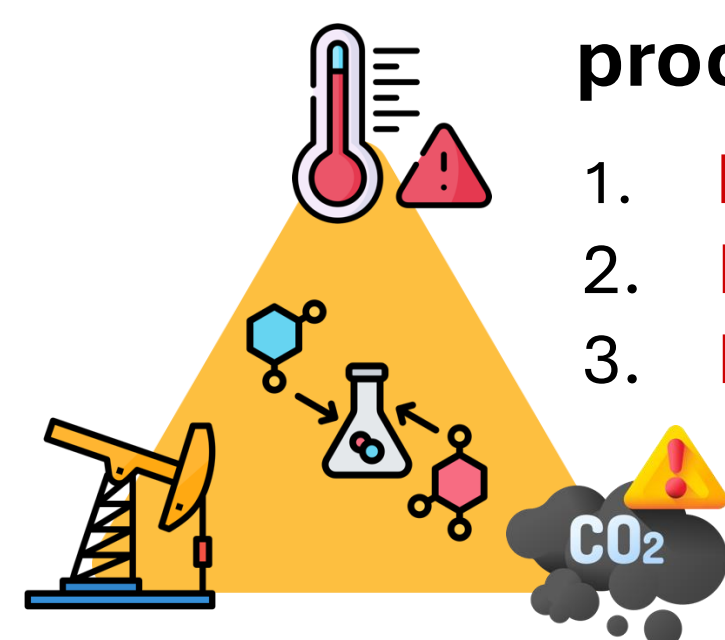


Introduction: Design of sustainable chemical processes

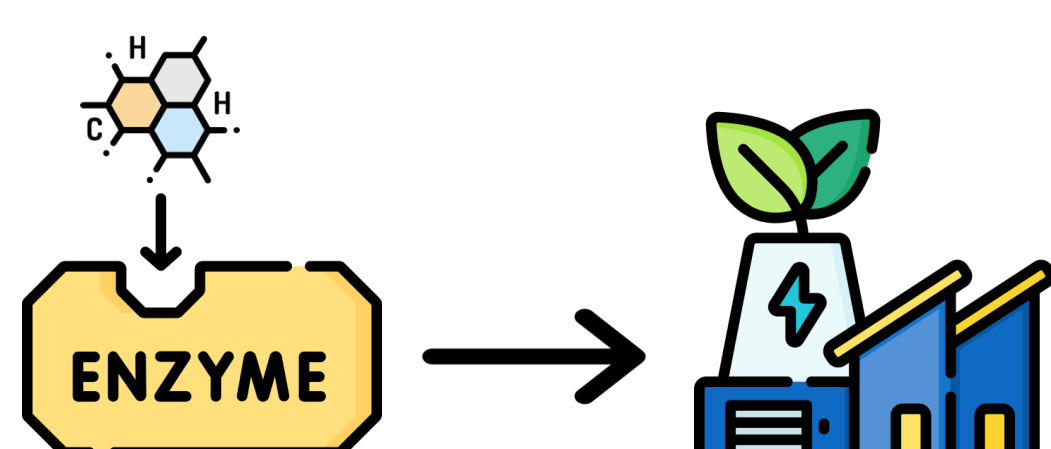
Need for sustainable processes^[1]

Standard chemical processes entail:

