Chemical Data Is Context-Dependent

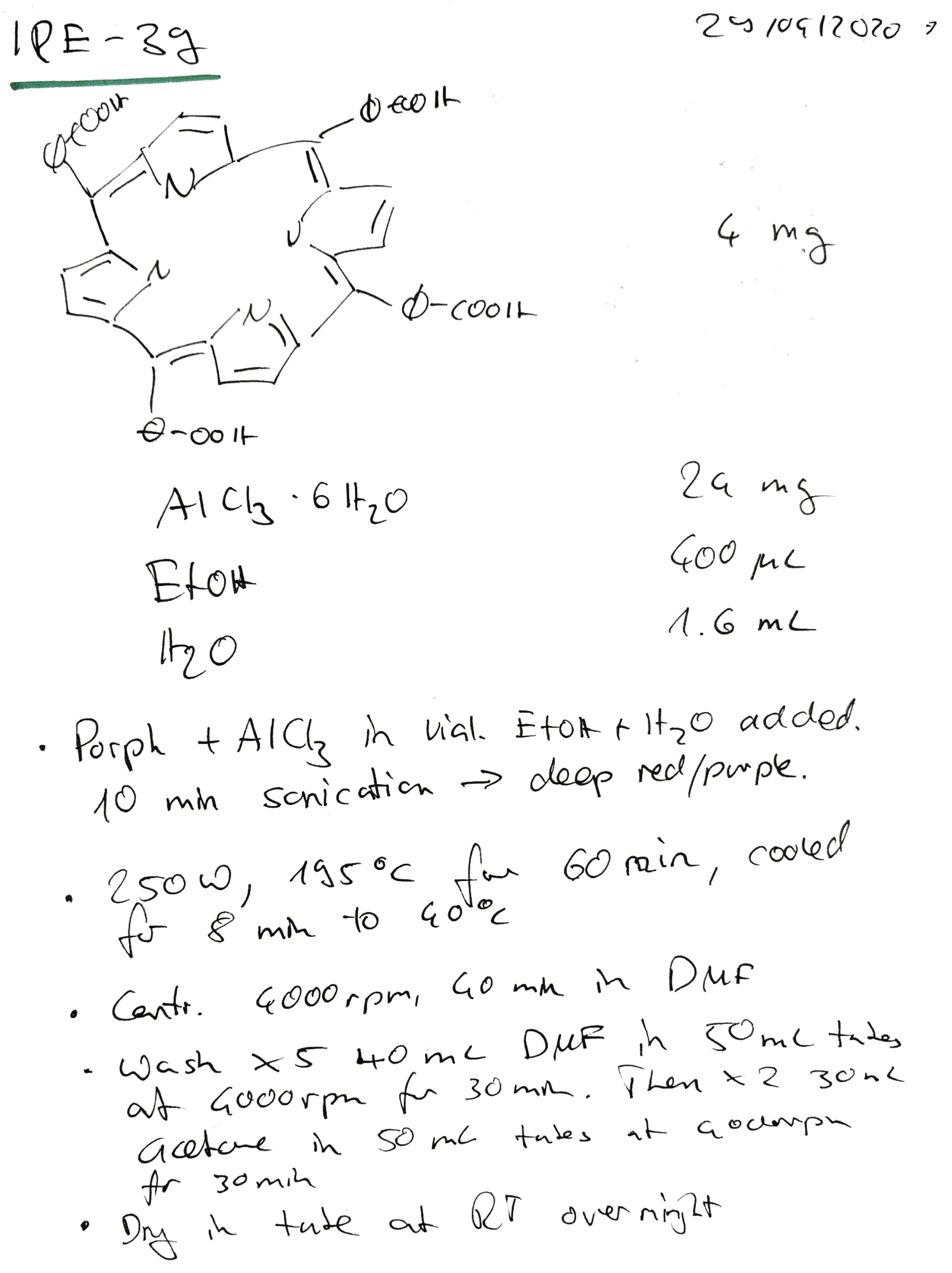
Processing history of materials, synthesis procedure, ….

Too flexible for tabular data

Nat. Chem. **2022**, *14* (4), 365–376.

Much of Chemical Data Is in Fuzzy Form

a



IPE-3g

29/09/2020

ϕ-COOH

ϕ-COOH

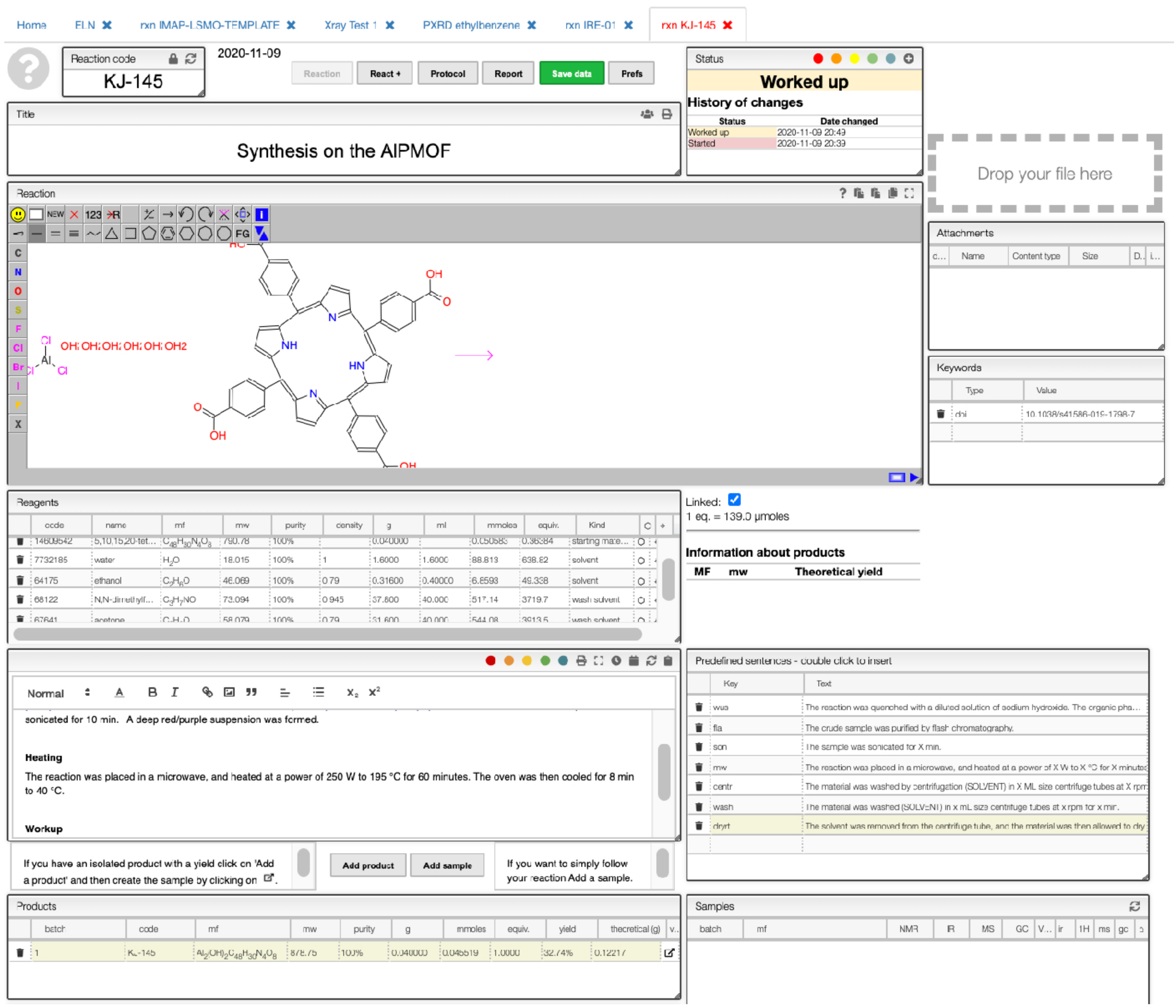
ϕ-COOH

ϕ-COOH

	4 mg
$AlCl_3 \cdot 6H_2O$	24 mg
EtOH	400 μL
H_2O	1.6 mL

- Porph + $AlCl_3$ in vial. EtOH + H_2O added. 10 min sonication → deep red/purple.
- 250 W, 195 °C for 60 min, cooled for 8 min to 40 °C
- Centr. 4000 rpm, 40 min in DMF
- Wash ×5 40 mL DMF in 50 mL tubes at 4000 rpm for 30 min. Then ×2 30 mL acetone in 50 mL tubes at 4000 rpm for 30 min
- Dry in tube at RT overnight

b



Reaction code: KJ-145

2020-11-09

Status: Worked up

History of changes

Status	Date changed
Worked up	2020-11-09 20:49
Started	2020-11-09 20:39

Title: Synthesis on the AlPMOF

Reagents

code	name	mf	mw	purity	density	g	ml	mmoles	equiv.	Kind
14609542	5,10,15,20-tet...	$C_{48}H_{30}N_4O_8$	790.78	100%		0.040000		0.050585	0.36384	starting mate...
7732185	water	H_2O	18.015	100%	1	1.6000	1.6000	88.813	638.82	solvent
64175	ethanol	C_2H_6O	46.069	100%	0.79	0.31600	0.40000	6.8593	49.338	solvent
68122	N,N-dimethylf...	C_3H_7NO	73.094	100%	0.945	37.800	40.000	517.14	3719.7	wash solvent
67641	acetone	C_3H_6O	58.079	100%	0.79	31.600	40.000	544.08	3913.5	wash solvent

Linked: ☑
1 eq. = 139.0 μmoles

Information about products

MF	mw	Theoretical yield

sonicated for 10 min. A deep red/purple suspension was formed.

Heating

The reaction was placed in a microwave, and heated at a power of 250 W to 195 °C for 60 minutes. The oven was then cooled for 8 min to 40 °C.

Workup

Predefined sentences - double click to insert

Key	Text
wua	The reaction was quenched with a diluted solution of sodium hydroxide. The organic pha...
fla	The crude sample was purified by flash chromatography.
son	The sample was sonicated for X min.
mw	The reaction was placed in a microwave, and heated at a power of X W to X °C for X minutes
centr	The material was washed by centrifugation (SOLVENT) in X ML size centrifuge tubes at X rpm
wash	The material was washed (SOLVENT) in x mL size centrifuge tubes at x rpm for x min.
dryrt	The solvent was removed from the centrifuge tube, and the material was then allowed to dry

If you have an isolated product with a yield click on 'Add a product' and then create the sample by clicking on .

If you want to simply follow your reaction Add a sample.

Products

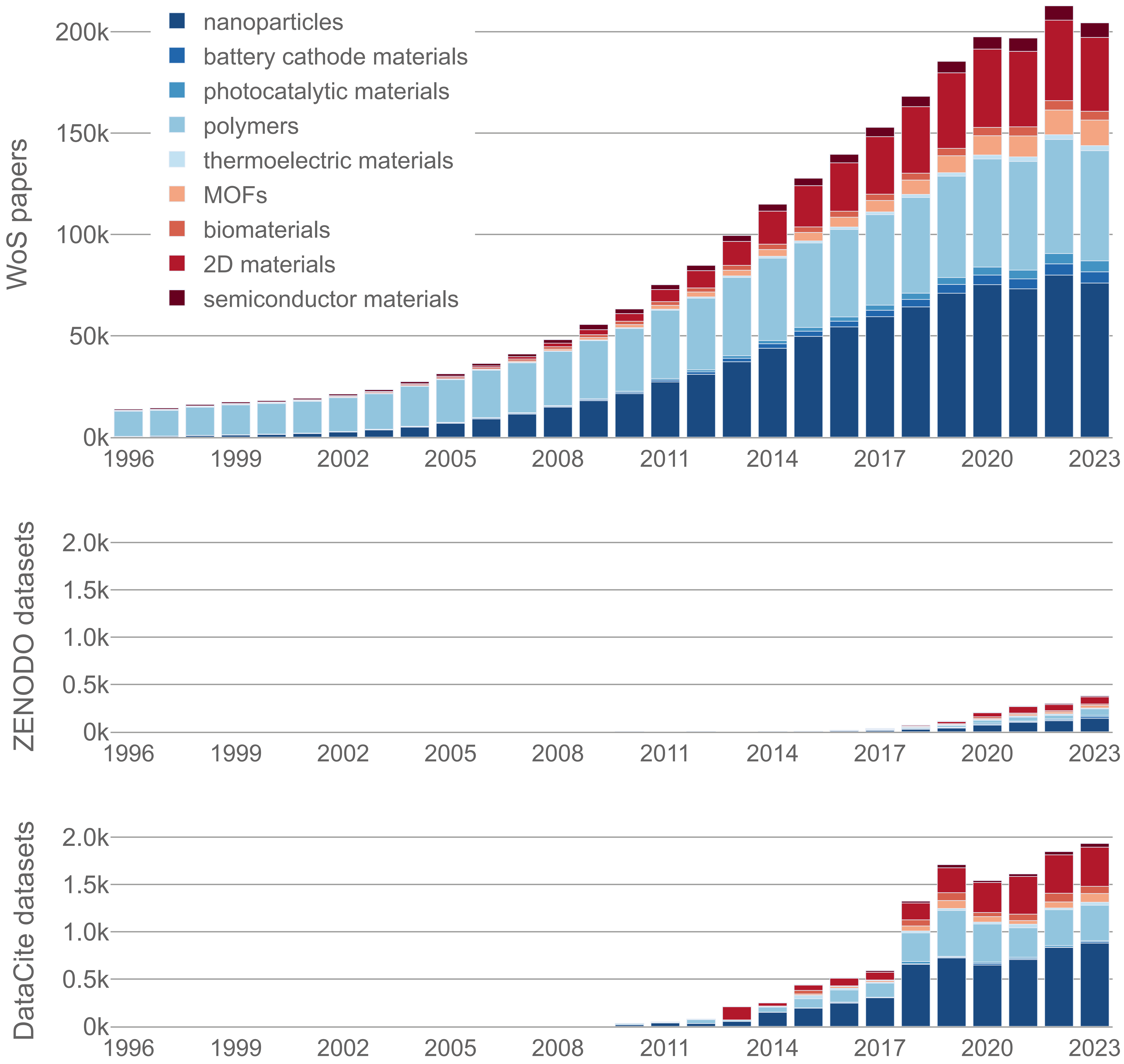
batch	code	mf	mw	purity	g	mmoles	equiv.	yield	theoretical (g)
1	KJ-145	$Al_2(OH)_2C_{48}H_{30}N_4O_8$	878.75	100%	0.040000	0.045519	1.0000	32.74%	0.12217

Samples

batch	mf	NMR	IR	MS	GC

Nat. Chem. **2022**, *14* (4), 365–376.
Images: E. PETERSEN/SCIENCE, University of Cambridge

We Need To Care About This Fuzzy Data



Matextract.pub and arXiv:2407.16867

Finding the Best Pancake Recipe in the Conventional Way

Correlating descriptors with ratings

Machine learning models can correlate descriptors of our recipes to ratings

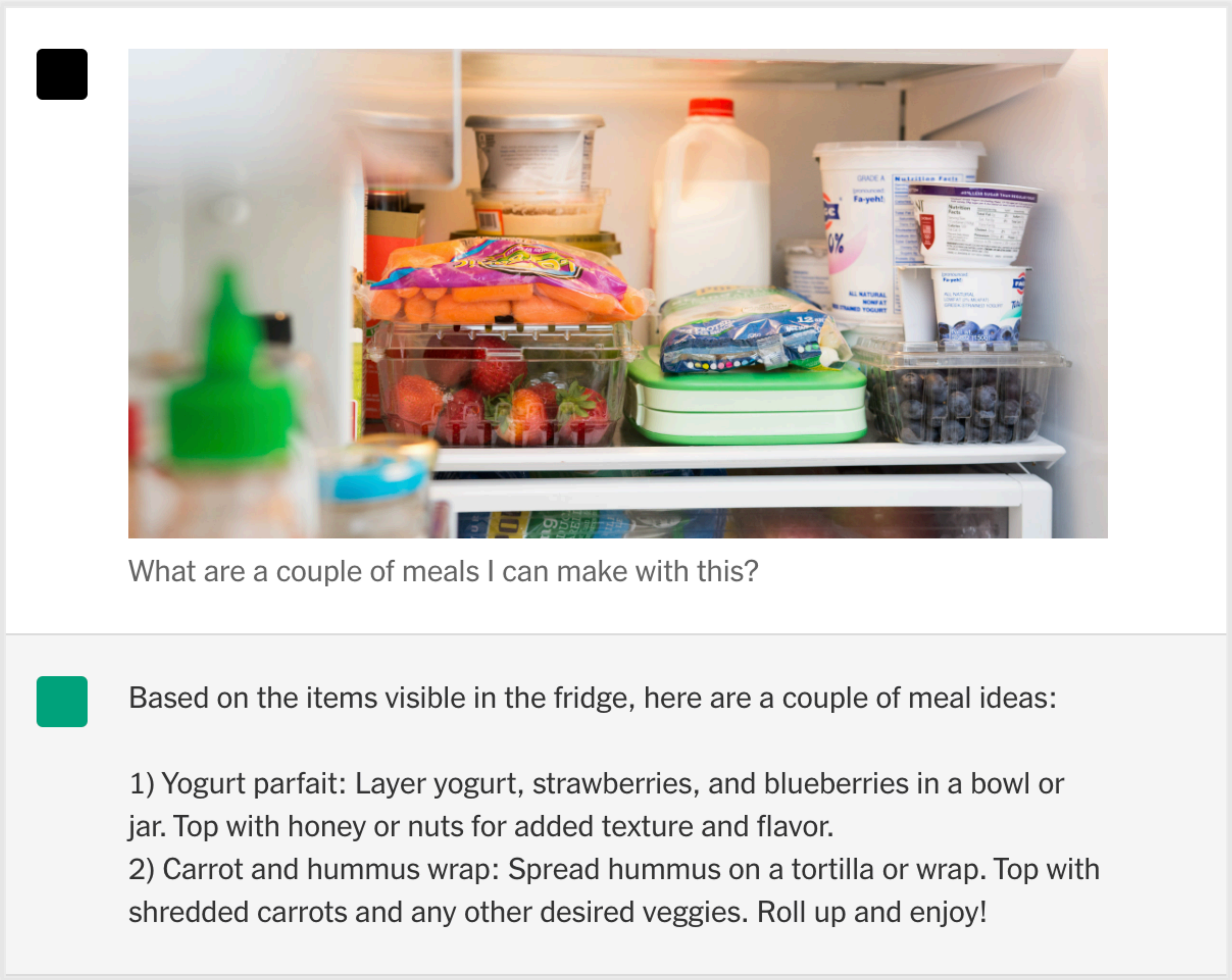
In this way, we could score new recipes

Time ⏰	*Banana* 🍌	*Milk* 🥛	*Egg* 🥚	*Rating* 🤩
30	1	100	0	4
10	0	129	0,5	5
5	2	140	1	3,4
12	2	80	2	2

Making pancakes: *Generalized Learning on Scale*

Large-language models can answer natural language queries.

It can do *without being explicitly trained to do so.*

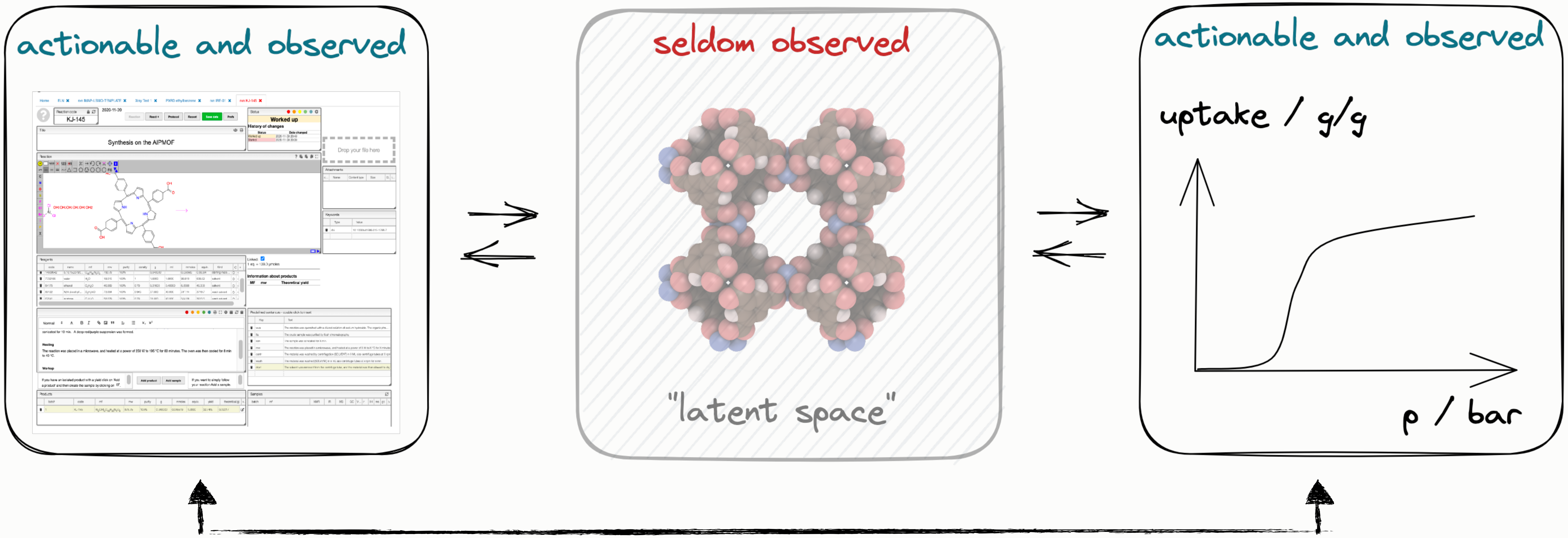


What are a couple of meals I can make with this?

Based on the items visible in the fridge, here are a couple of meal ideas:

1) Yogurt parfait: Layer yogurt, strawberries, and blueberries in a bowl or jar. Top with honey or nuts for added texture and flavor.
2) Carrot and hummus wrap: Spread hummus on a tortilla or wrap. Top with shredded carrots and any other desired veggies. Roll up and enjoy!

Actionable Machine Learning Powered By Large Language Models



Reaction code
KJ-145

Title
Synthesis on the AlPMOF

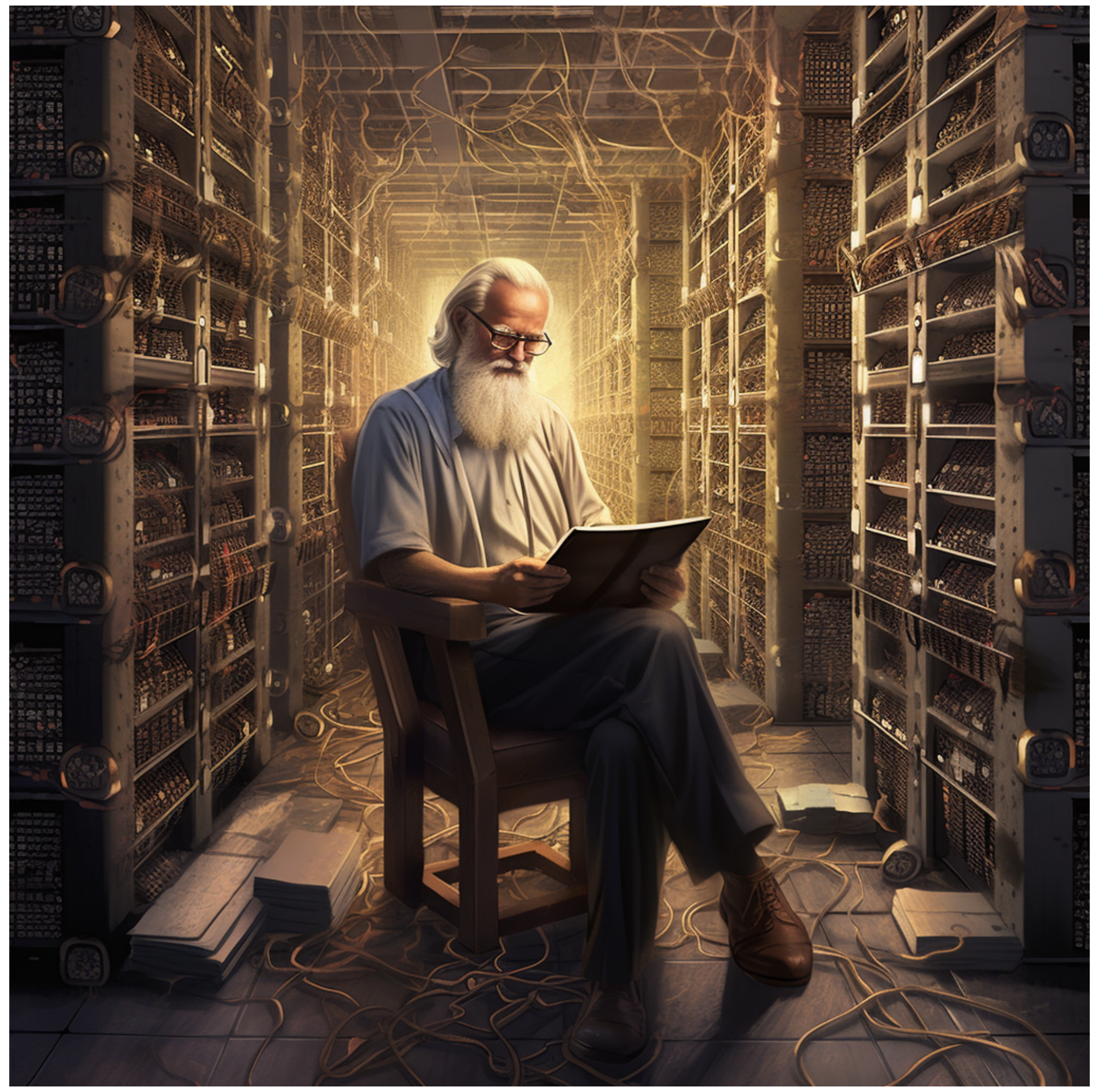
Status
Worked up
History of changes

“Conventional” ML works on data we have little direct control over.

Transforming Chemistry With Transformers

Kevin M Jablonka (HIPOLE Jena)
@kmjablonka l mail@kjablonka.com l jablonkagroup.uni-jena.de

LLMs Can Play Many Roles



Storage of fuzzy knowledge
Making experience and knowledge accessible

Director
Orchestrating tools and creating novel interfaces

Curator
Extracting structured data

Digital Discovery **2023**, 2 (5), 1233–1250

Predictions as Text-Completion

1. Tabular dataset with string representation of system

2. Transform into sentences

3. Fine-Tune LLM (e.g. , GPT-3) to complete prompts

4. Query LLM to complete prompt

prompt	*completion*
What is the phase of $Co_1Cu_1Fe_1Ni_1V_1$?###	multi-phase
What is the phase of $Pu_{0.75}Zr_{0.25}$?	single-phase
What is the phase of BeFe?	multi-phase
What is the phase of LiTa?	multi-phase
What is the phase of $Nb_{0.5}Ta_{0.5}$?	single-phase
What is the phase of $Al_{0.1}W_{0.9}$?	single-phase
What is the phase of $Cr_{0.5}Fe_{0.5}$?	single-phase
What is the phase of $Al_1Co_1Cr_1Cu_1Fe_1Ni_1Ti_1$	multi-phase
What is the phase of $Cu_{0.5}Mn_{0.5}$?	single-phase
What is the phase of OsU?@@@	multi-phase

fine-tuned LLM

Nature Machine Intelligence **2024,** 6, 161–169. /w Andres Ortega, Philippe Schwaller, and Berend Smit
First in AI for Accelerated Materials Design NeurIPS 2022 Workshop, 2022 (before ChatGPT)

Outperforms the State-of-the-Art

Domain-specific model (RF)

Tree-based model with hand-tuned features on about 1000 points

npj Computational Materials **2020**, 6 (1).

Automatminer

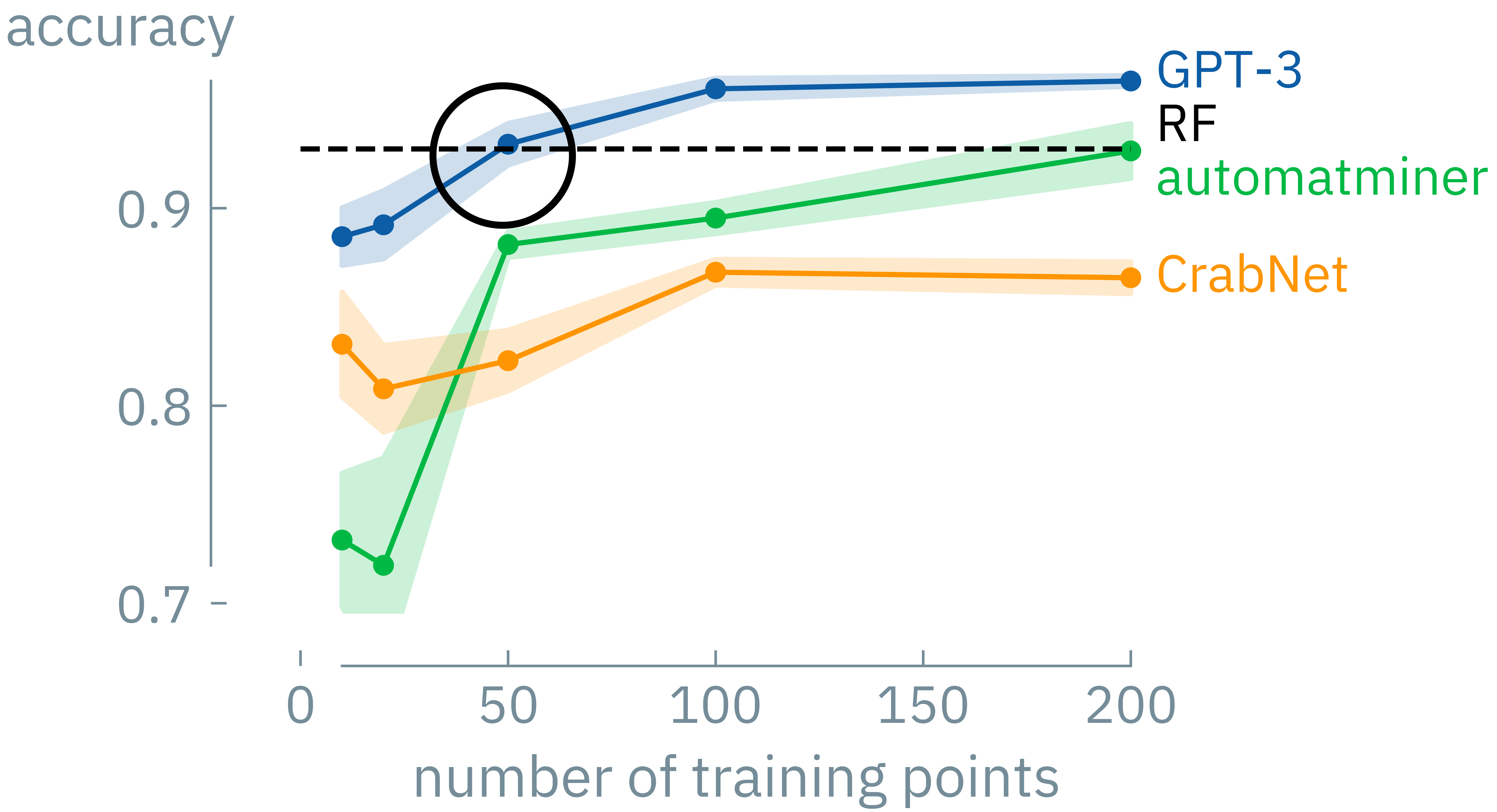
Automatic machine learning that optimizes featurizers and models

npj Computational Materials **2020,** 6 (138).

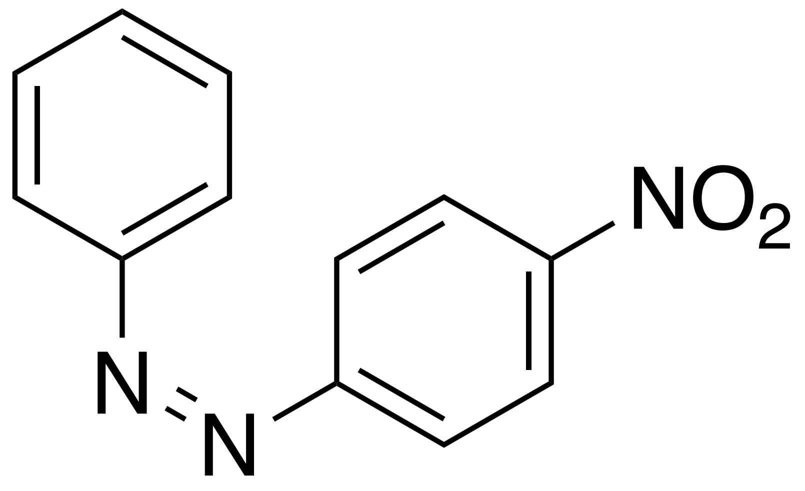
CrabNet

Composition-based transformer model

npj Computational Materials **2021**, 7 (77).



Works for Different Representations



O=[N+](C(C=C1)=CC=C1/N=N\C2=CC=CC=C2)[O-]

SMILES

[O][=N+1][Branch2][Ring1][Branch1][C][Branch1][Ring1][C][=C][=C][C][=C][Ring1][Ring2][/N][=N][\\C][=C][C][=C][C][=C][Ring1][=Branch1][O-1]

SELFIES

InChI=1S/C12H9N3O2/c16-15(17)12-8-6-11(7-9-12)14-13-10-4-2-1-3-5-10/h1-9H/b14-13-

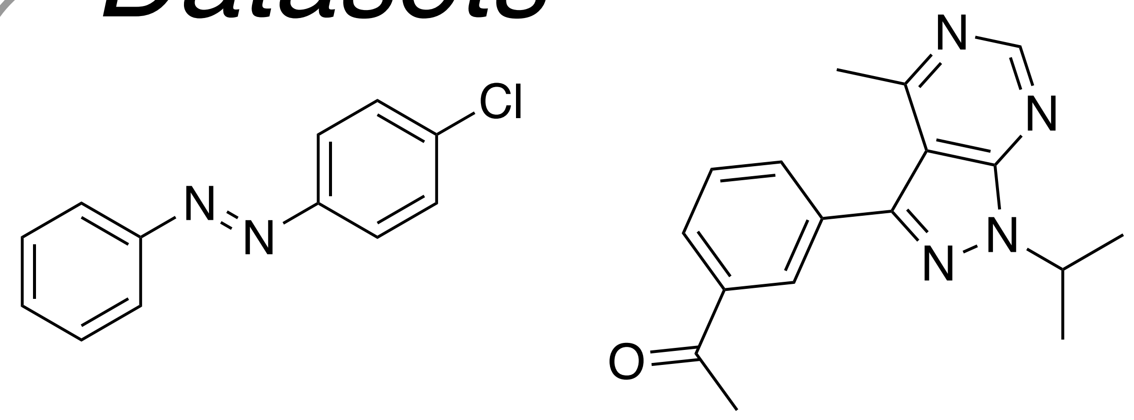
InChI

(*Z*)-1-(4-nitrophenyl)-2-phenyldiazene

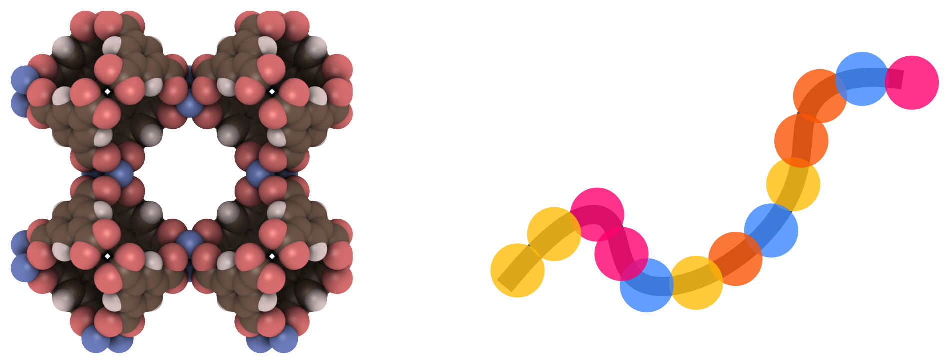
IUPAC name

Across Chemical Space

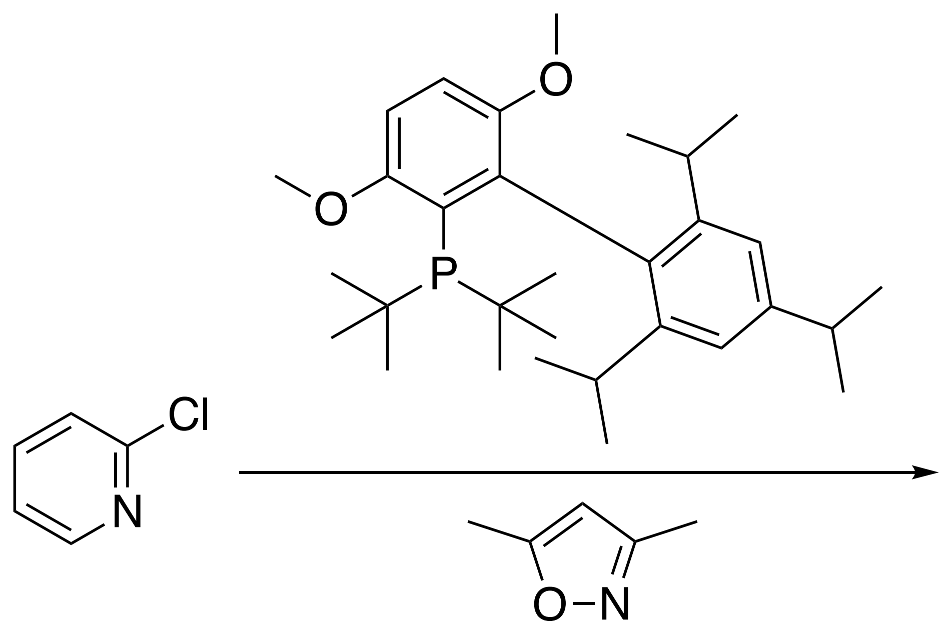
Datasets



Molecules



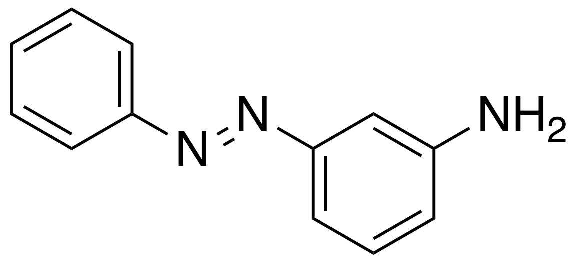
Materials



Reactions

Tasks

“What is the transition wavelength of 2-phenyldiazenylaniline”



↓

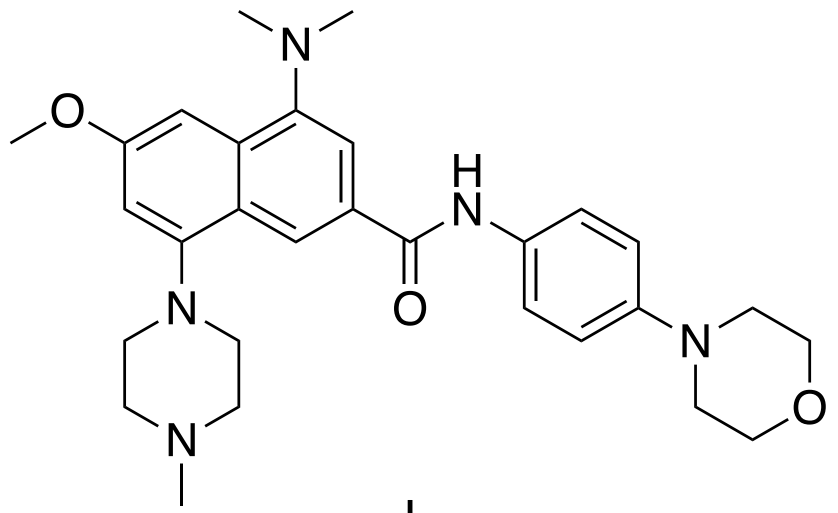
GPT -3

↓

“low”

Classification

“What is the lipophilicity of COc1cc(N2CCN(C)CC2)c3nc(cc(N(C)C)c3c1)C(=O)Nc4ccc(cc4)N5CCOCC5?”