1. **High** temperatures
2. **High** emissions
3. **Fossil**-based

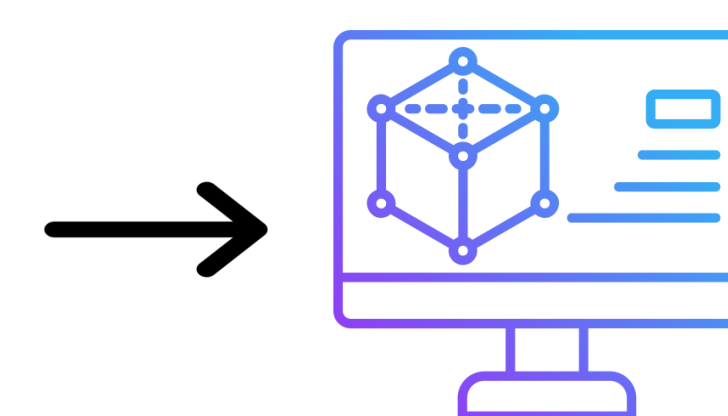


The use of enzymes can address these drawbacks



AI-assisted reaction design

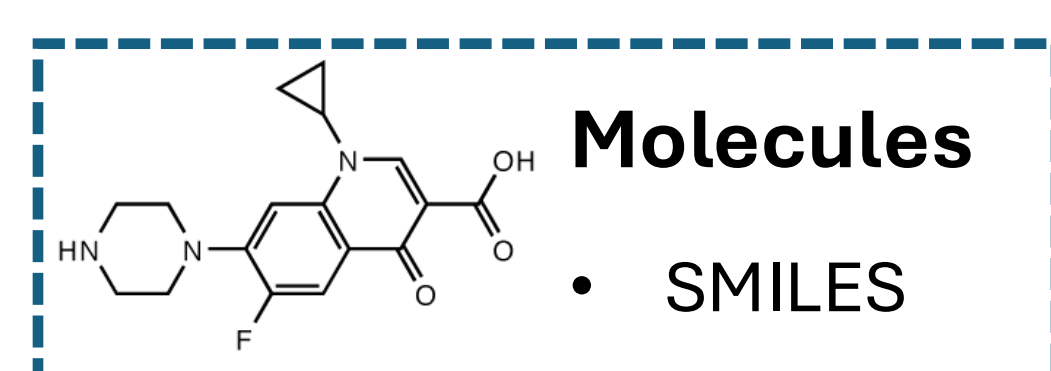
Substrates
Reactants
Enzymes
Products
...



Discovery of new **sustainable** pathways?

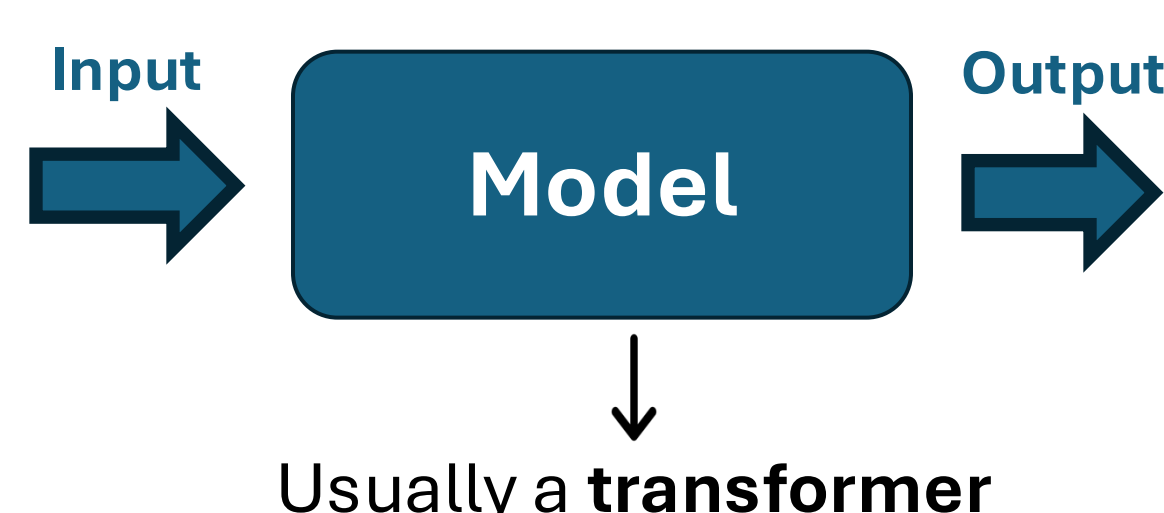
Specialized Machine Learning models^[2,3]

(Some) data representation and tasks:



Enzymes

- Text description
- Enzyme Commission (EC) number



Prediction of:
Substrate
EC number
Product

Large Language Models (LLMs)

Large* models trained on large* datasets
(+ best engineering practices)

Emergent capabilities

Knowledge retrieval
Learning from examples
Basic reasoning
...

Impact in chemistry:

Compound captioning
Property prediction
De-novo design
...

* Hundreds of billions of tokens and parameters



What is the potential of fine-tuned LLMs for biochemical tasks?