↓

GPT -3

↓

3.3

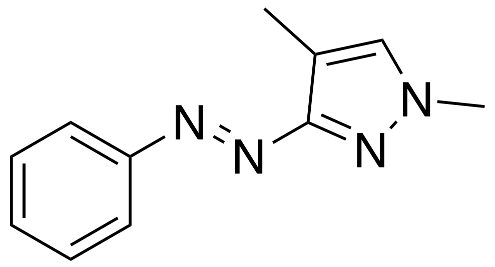
Regression

“What is a molecule with E isomer transition wavelength of 325 nm, Z isomer transition wavelength of 286 nm?”

↓

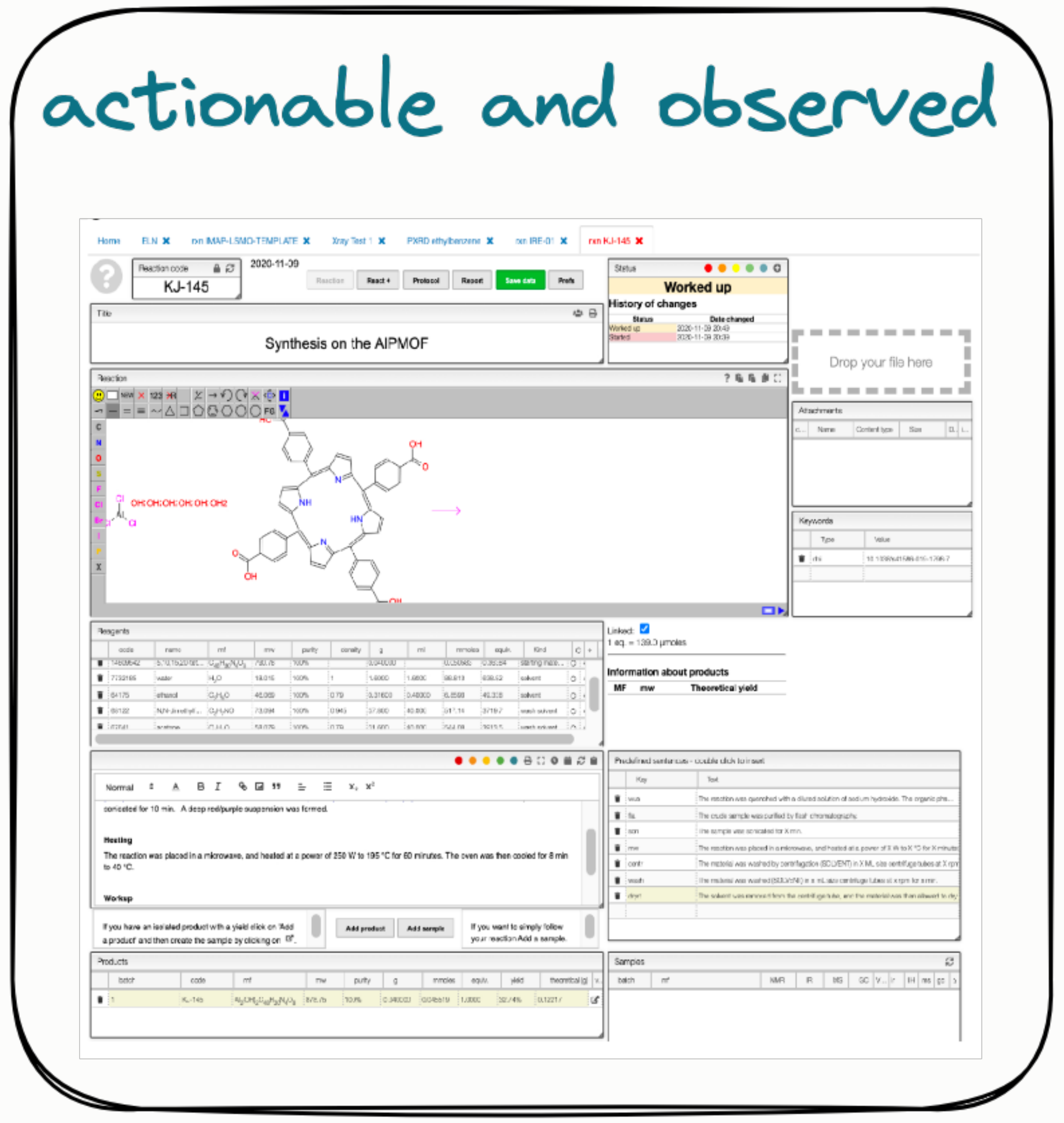
GPT -3

↓



Inverse Design

Actionable Machine Learning Powered By Large Language Models



Reaction code: KJ-145

2020-11-09

Status: Worked up

History of changes

Status	Date changed
Worked up	2020-11-09 20:49
Started	2020-11-09 20:39

Title: Synthesis on the AlPMOF

Linked: ☑

1 eq. = 139.3 µmoles

Information about products

MF	mw	Theoretical yield

sonicated for 10 min. A deep red/purple suspension was formed.

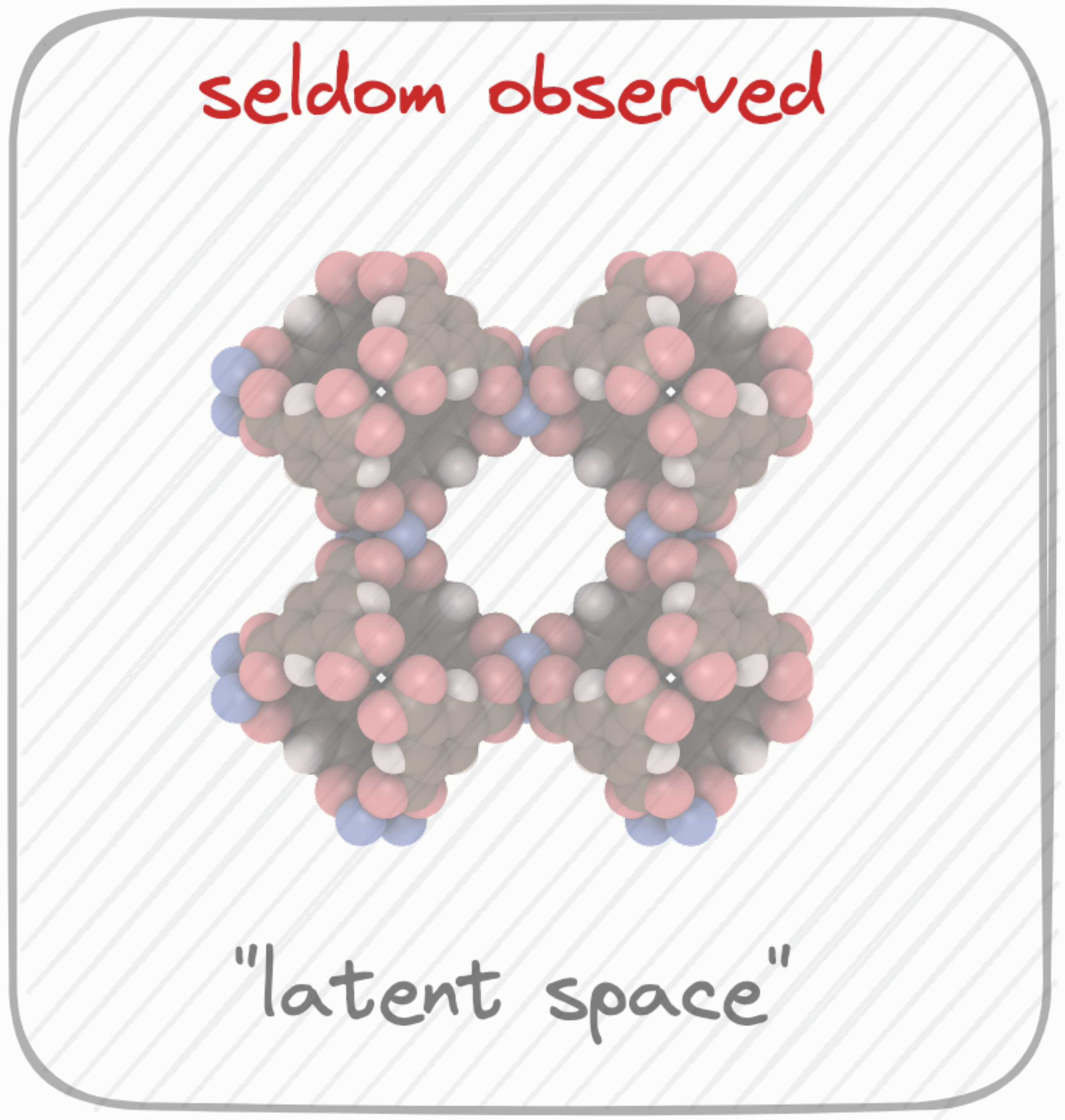
Heating

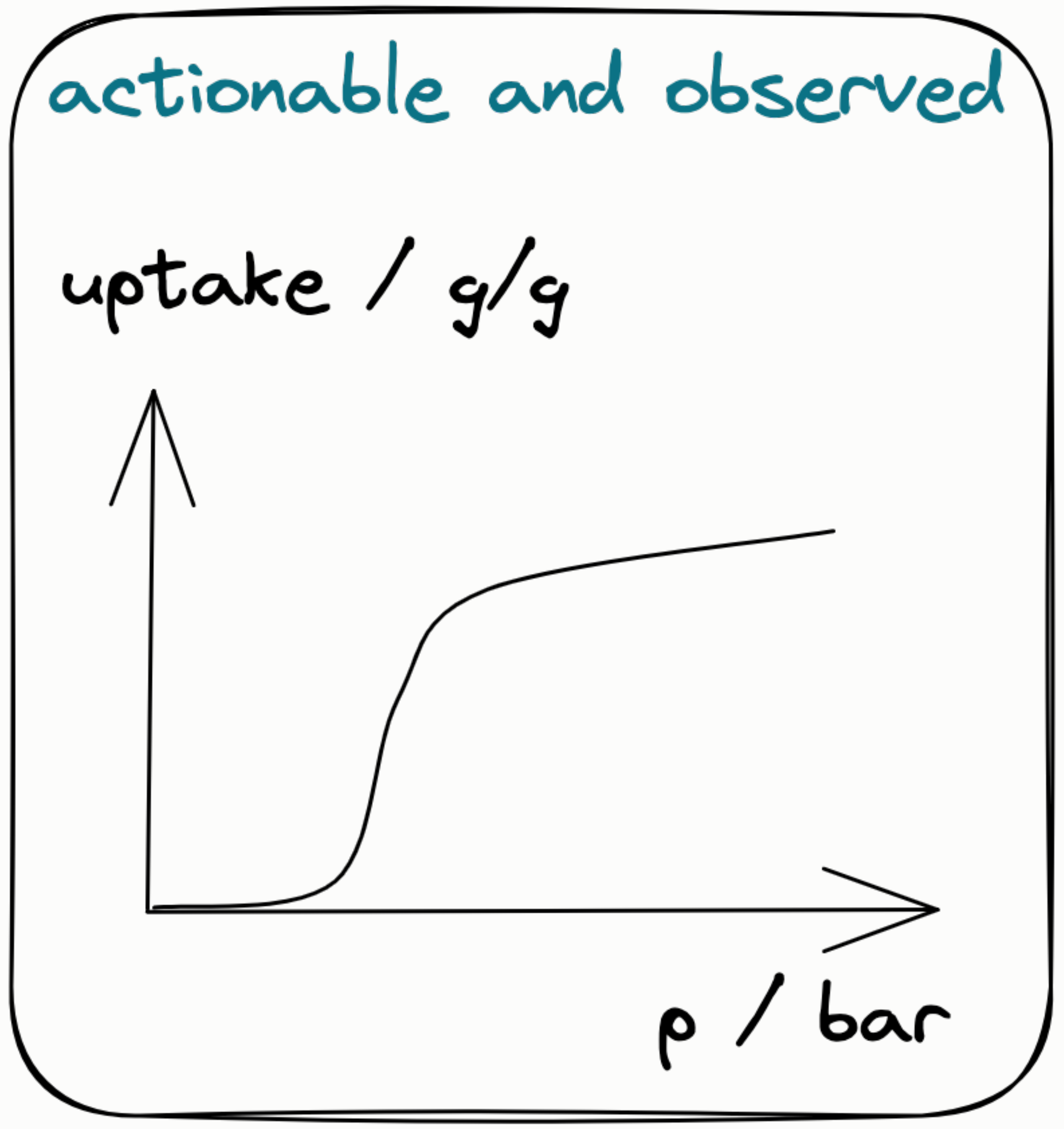
The reaction was placed in a microwave, and heated at a power of 250 W to 195 °C for 60 minutes. The oven was then cooled for 8 min to 40 °C.

Workup

If you have an isolated product with a yield click on 'Add a product' and then create the sample by clicking on .

If you want to simply follow your reaction Add a sample.



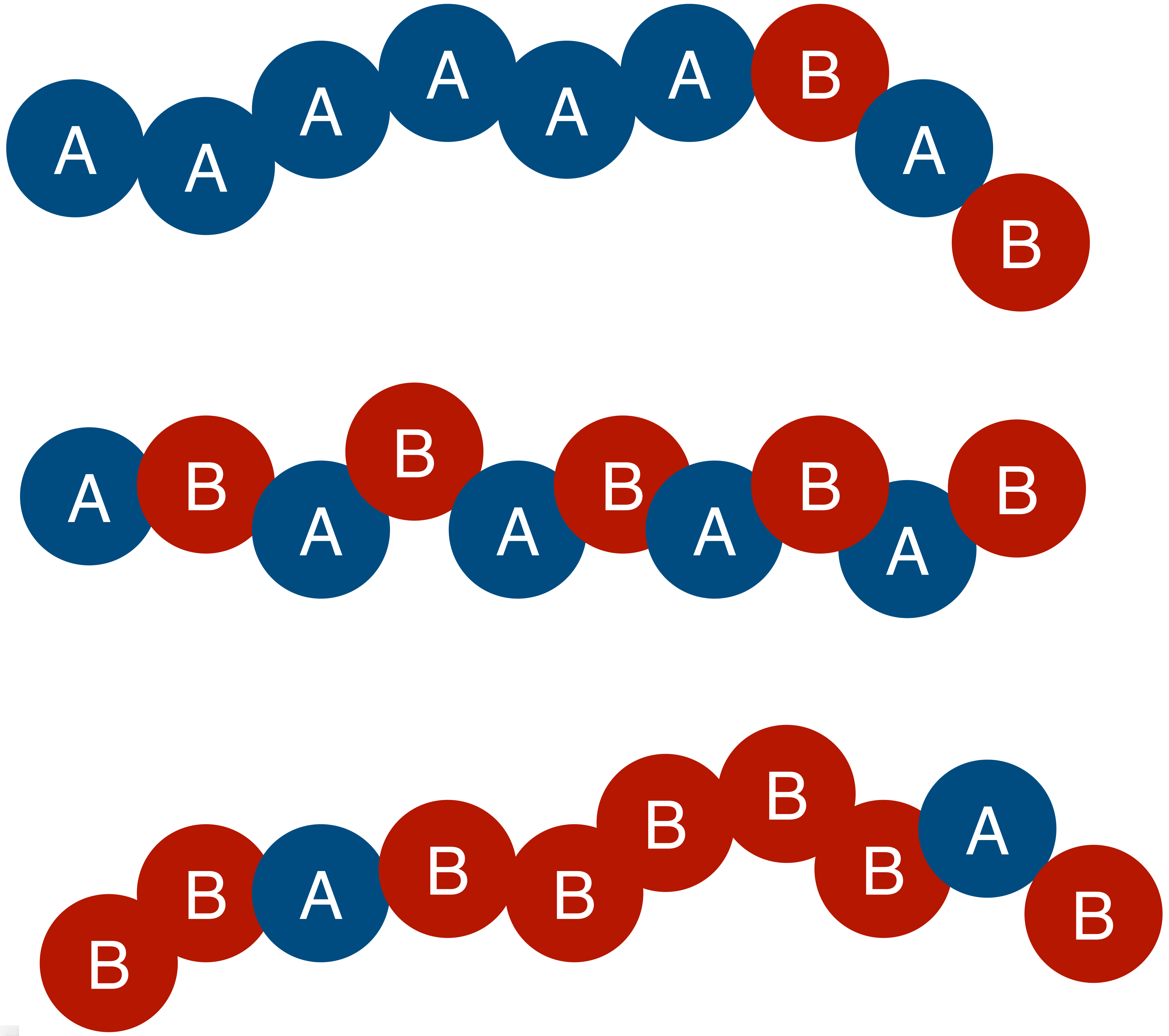


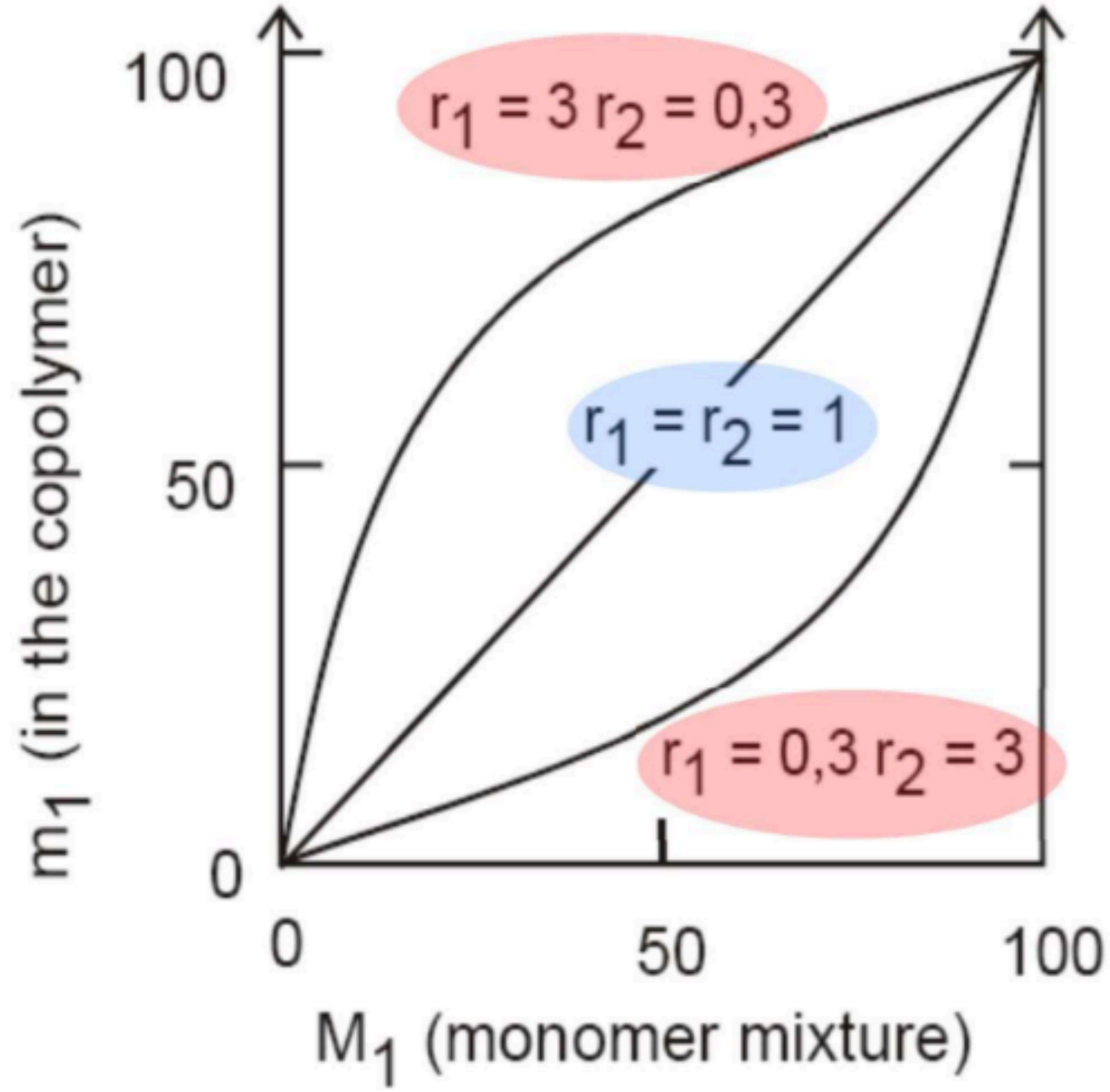
"Conventional" ML works on data we have little direct control over.

Curator
Extracting structured data

Reactivity Ratio Determines Co-Polymer Architecture

A B





Mara Schilling-Wilhelmi

Structured Data Extraction for Copolymerization Reactions

A B

Depends on copolymerization mechanism

Depends on solvent

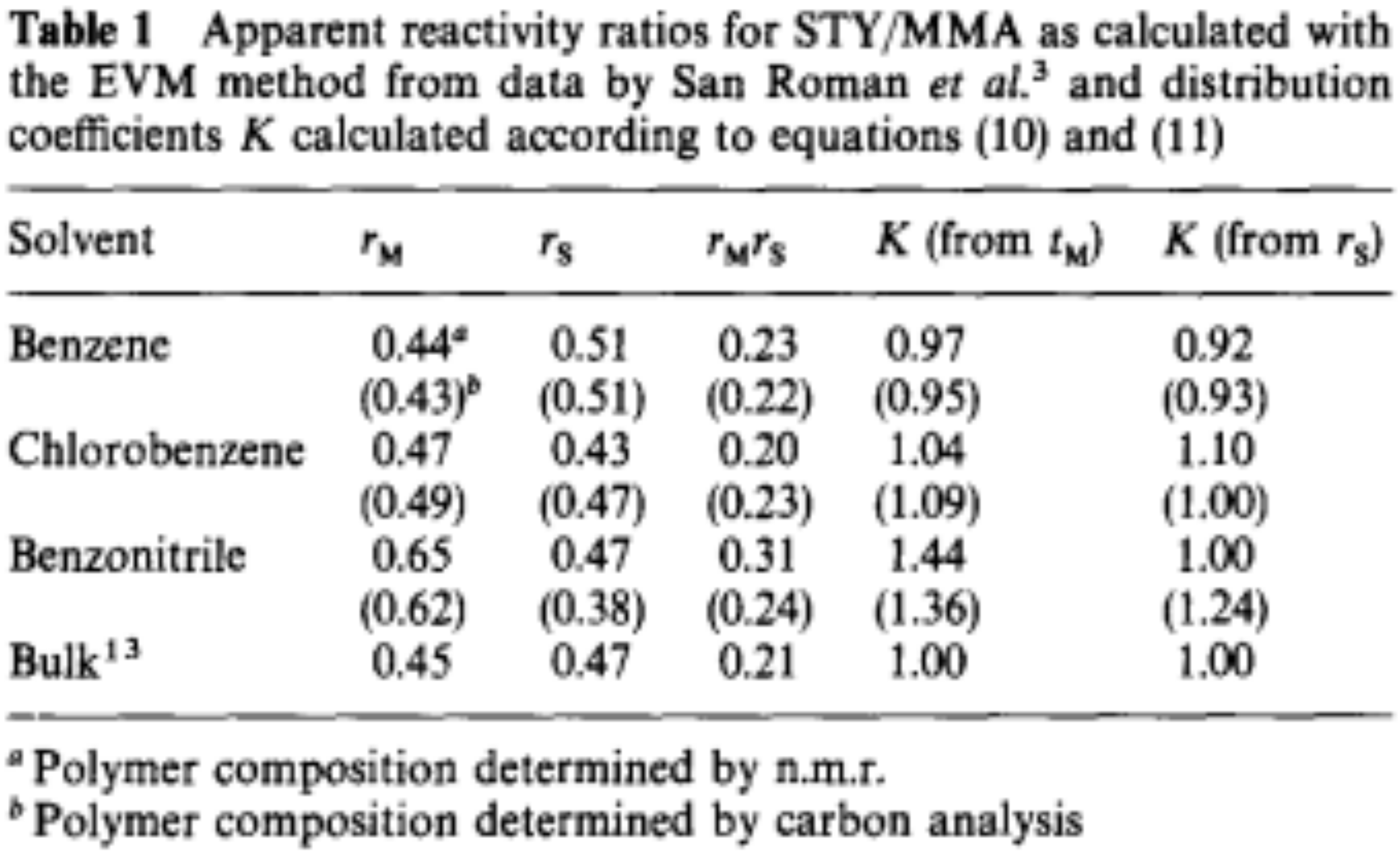


Table 1 Apparent reactivity ratios for STY/MMA as calculated with the EVM method from data by San Roman *et al.*[3] and distribution coefficients *K* calculated according to equations (10) and (11)

Solvent	r_M	r_S	$r_M r_S$	K (from t_M)	K (from r_S)
Benzene	0.44[a] (0.43)[b]	0.51 (0.51)	0.23 (0.22)	0.97 (0.95)	0.92 (0.93)
Chlorobenzene	0.47 (0.49)	0.43 (0.47)	0.20 (0.23)	1.04 (1.09)	1.10 (1.00)
Benzonitrile	0.65 (0.62)	0.47 (0.38)	0.31 (0.24)	1.44 (1.36)	1.00 (1.24)
Bulk[13]	0.45	0.47	0.21	1.00	1.00

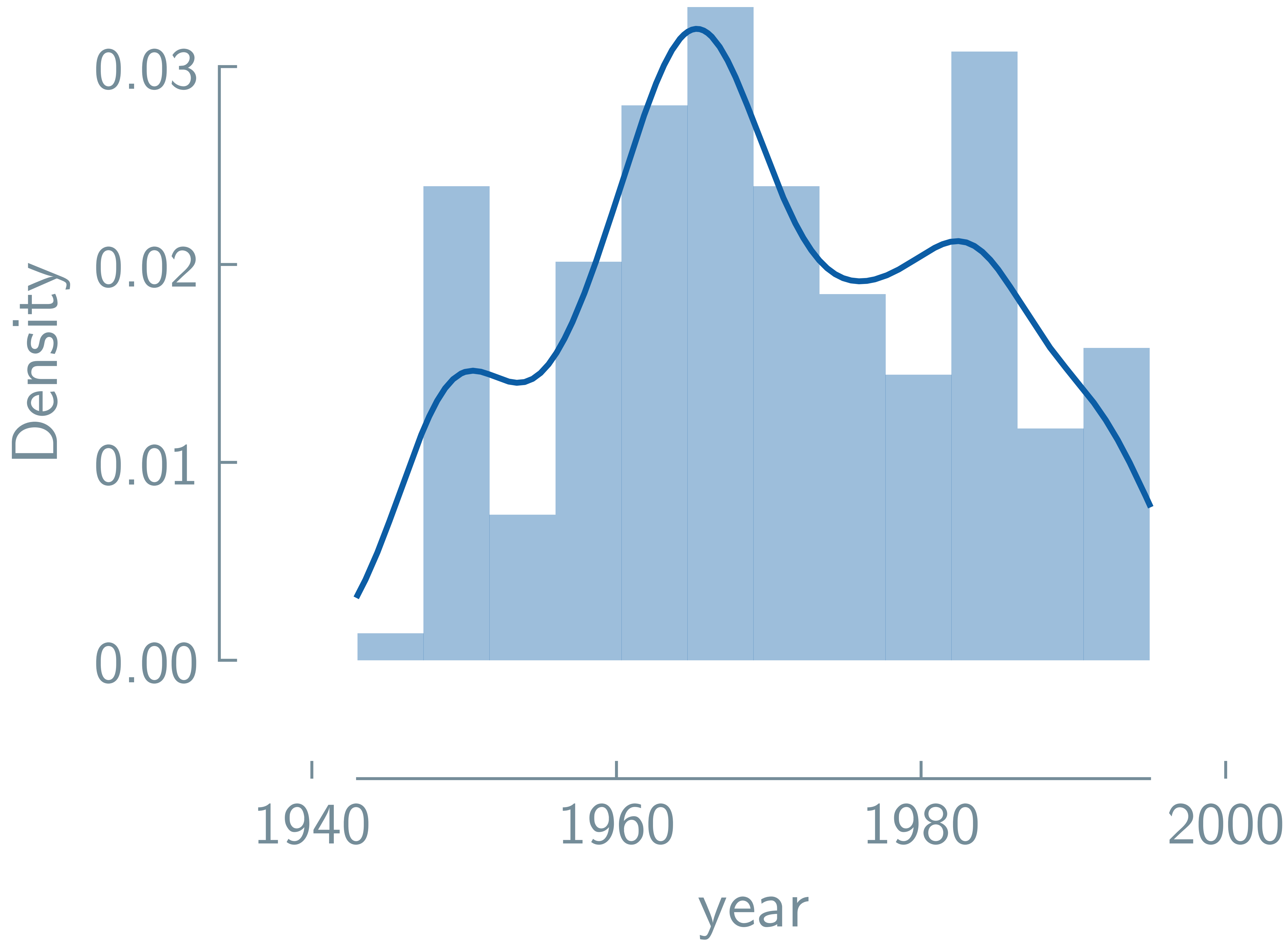
[a] Polymer composition determined by n.m.r.
[b] Polymer composition determined by carbon analysis

Polymer **1993**, 34 (5), 1032–1037.

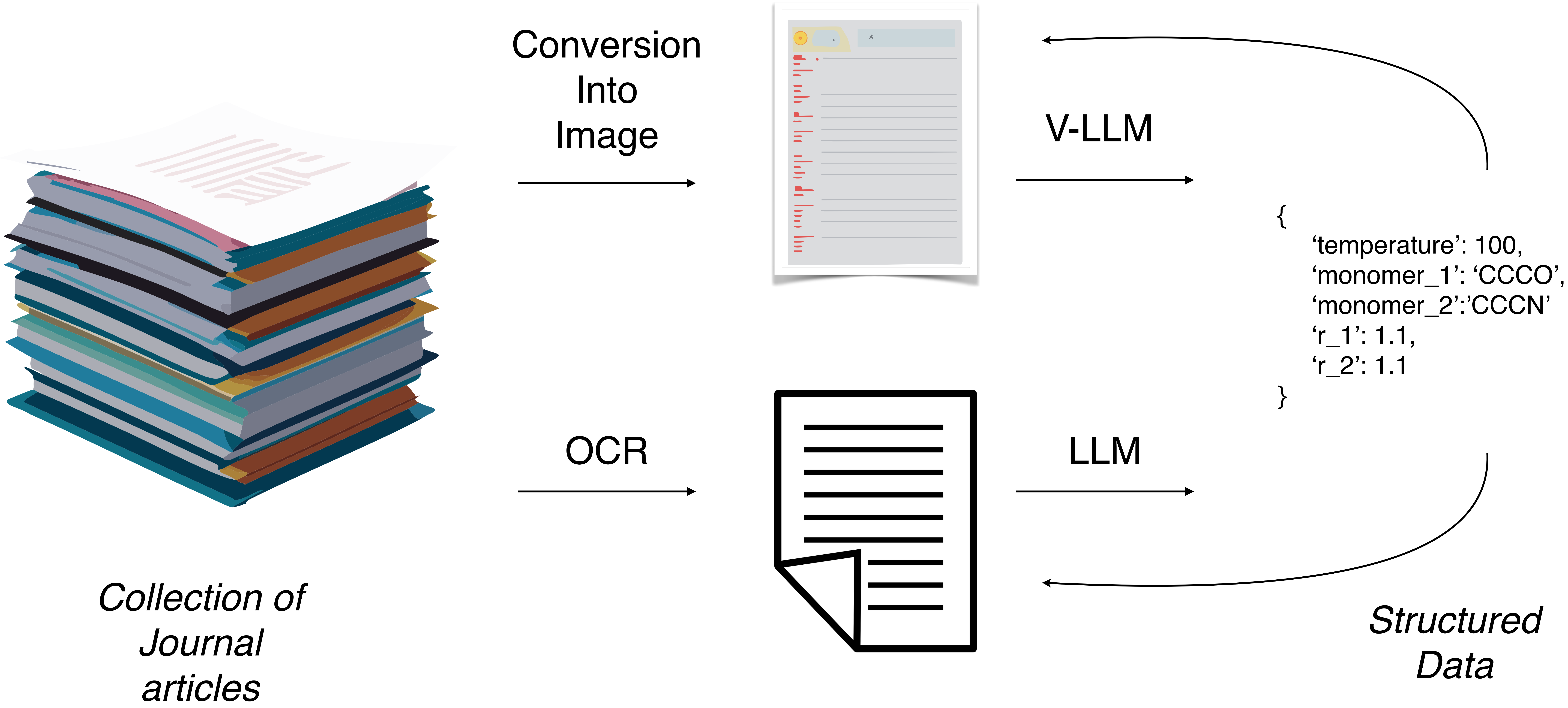
Structured Data Extraction for Copolymerization Reactions

Have to deal with old papers

Have to deal with rotated pages, …



Comparing VLMs and LLMs



Conversion Into Image

V-LLM

OCR

LLM

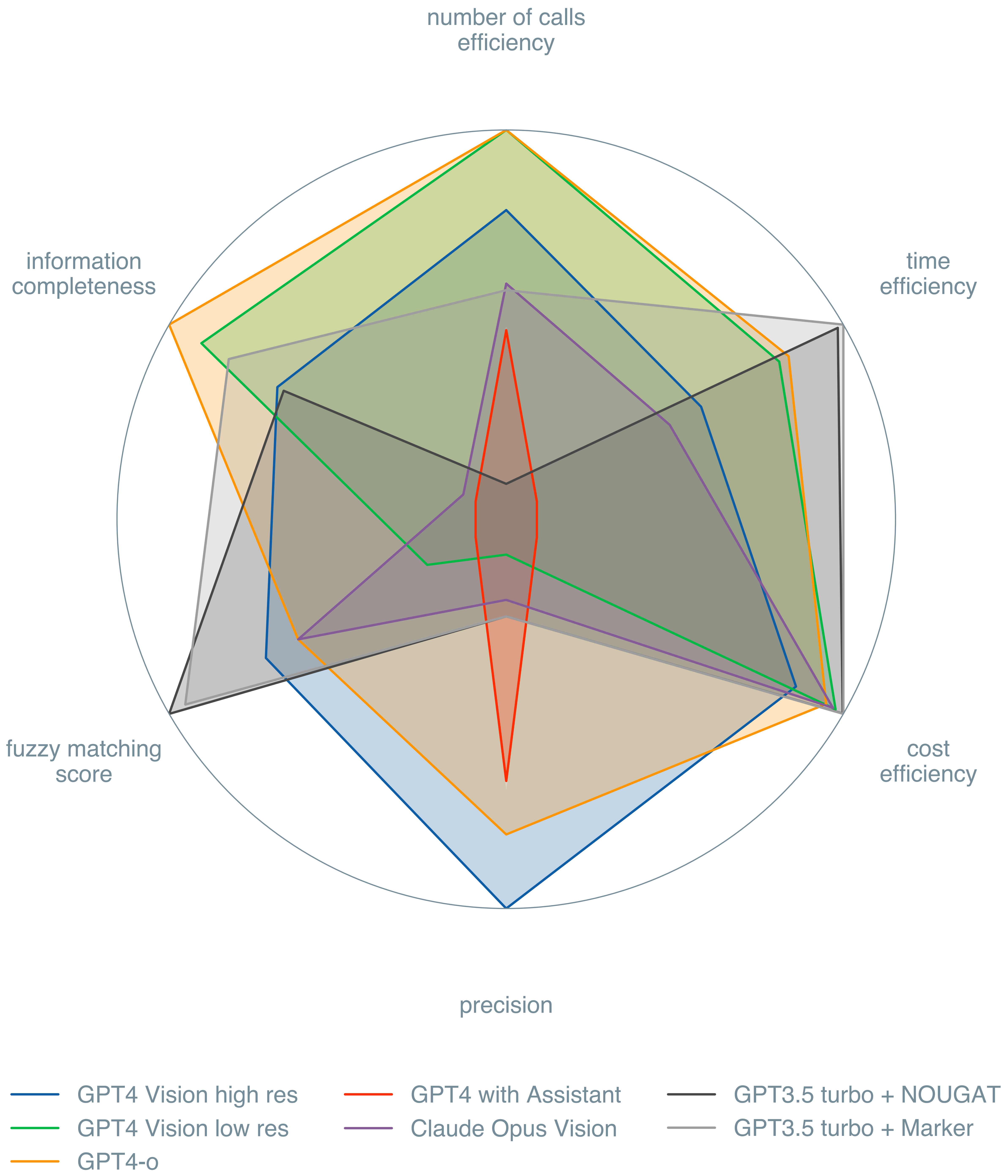
{
‘temperature’: 100,
‘monomer_1’: ‘CCCO’,
‘monomer_2’:’CCCN’
‘r_1’: 1.1,
‘r_2’: 1.1
}