Methods and results

Our approach:

Dataset: Brenda^[5]

Models: Llama-3.1 family

- Strategies:**
- Fine-Tuning (FT)
 - In-Context Learning (ICL)

Our tasks:

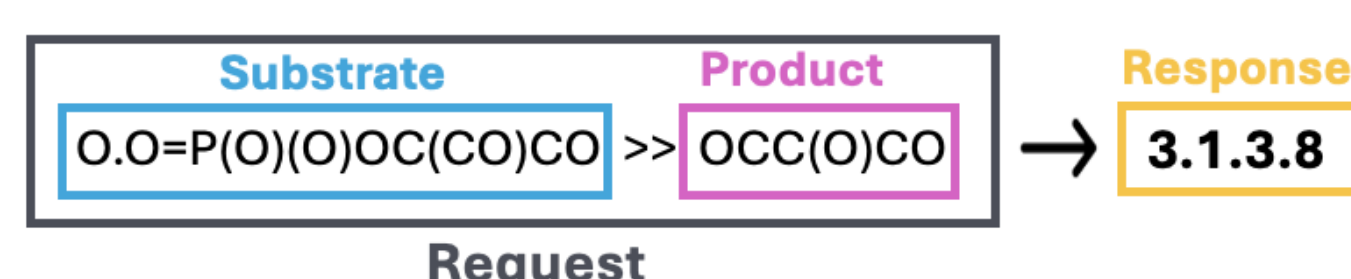
- EC number classification (EC)
- Product prediction (Forward synthesis, FS)
- Substrate prediction (Retrosynthesis, RS)

We formulate natural language questions from the dataset
"Given substrate **S** and product **P** in SMILES notation, provide the EC number!"

System: You are an AI assistant with expertise in biochemistry. When provided with [TASK] you should be able to [OBJECTIVE]. Please strictly stick to [RULES] and provide [RESPONSE] to [REQUEST].

User: [REQUEST]

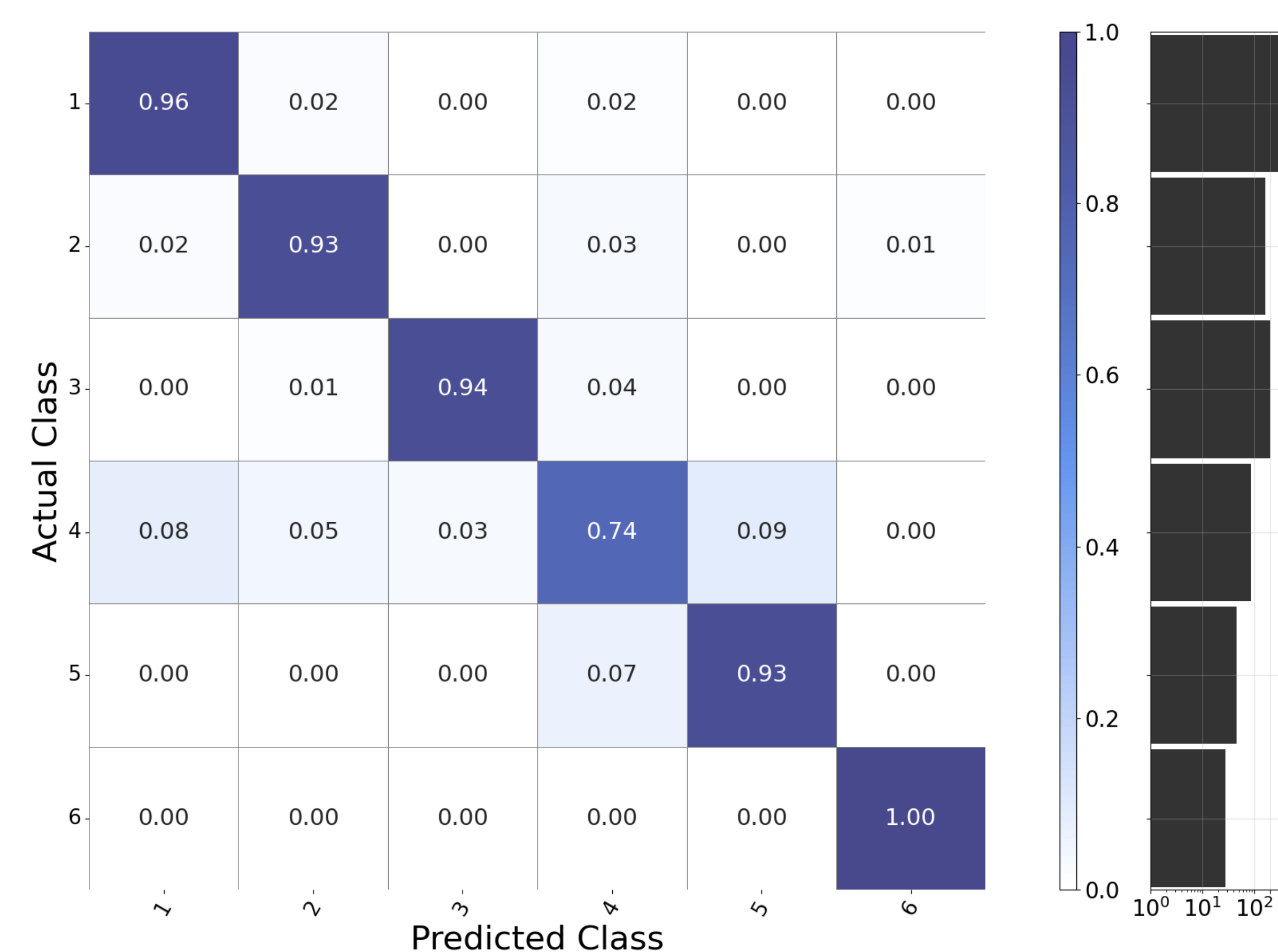
Assistant: [RESPONSE]