Collection of Journal articles

Structured Data

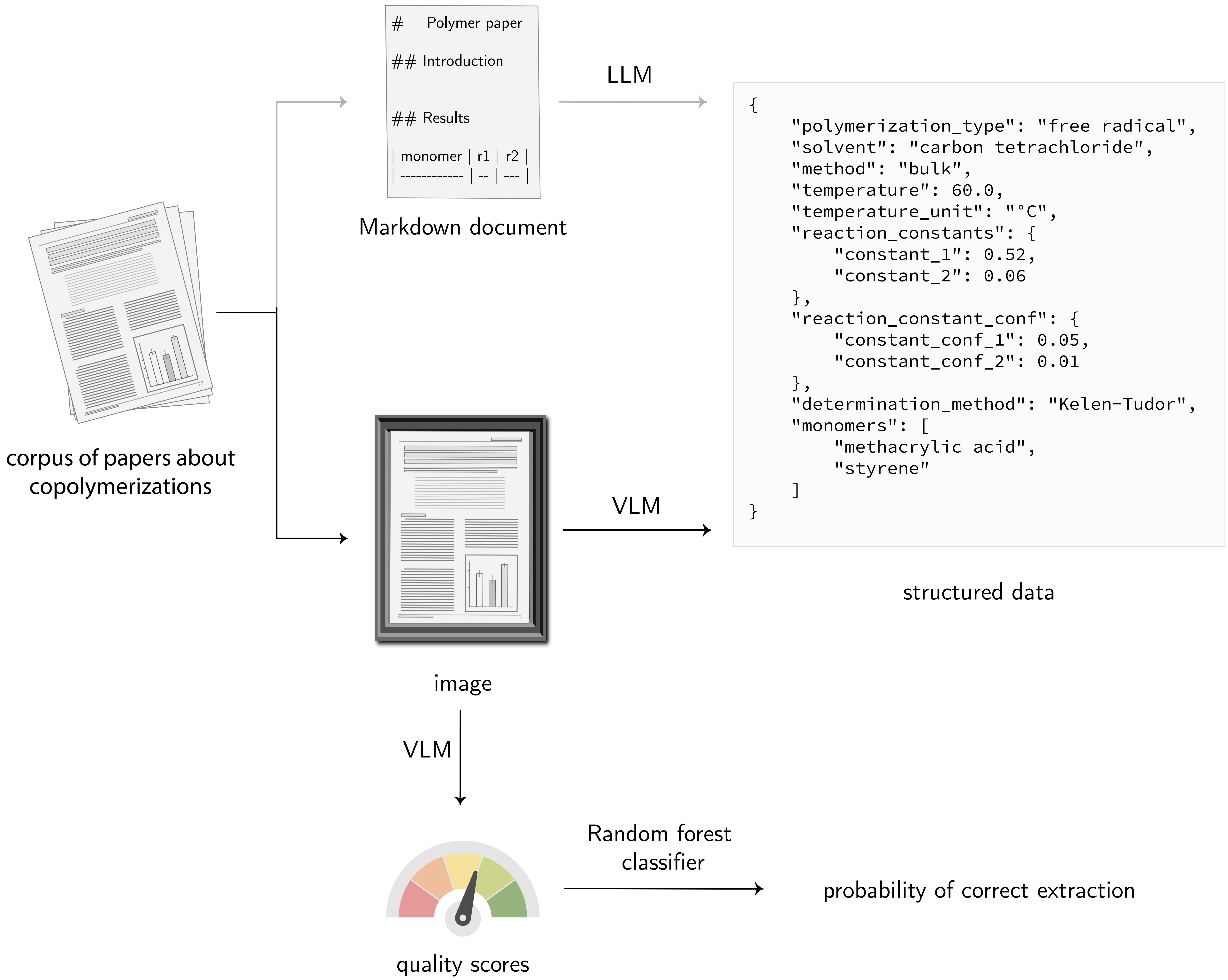
Comparing VLMs and LLMs

Only the vision models have acceptable performance



Quality Classifier Can Futher Boost Performance

Boost from 77% accuracy to 94%



corpus of papers about copolymerizations

```
#    Polymer paper

## Introduction


## Results

| monomer | r1 | r2 |
| ------------ | -- | --- |
```

Markdown document

LLM

```
{
    "polymerization_type": "free radical",
    "solvent": "carbon tetrachloride",
    "method": "bulk",
    "temperature": 60.0,
    "temperature_unit": "°C",
    "reaction_constants": {
        "constant_1": 0.52,
        "constant_2": 0.06
    },
    "reaction_constant_conf": {
        "constant_conf_1": 0.05,
        "constant_conf_2": 0.01
    },
    "determination_method": "Kelen-Tudor",
    "monomers": [
        "methacrylic acid",
        "styrene"
    ]
}
```

structured data

VLM

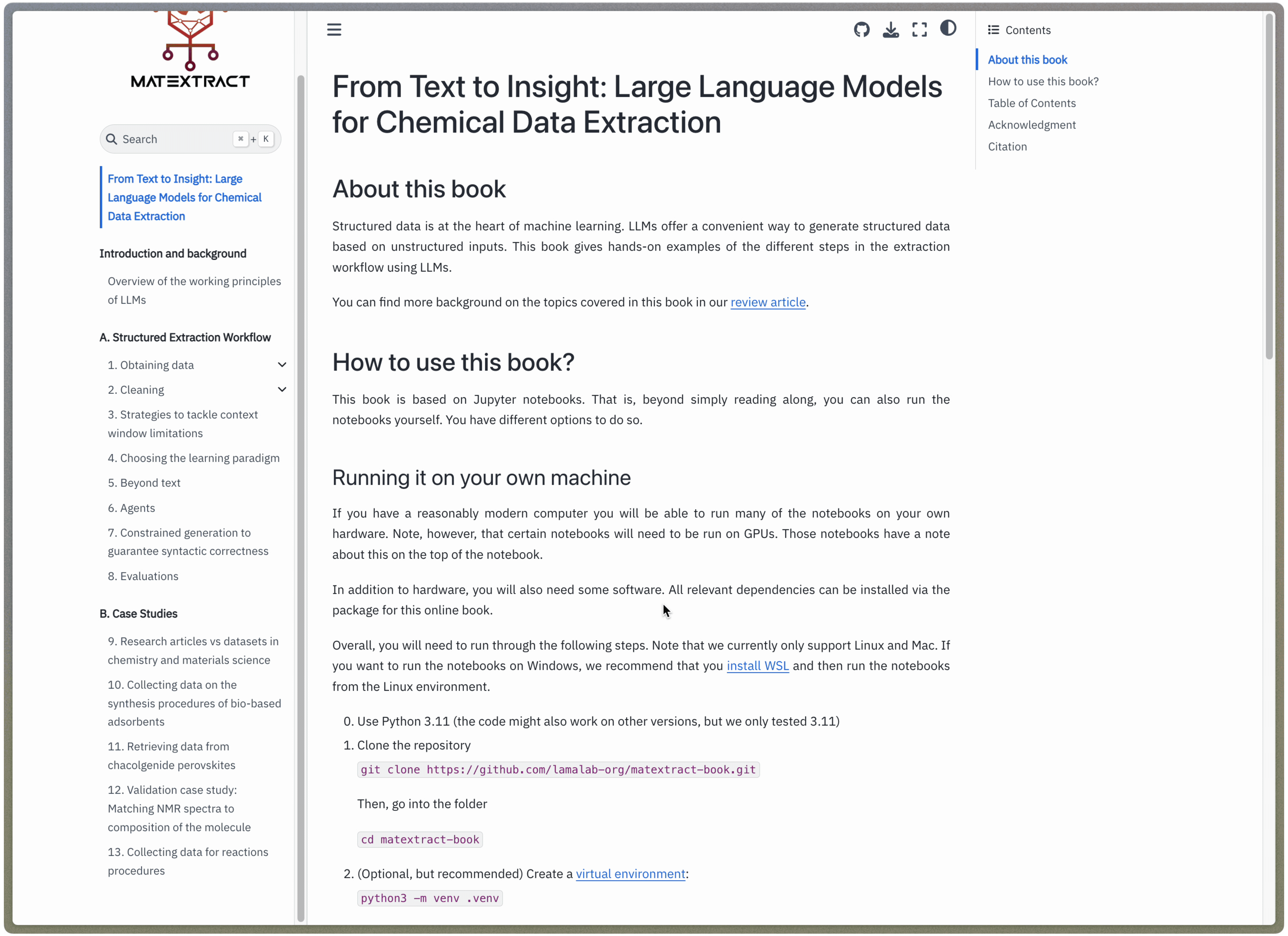
image

VLM

quality scores

Random forest classifier

probability of correct extraction



From Text to Insight: Large Language Models for Chemical Data Extraction

Introduction and background

- Overview of the working principles of LLMs

A. Structured Extraction Workflow

- 1. Obtaining data
- 2. Cleaning
- 3. Strategies to tackle context window limitations
- 4. Choosing the learning paradigm
- 5. Beyond text
- 6. Agents
- 7. Constrained generation to guarantee syntactic correctness
- 8. Evaluations

B. Case Studies

- 9. Research articles vs datasets in chemistry and materials science
- 10. Collecting data on the synthesis procedures of bio-based adsorbents
- 11. Retrieving data from chacolgenide perovskites
- 12. Validation case study: Matching NMR spectra to composition of the molecule
- 13. Collecting data for reactions procedures

Contents

- About this book
- How to use this book?
- Table of Contents
- Acknowledgment
- Citation

From Text to Insight: Large Language Models for Chemical Data Extraction

About this book

Structured data is at the heart of machine learning. LLMs offer a convenient way to generate structured data based on unstructured inputs. This book gives hands-on examples of the different steps in the extraction workflow using LLMs.

You can find more background on the topics covered in this book in our review article.

How to use this book?

This book is based on Jupyter notebooks. That is, beyond simply reading along, you can also run the notebooks yourself. You have different options to do so.

Running it on your own machine

If you have a reasonably modern computer you will be able to run many of the notebooks on your own hardware. Note, however, that certain notebooks will need to be run on GPUs. Those notebooks have a note about this on the top of the notebook.

In addition to hardware, you will also need some software. All relevant dependencies can be installed via the package for this online book.

Overall, you will need to run through the following steps. Note that we currently only support Linux and Mac. If you want to run the notebooks on Windows, we recommend that you install WSL and then run the notebooks from the Linux environment.

0. Use Python 3.11 (the code might also work on other versions, but we only tested 3.11)
1. Clone the repository

```
git clone https://github.com/lamalab-org/matextract-book.git
```

   Then, go into the folder

```
cd matextract-book
```

2. (Optional, but recommended) Create a virtual environment:

```
python3 -m venv .venv
```

Mara Schilling-Wilhelmi

Martiño Ríos García

Sherjeel Shabih

Pepe Márquez

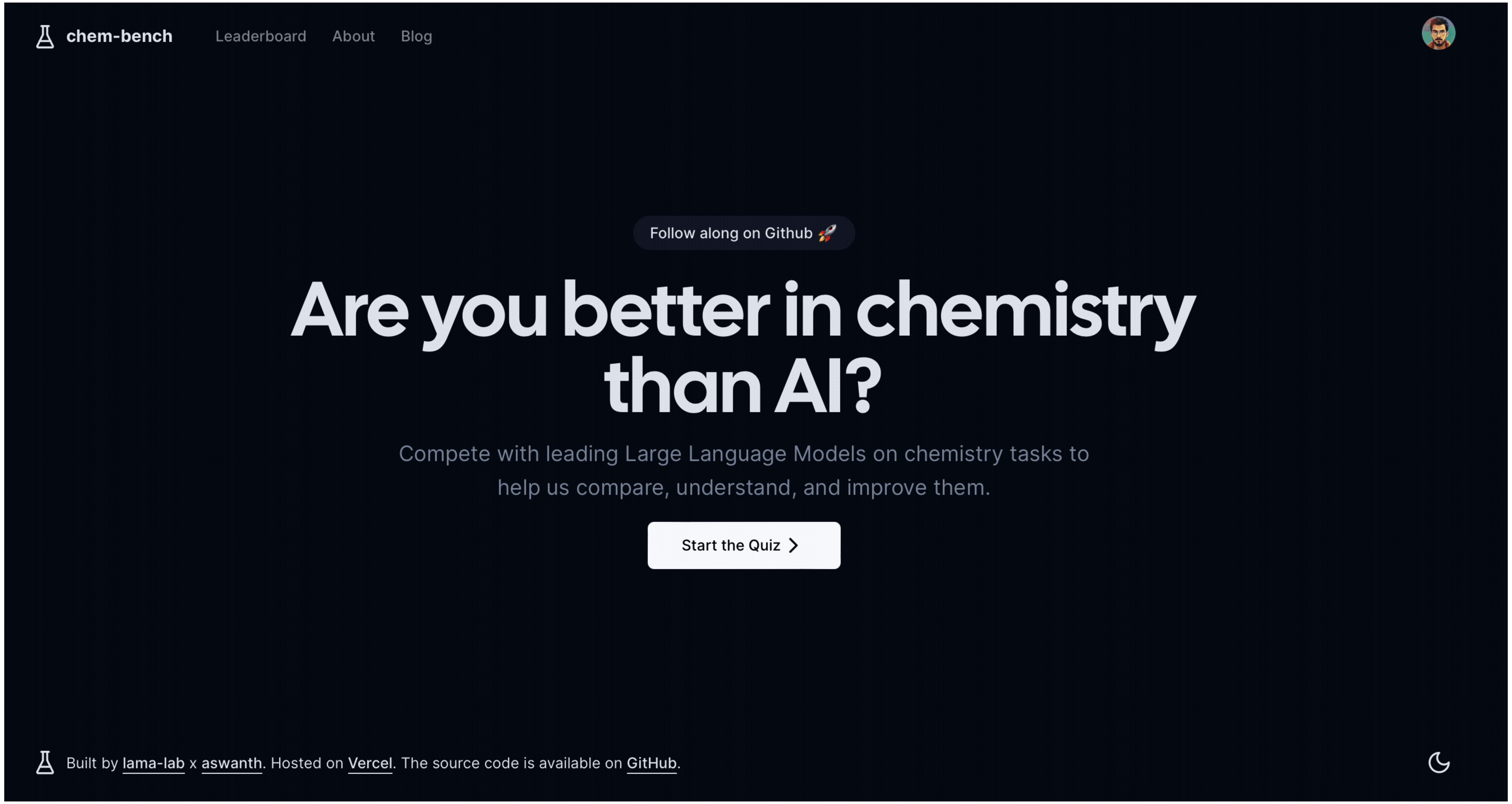
Christoph T Koch

Santiago Miret

María Victoria Gil

Matextract.pub and arXiv:2407.16867

How Good Are Current Models Compared to Experts?

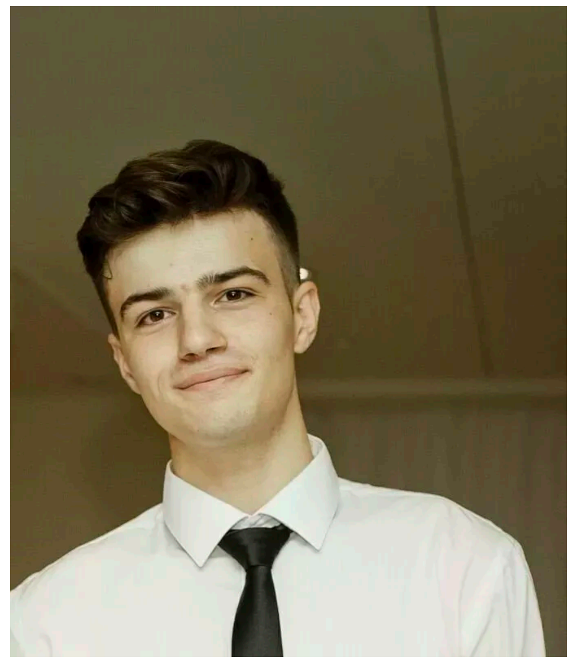


Are you better in chemistry than AI?

Compete with leading Large Language Models on chemistry tasks to help us compare, understand, and improve them.

Built by lama-lab x aswanth. Hosted on Vercel. The source code is available on GitHub.

Test yourself at chembench.org

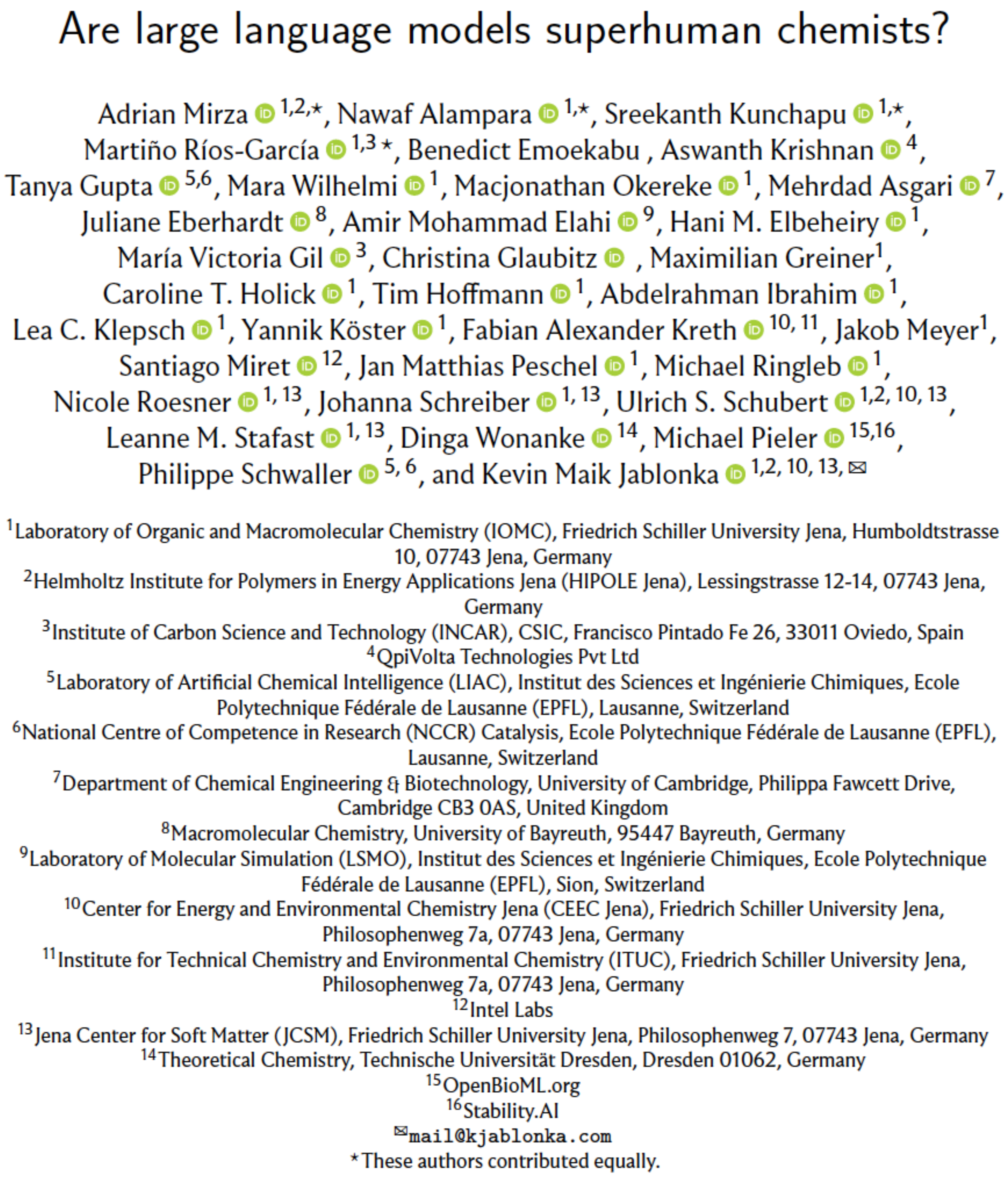


Adrian Mirza

Nawaf Alampara

Martiño Ríos García

Sreekanth Kunchapu



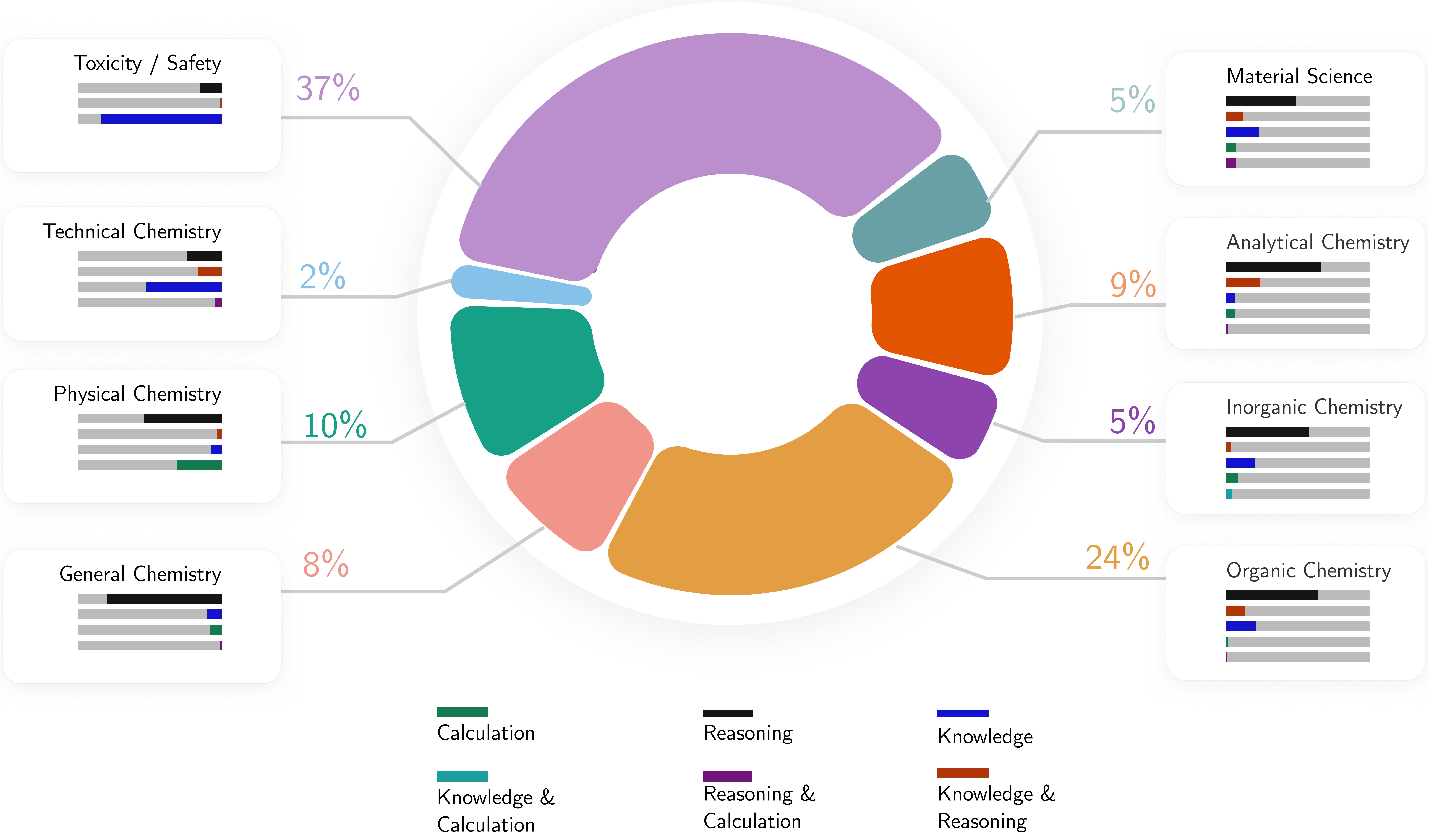
Are large language models superhuman chemists?

Adrian Mirza[1,2,*], Nawaf Alampara[1,*], Sreekanth Kunchapu[1,*],
Martiño Ríos-García[1,3*], Benedict Emoekabu, Aswanth Krishnan[4],
Tanya Gupta[5,6], Mara Wilhelmi[1], Macjonathan Okereke[1], Mehrdad Asgari[7],
Juliane Eberhardt[8], Amir Mohammad Elahi[9], Hani M. Elbeheiry[1],
María Victoria Gil[3], Christina Glaubitz, Maximilian Greiner[1],
Caroline T. Holick[1], Tim Hoffmann[1], Abdelrahman Ibrahim[1],
Lea C. Klepsch[1], Yannik Köster[1], Fabian Alexander Kreth[10, 11], Jakob Meyer[1],
Santiago Miret[12], Jan Matthias Peschel[1], Michael Ringleb[1],
Nicole Roesner[1, 13], Johanna Schreiber[1, 13], Ulrich S. Schubert[1,2, 10, 13],
Leanne M. Stafast[1, 13], Dinga Wonanke[14], Michael Pieler[15,16],
Philippe Schwaller[5, 6], and Kevin Maik Jablonka[1,2, 10, 13, ✉]

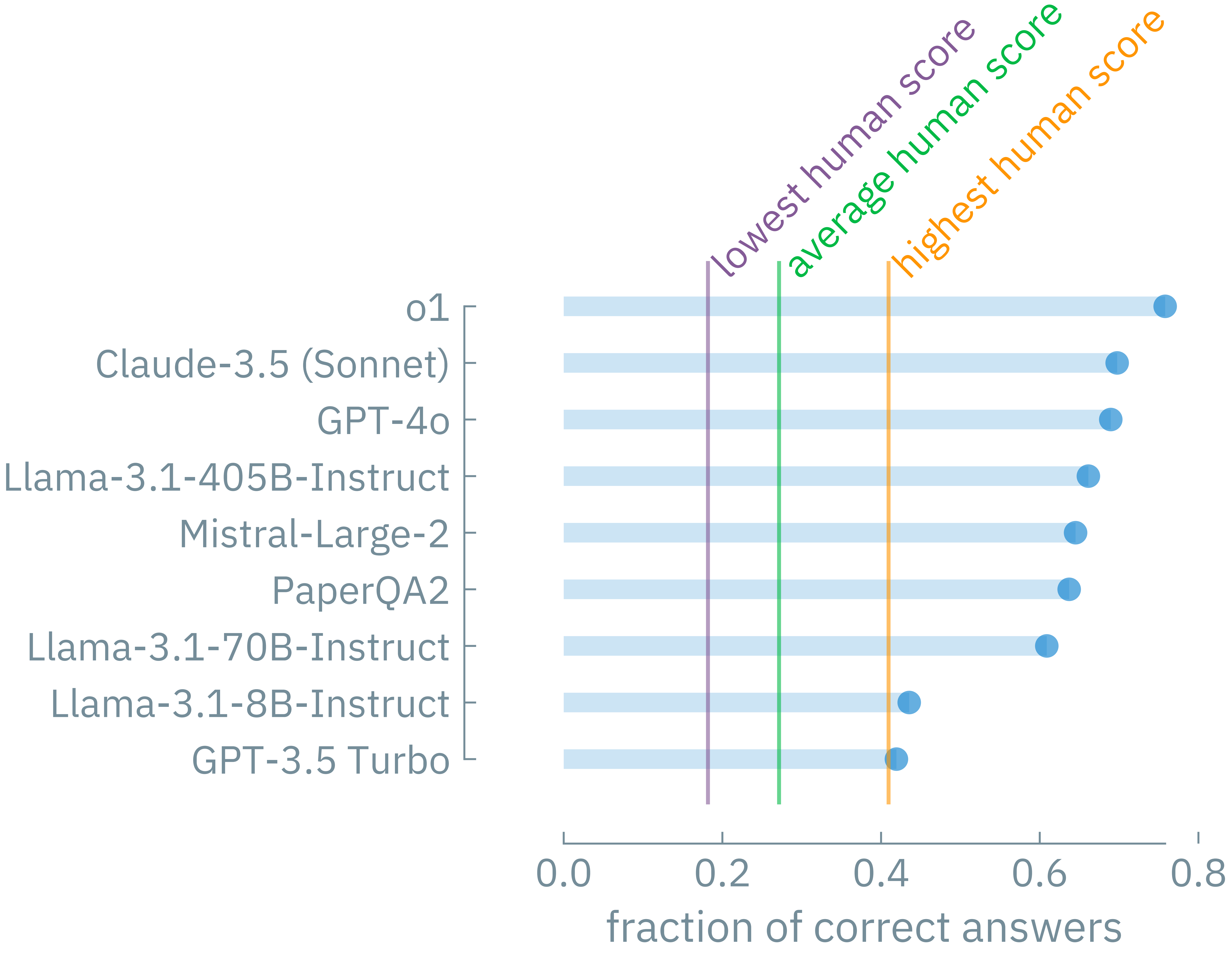
[1] Laboratory of Organic and Macromolecular Chemistry (IOMC), Friedrich Schiller University Jena, Humboldtstrasse 10, 07743 Jena, Germany
[2] Helmholtz Institute for Polymers in Energy Applications Jena (HIPOLE Jena), Lessingstrasse 12-14, 07743 Jena, Germany
[3] Institute of Carbon Science and Technology (INCAR), CSIC, Francisco Pintado Fe 26, 33011 Oviedo, Spain
[4] QpiVolta Technologies Pvt Ltd
[5] Laboratory of Artificial Chemical Intelligence (LIAC), Institut des Sciences et Ingénierie Chimiques, Ecole Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland
[6] National Centre of Competence in Research (NCCR) Catalysis, Ecole Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland
[7] Department of Chemical Engineering & Biotechnology, University of Cambridge, Philippa Fawcett Drive, Cambridge CB3 0AS, United Kingdom
[8] Macromolecular Chemistry, University of Bayreuth, 95447 Bayreuth, Germany
[9] Laboratory of Molecular Simulation (LSMO), Institut des Sciences et Ingénierie Chimiques, Ecole Polytechnique Fédérale de Lausanne (EPFL), Sion, Switzerland
[10] Center for Energy and Environmental Chemistry Jena (CEEC Jena), Friedrich Schiller University Jena, Philosophenweg 7a, 07743 Jena, Germany
[11] Institute for Technical Chemistry and Environmental Chemistry (ITUC), Friedrich Schiller University Jena, Philosophenweg 7a, 07743 Jena, Germany
[12] Intel Labs
[13] Jena Center for Soft Matter (JCSM), Friedrich Schiller University Jena, Philosophenweg 7, 07743 Jena, Germany
[14] Theoretical Chemistry, Technische Universität Dresden, Dresden 01062, Germany
[15] OpenBioML.org
[16] Stability.AI
✉ `mail@kjablonka.com`
* These authors contributed equally.

arXiv:2404.01475

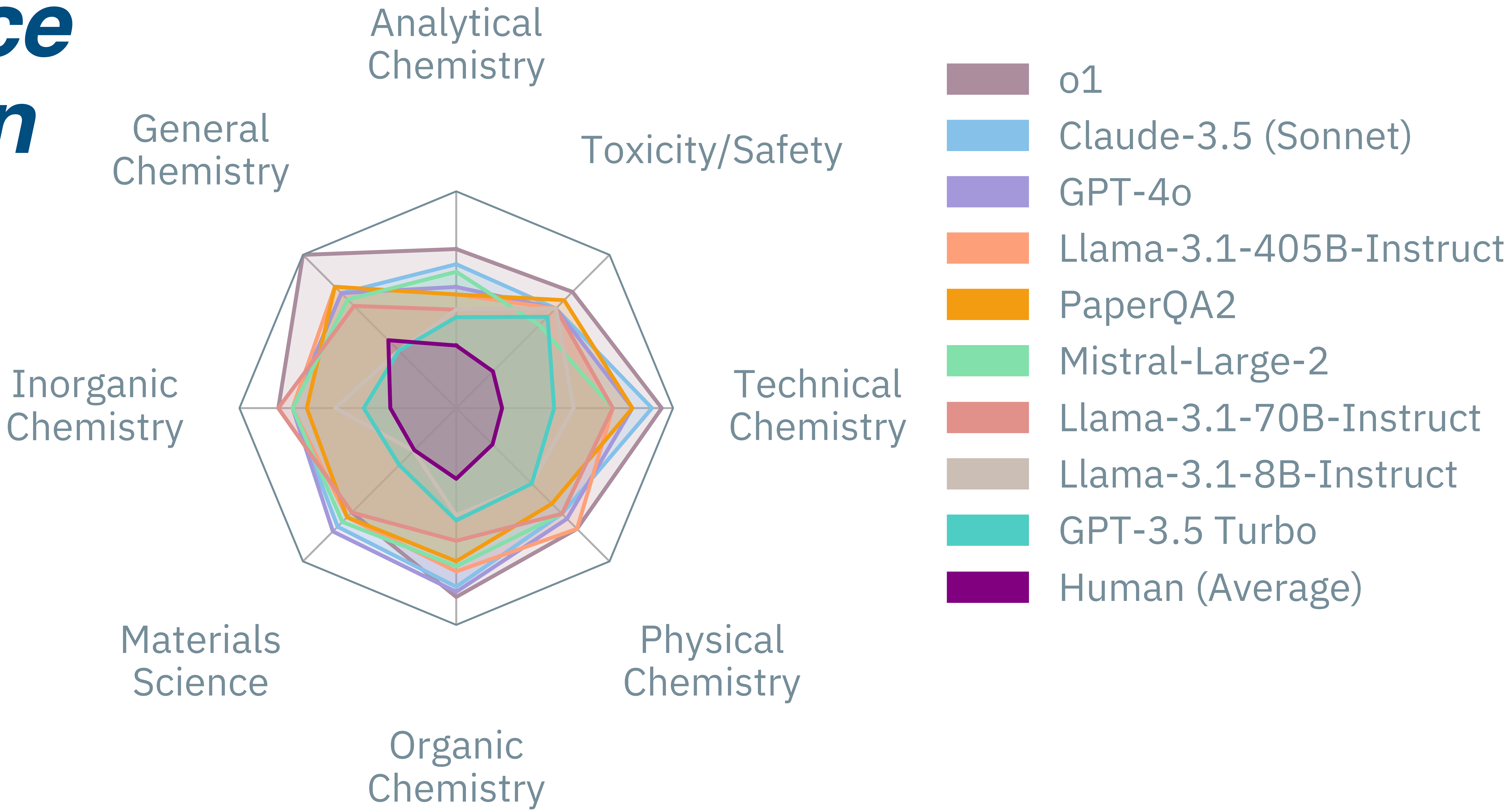
Large Corpus Testing Diverse Skills



On Average: Superhuman Performance



But Performance Depends on Topic



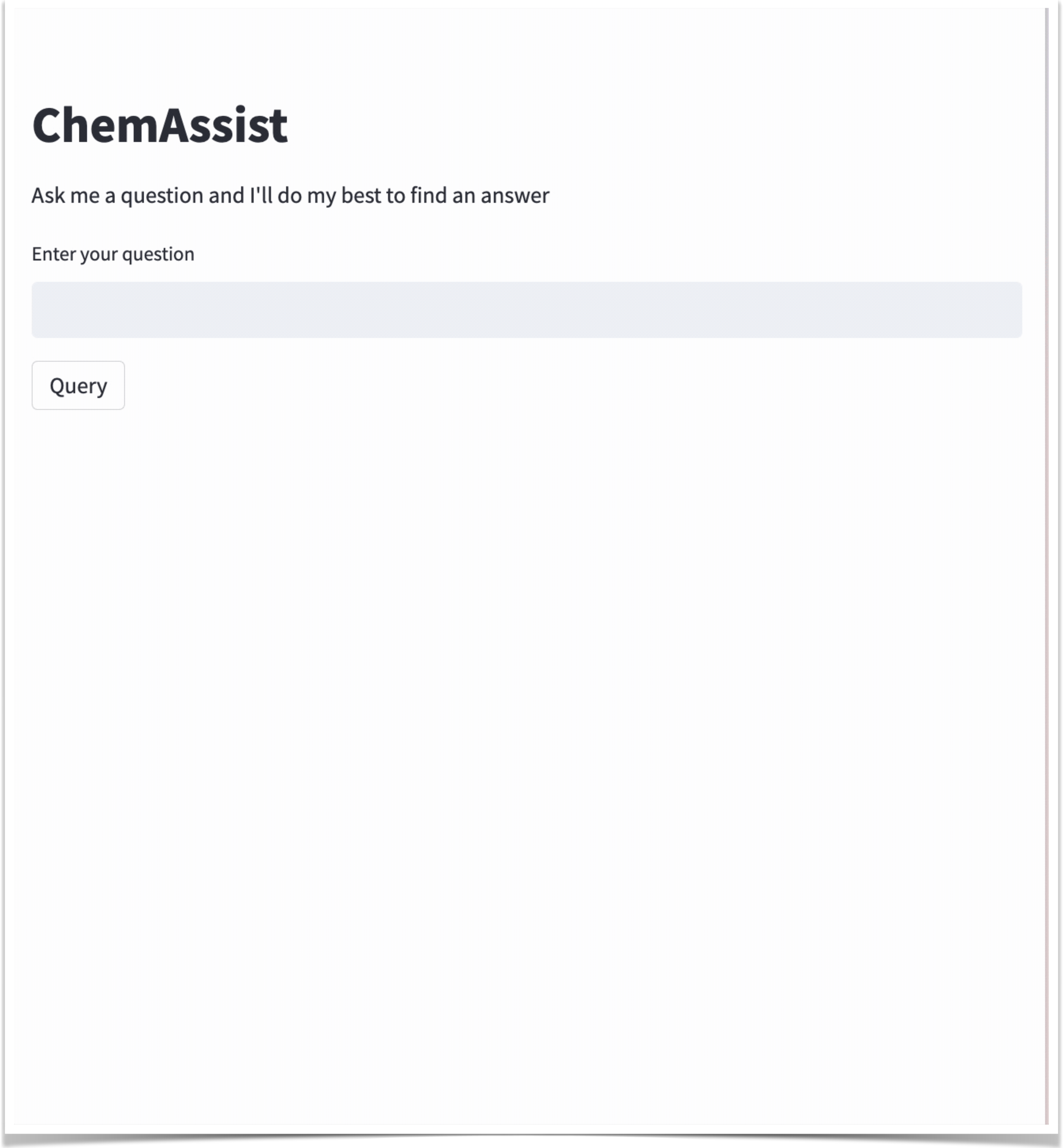
Connecting Tools

Conventional user interfaces are rigid chemistry is flexible

LLMs can do semantic search

LLMs can dynamically create interface components

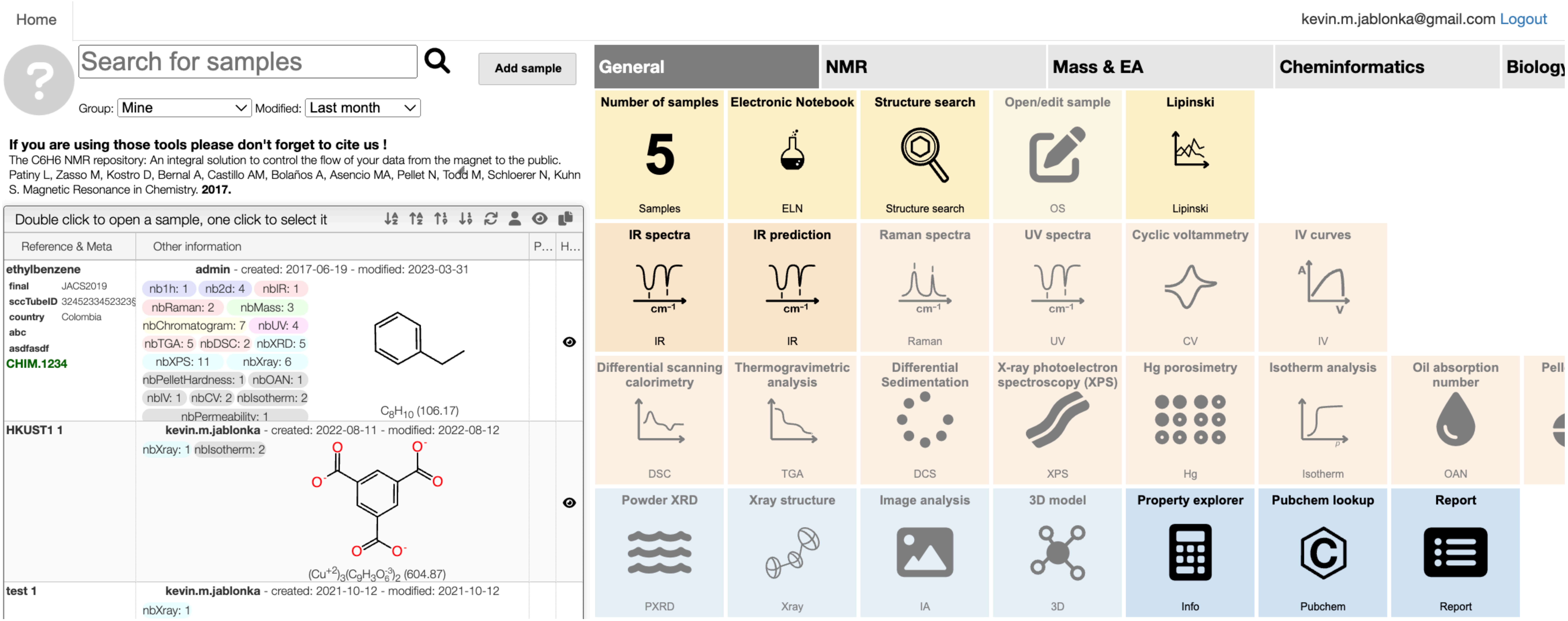
LLMs can reason about the data



ChemAssist

Ask me a question and I'll do my best to find an answer

Enter your question



Group: Mine
Modified: Last month

If you are using those tools please don't forget to cite us !