We fine-tune LLMs using Low-Rank Adaptation (LoRA)^[6] on different scenarios:

1. **Single-task (ST):** One model is fine-tuned per task
2. **Multitask (MT):** One model is fine-tuned for all tasks
3. **Low-data regime:** What if data availability is low?
4. **Generalization:** Does FT degrade general performance?

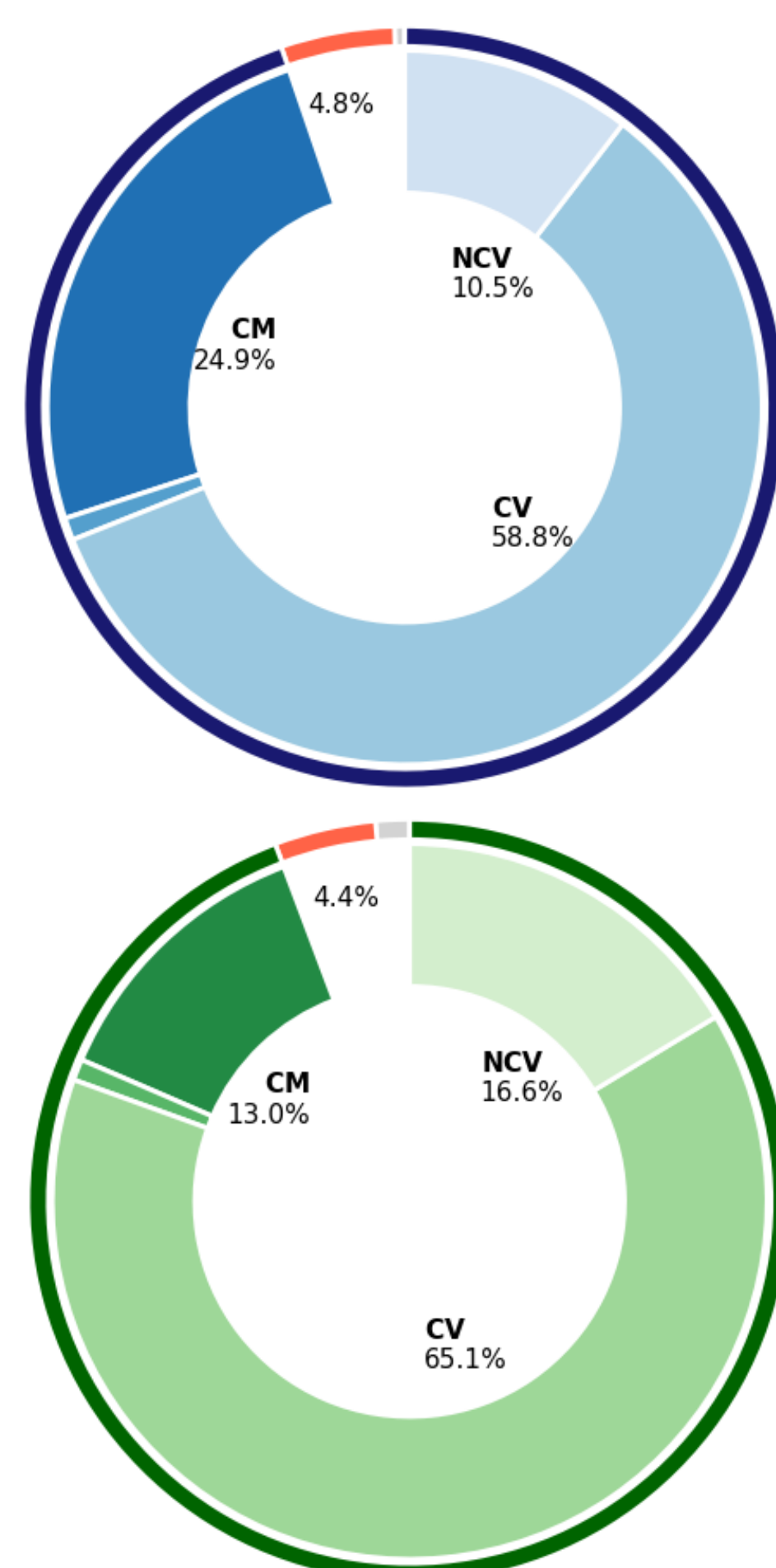
1 EC class prediction accuracy



ICL baseline (%):

EC class	FS Match	RS Match
29.6	<0.1	<0.1

FS and RS prediction



Legend:
Wrong generation (grey)
Invalid (red)
CM: Canonical Match
NCM: Non-Canonical Match
CV: Canonical Valid
NCV: Non-Canonical Valid

2 MT performance improvement

Task	Metric (%)	Llama-3.1 70B		
		ST	MT	Δ
EC	Accuracy EC1 ↑	91.7	86.4	-5.3
	Accuracy EC2 ↑	61.7	65.1	+3.7
	Accuracy EC3 ↑	49.2	48.9	-0.3
FS	Match ↑	25.9	33.8	+7.9
	Match+(TS=1) ↑	33.0	44.4	+11.4
	Match+(TS>0.95) ↑	34.2	45.4	+11.2
RS	Invalid ↓	4.8	4.9	+0.1
	Match ↑	13.9	19.2	+5.3
	Match+(TS=1) ↑	20.6	30.1	+9.5
	Match+(TS>0.95) ↑	36.1	45.4	+9.3
	Invalid ↓	4.4	3.0	-1.4

Tanimoto score **TS** ∈ [0,1] is used to assess structural similarity of predicted molecules

Conclusions

1. Fine-tuning greatly improves performance over in-context learning (ICL)
2. MT enhances performance across all tasks compared to ST scenarios
3. Fine-tuned LLMs have potential in low-data regimes (< 1000 samples)
4. ST fine-tuning improves performance on unseen related tasks

References

- [1] Paper Sheldon, Roger Arthur. "Green Chemistry and Biocatalysis: Engineering a Sustainable Future." *Catalysis Today* 431 (2024)
- [2] Probst, D., Manica, M., Teukam, Y. G. N., Castrogiovanni, A., Paratore, F., and Laino, T. "Biocatalysed Synthesis Planning Using Data-Driven Learning." *Nature Communications* 13, no. 1 (2022)
- [3] Qian et al. "A General Model for Predicting Enzyme Functions Based on Enzymatic Reactions." *Journal of Cheminformatics* 16, no. 1 (2024)
- [4] Brown et al. "Language Models Are Few-Shot Learners." *arXiv*, (2020)
- [5] Antje C. Dietmar S. Ida, S. Brenda, enzyme data and metabolic information—pubmed. *Nucleic Acids Res.* 30,47–49, 2002.
- [6] E. J. Hu et al. "Low-rank adaptation of large language models." *arXiv*, (2021)