The C6H6 NMR repository: An integral solution to control the flow of your data from the magnet to the public. Patiny L, Zasso M, Kostro D, Bernal A, Castillo AM, Bolaños A, Asencio MA, Pellet N, Todd M, Schloerer N, Kuhn S. Magnetic Resonance in Chemistry. **2017.**

Double click to open a sample, one click to select it

Reference & Meta	Other information	P…	H…
ethylbenzene **final** JACS2019 **sccTubeID** 3245233452323 **country** Colombia **abc** **asdfasdf** **CHIM.1234**	**admin** - created: 2017-06-19 - modified: 2023-03-31 nb1h: 1 nb2d: 4 nbIR: 1 nbRaman: 2 nbMass: 3 nbChromatogram: 7 nbUV: 4 nbTGA: 5 nbDSC: 2 nbXRD: 5 nbXPS: 11 nbXray: 6 nbPelletHardness: 1 nbOAN: 1 nbIV: 1 nbCV: 2 nbIsotherm: 2 nbPermeability: 1 C_8H_{10} (106.17)		👁
HKUST1 1	**kevin.m.jablonka** - created: 2022-08-11 - modified: 2022-08-12 nbXray: 1 nbIsotherm: 2 $(Cu^{+2})_3(C_9H_3O_6^{-3})_2$ (604.87)		👁
test 1	**kevin.m.jablonka** - created: 2021-10-12 - modified: 2021-10-12 nbXray: 1		

Nature Chemistry **2022**, 14 (4), 365–376.

Chemotion
All
IUPAC, InChI, SMILES, RInChI
0
Nicole Jung
Collections
All
chemotion-repository.net
Test Collection Data
Lamella Example
MOF Test collection
SurMoF Test Collection
MOF examples
Polymers
Linker
Metall-Precursor
Calculation Doptants
MOF collection
Info_Exchange IFG_IBCS_IOC
Extracted data
Calculation OLED
Calculation TADF
Lamella Test Collection
New Collection
My Data
Joachim
Fabian
My shared collections
Shared with me
Synchronized with me
Inbox 14
15(0)
36(0)
0(0)
0(0)
From
To
NJ-R1540
NJ-R1541
NJ-R1539
NJ-R1538
NJ-R1542 Br CF3 dienophile
NJ-R1546 CF3 dienophile precursor
NJ-R1544 Br NAc dienophile
NJ-R1547 Br NAc dienophile precursor
1 - 0
«
1
›
»
Show
15
NJ-R1546 CF3 dienophile precursor
CH3
H_3C
Br
HO
O
NH_2
F
O
40 °C, 24 hr
H_3C
HN
71%
Scheme
Properties
Literature
Analyses
Solvents
Conditions

Starting materials	Ref	L/S/T/RAmount						Conc		Equiv
A NJ-2819 1,4-dimethoxynaphthalene	●	s t	188.0	mg	0.00	ml	0.9988 mmol	n.d.	mmol/l	1.000
B NJ-2820 5-amino-2-bromobenzoic acid	○	s t	215.8	mg	0.00	ml	0.9988 mmol	n.d.	mmol/l	1.000
Reactants		Reagents								
C (2,2,2-trifluoroacetyl) 2,2,2-t...	○	s t	2098	mg	1.39	ml	9.988 mmol	n.d.	mmol/l	10.00
Products		**L/S/T/RAmount**						**Conc**		**Yield**
P1 NJ-2821 **JB-R401-...** C21H15BrF3NO4		s r	341.0	mg	0.00	ml	0.7071 mmol	n.d.	mmol/l	71%

Name
CF3 dienophile precursor

Status
Successful

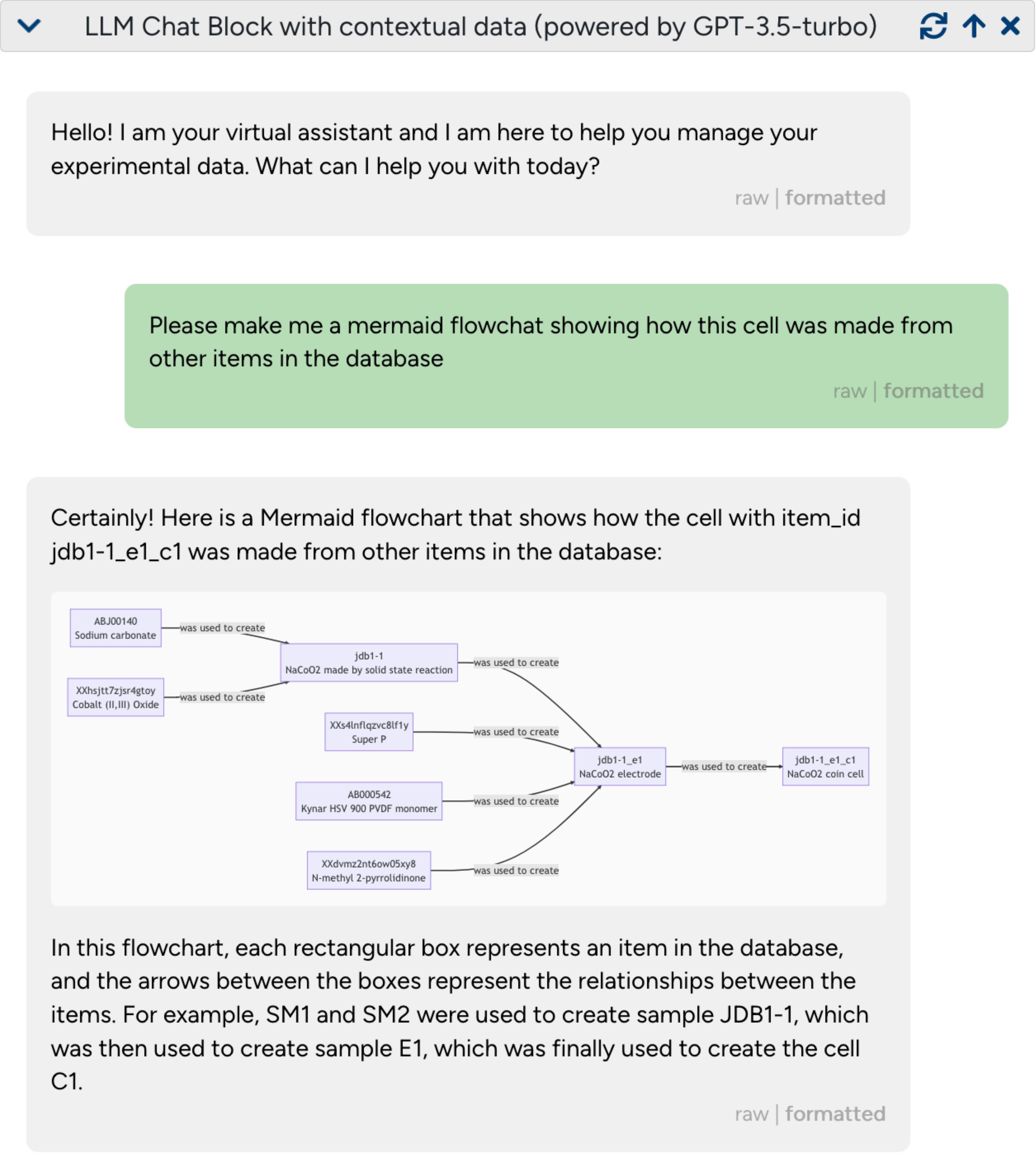
Temperature
40 °C

Start
DD/MM/YYYY hh:mm:ss

Stop
DD/MM/YYYY hh:mm:ss

Duration
Duration
24 Hour(s)

Chatting With Your Data



LLM Chat Block with contextual data (powered by GPT-3.5-turbo)

Hello! I am your virtual assistant and I am here to help you manage your experimental data. What can I help you with today?

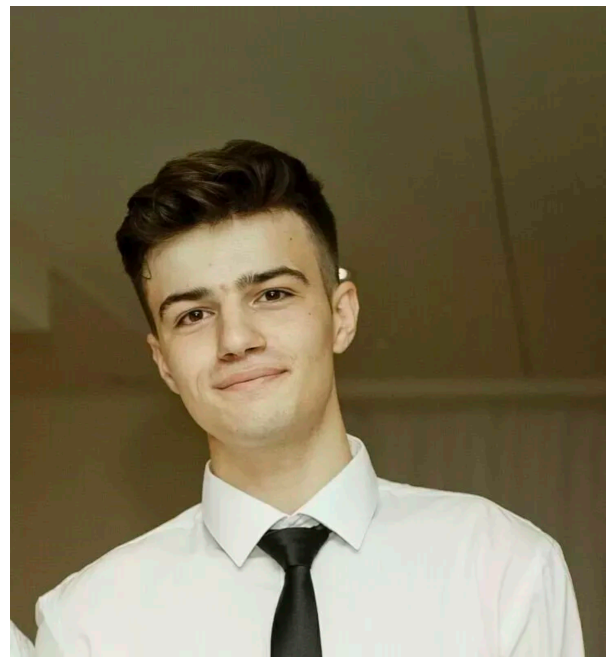
Please make me a mermaid flowchat showing how this cell was made from other items in the database

Certainly! Here is a Mermaid flowchart that shows how the cell with item_id jdb1-1_e1_c1 was made from other items in the database:

In this flowchart, each rectangular box represents an item in the database, and the arrows between the boxes represent the relationships between the items. For example, SM1 and SM2 were used to create sample JDB1-1, which was then used to create sample E1, which was finally used to create the cell C1.

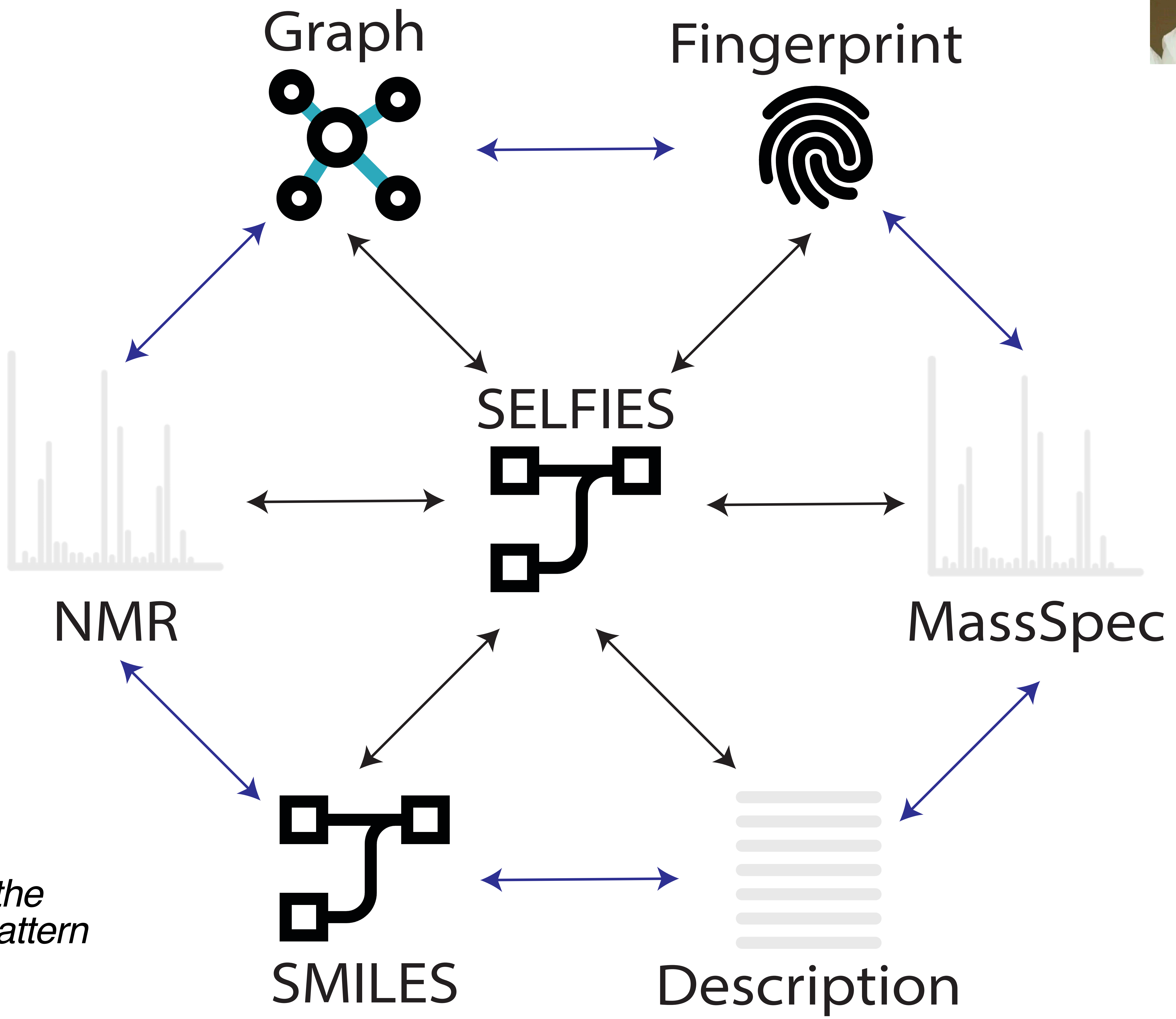
*Matthew Evans, Joshua D. Bocarsly, Ben E. Smith in Digital Discovery **2023**, 2 (5), 1233–1250*

Binding Molecular Embeddings



Adrian Mirza

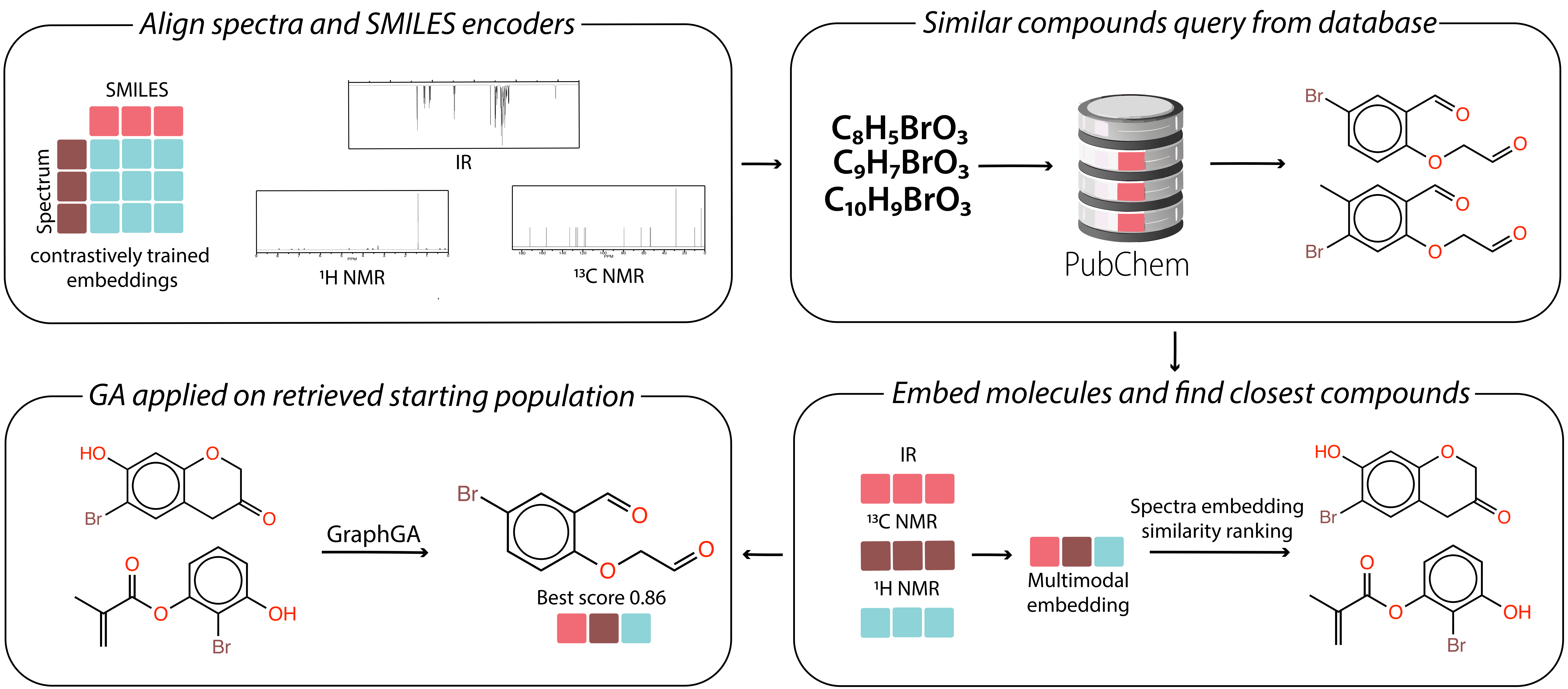
What if you can find the procedure that best matches the spectrum you want to see?



Architecture: Girdhar, Rohit, et al. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023

AI4Mat 2024 Spotlight

Building Matching Structures That Match Spectra



Align spectra and SMILES encoders

contrastively trained embeddings

Similar compounds query from database

$C_8H_5BrO_3$
$C_9H_7BrO_3$
$C_{10}H_9BrO_3$

Embed molecules and find closest compounds

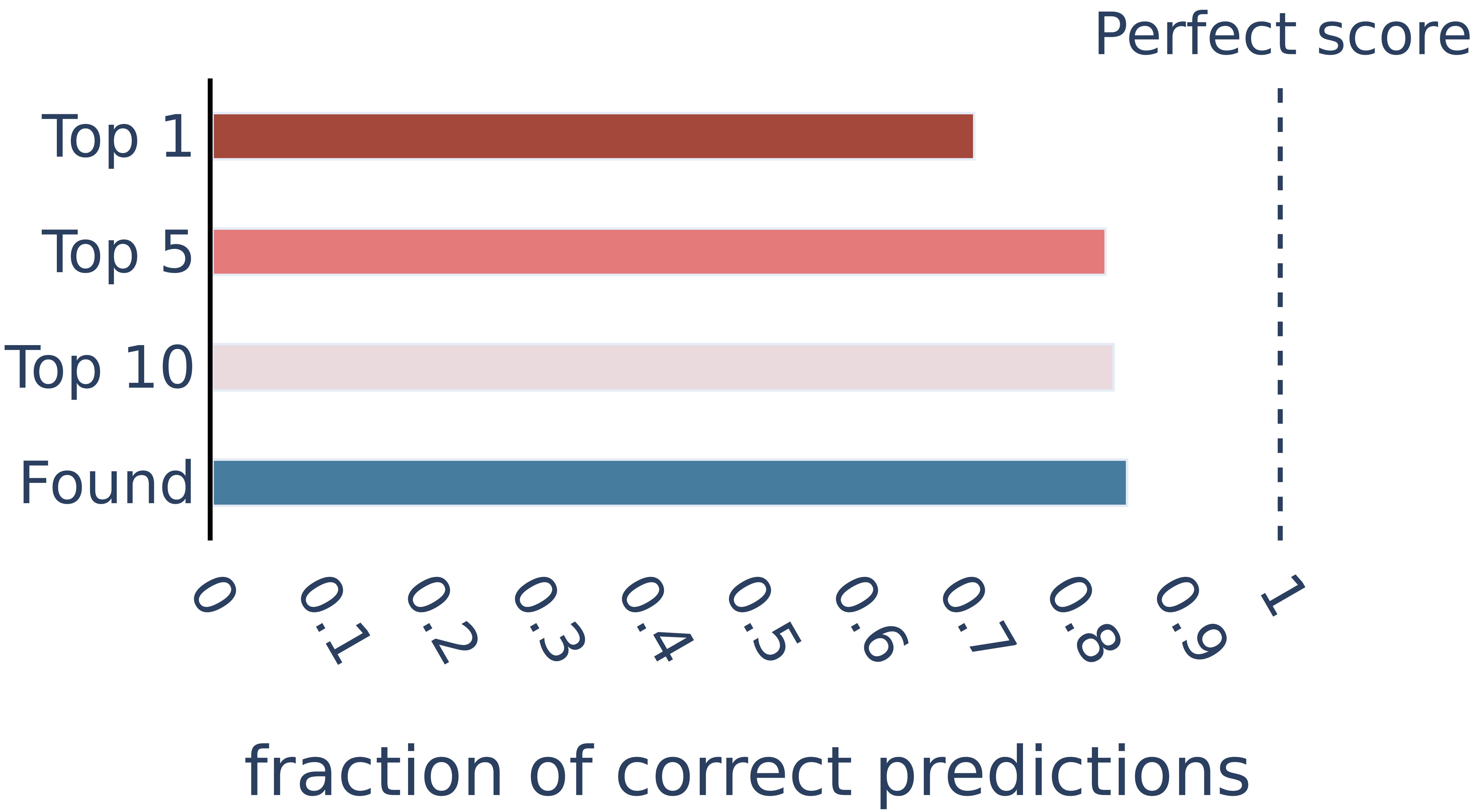
Multimodal embedding

Spectra embedding similarity ranking

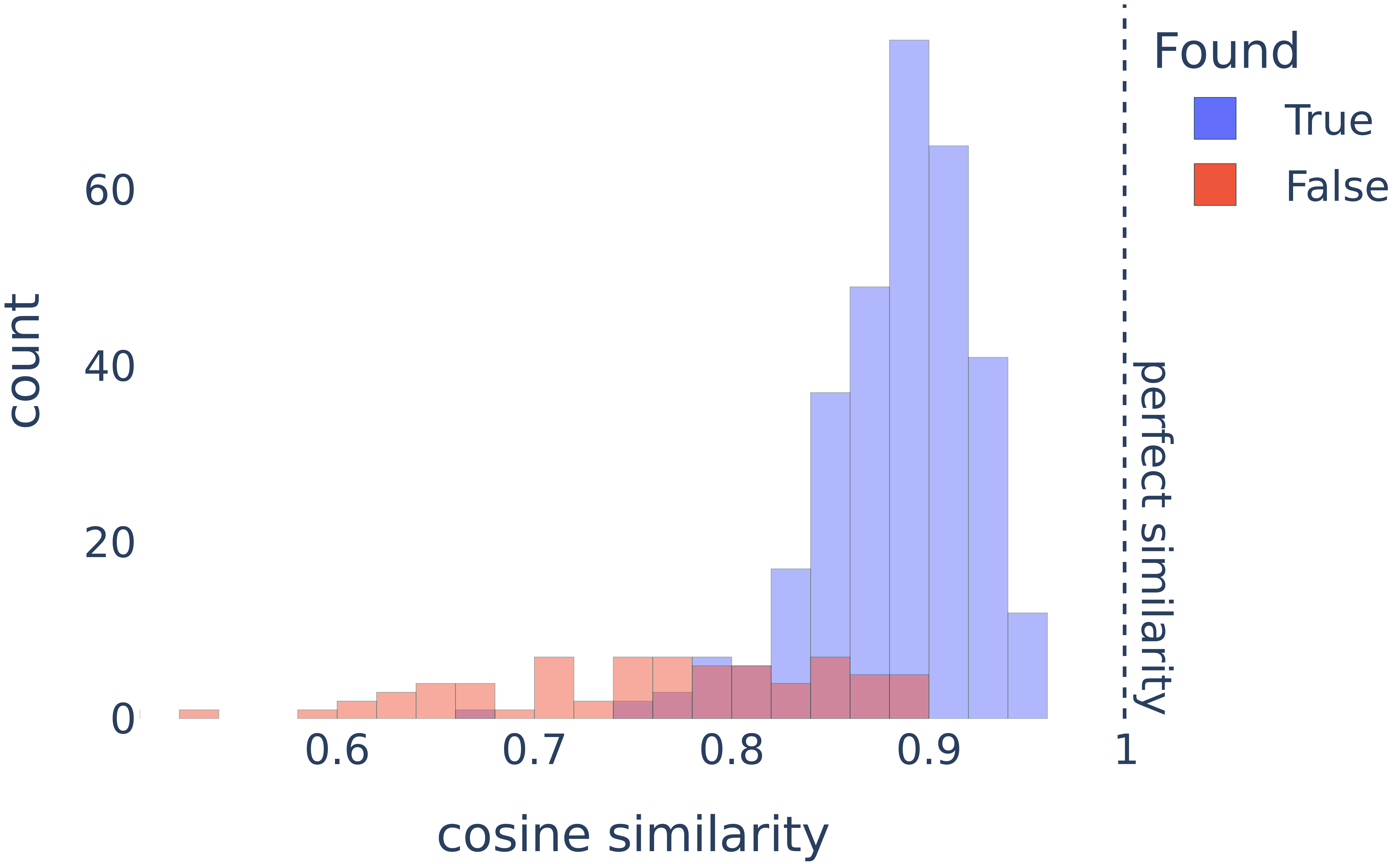
GA applied on retrieved starting population

Best score 0.86

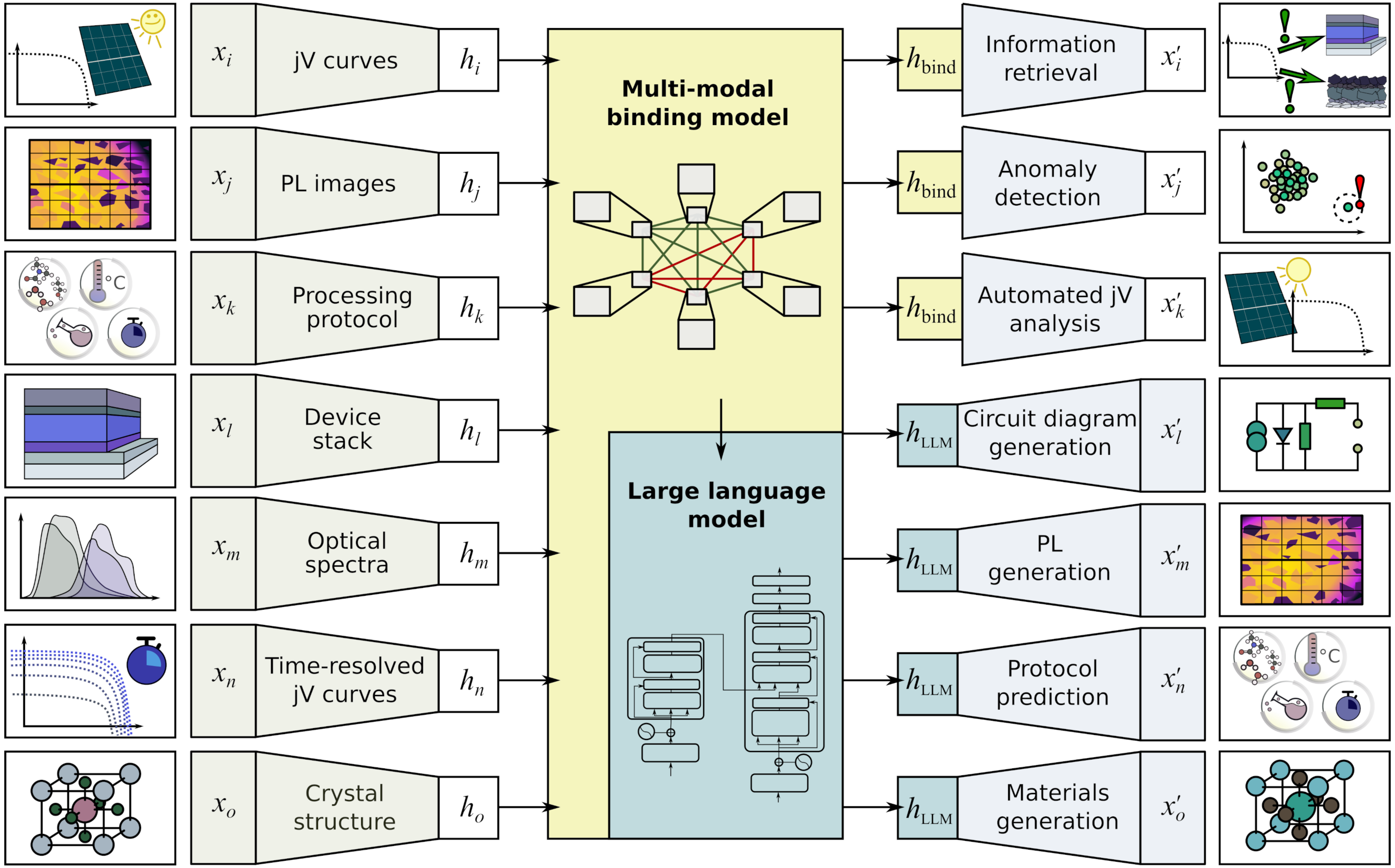
Spectrum To Structure



Spectrum To Structure



Multimodal Embeddings for Photovoltaics



SOL-AI (Helmholtz Foundation model initiative)

Testing the Limitations of Multimodal Models: MaCBench

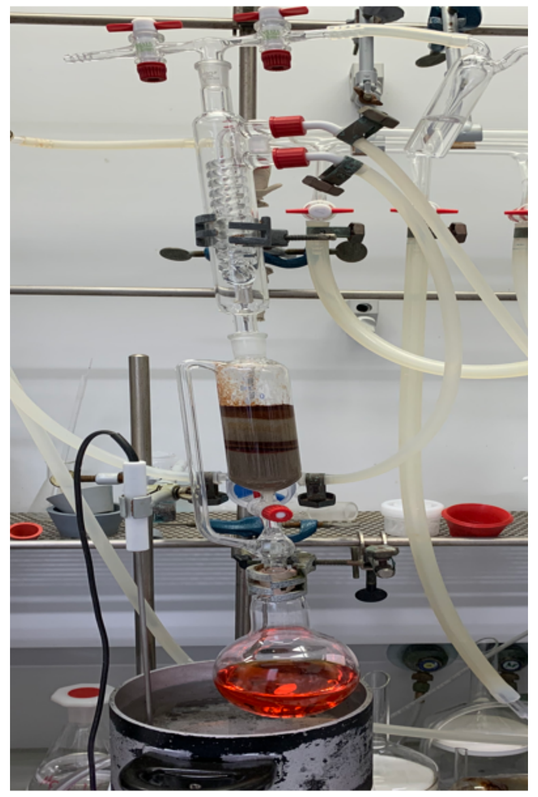
Nawaf Alampara

Martiño Ríos García

Mara Schilling-Wilhelmi

/w Anoop Krishnan (IIT Dehli)
NeurIPS AI4Mat 2024 Spotlight

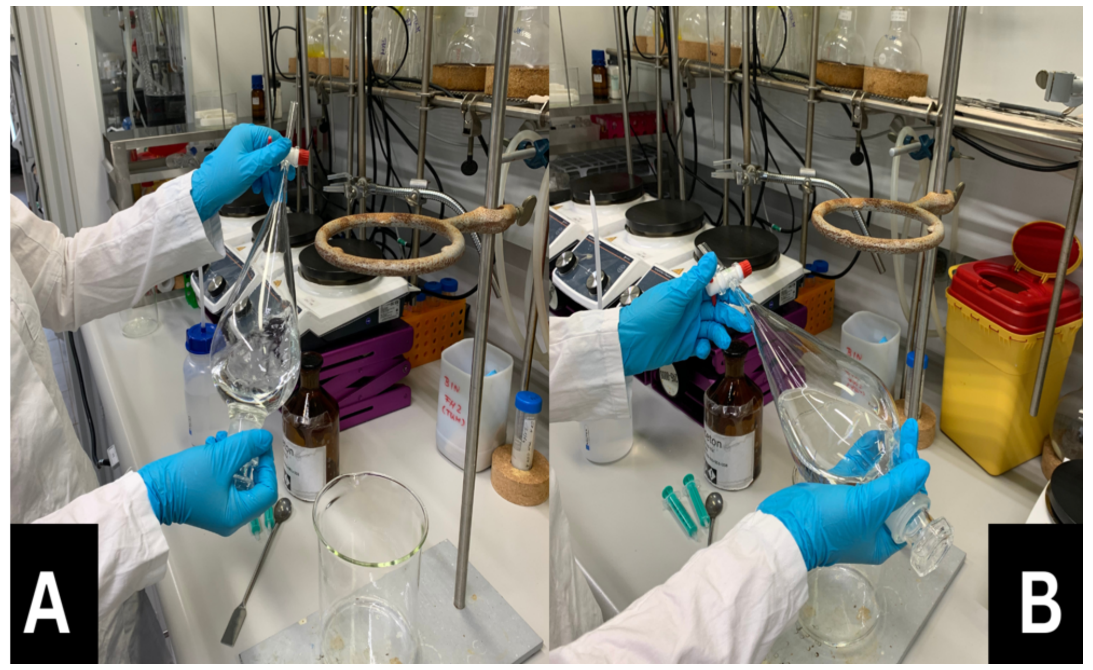
Lab scenarios



This is a photograph from a chemistry lab. The apparatus is in operation. What is the safety issue here?

Options:
- a) The setup should be placed perpendicular to the ground
- b) There is risk of the apparatus falling down
- c) The round bottom flask is not touching the oil bath
- d) Everything is correct, no safety risk

Lab scenario comparisons



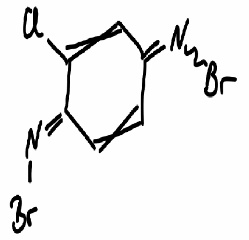
The chemist wants to separate two liquids with a separation funnel. Which is image, A or B, shows the correct protocol?

Glassware identification

In this photograph, what do you see?

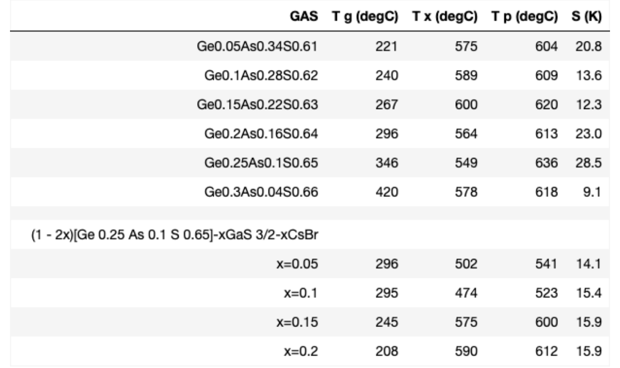
Options:
- a) Funnel
- b) Separatory Funnel
- c) Chromatography Column
- d) Burette

Handdrawn molecules



What is the SMILES of this molecule?
- a) C[Al](C)NC1=NC=CS1
- b) C(F)(F)(F)S(=O)NP(=S)(N)N
- c) C\1=C/C(=N\Br)/C(=C/C1=N/Br)Cl
- d) C/C(=C\C(C(C#C)[Si](C)(C)C(C)(C)C)(O)[Si](C)(C)C)/C#C

Table understanding

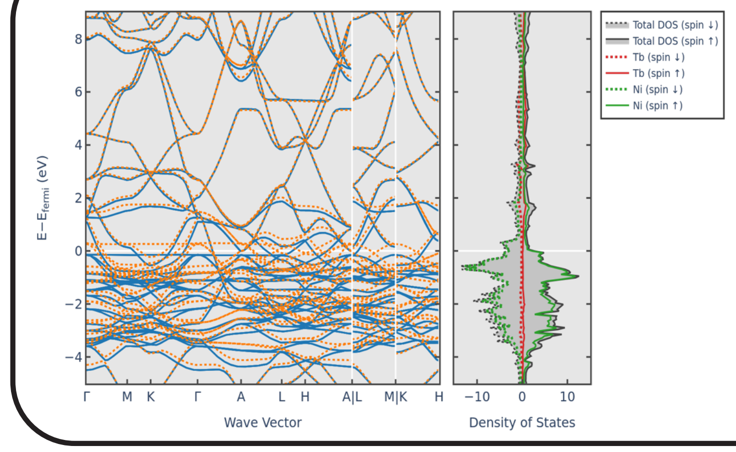


	GAS	T g (degC)	T x (degC)	T p (degC)	S (K)
	Ge0.05As0.34S0.61	221	575	604	20.8
	Ge0.1As0.28S0.62	240	589	609	13.6
	Ge0.15As0.22S0.63	267	600	620	12.3
	Ge0.2As0.16S0.64	296	564	613	23.0
	Ge0.25As0.1S0.65	346	549	636	28.5
	Ge0.3As0.04S0.66	420	578	618	9.1
	(1 - 2x)[Ge 0.25 As 0.1 S 0.65]-xGaS 3/2-xCsBr				
	x=0.05	296	502	541	14.1
	x=0.1	295	474	523	15.4
	x=0.15	245	575	600	15.9
	x=0.2	208	590	612	15.9

Does this table contain any material composition?

Do all the material compositions present in the table add up?

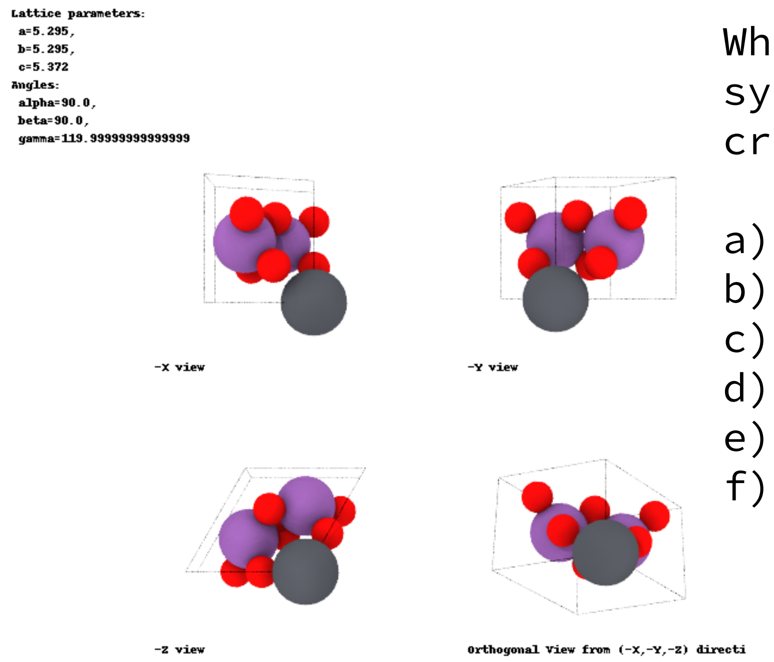
Band structures



What is the characteristic of the electronic structure in the image?

- a) metallic
- b) direct gap
- c) indirect gap

Reasoning about crystal structures

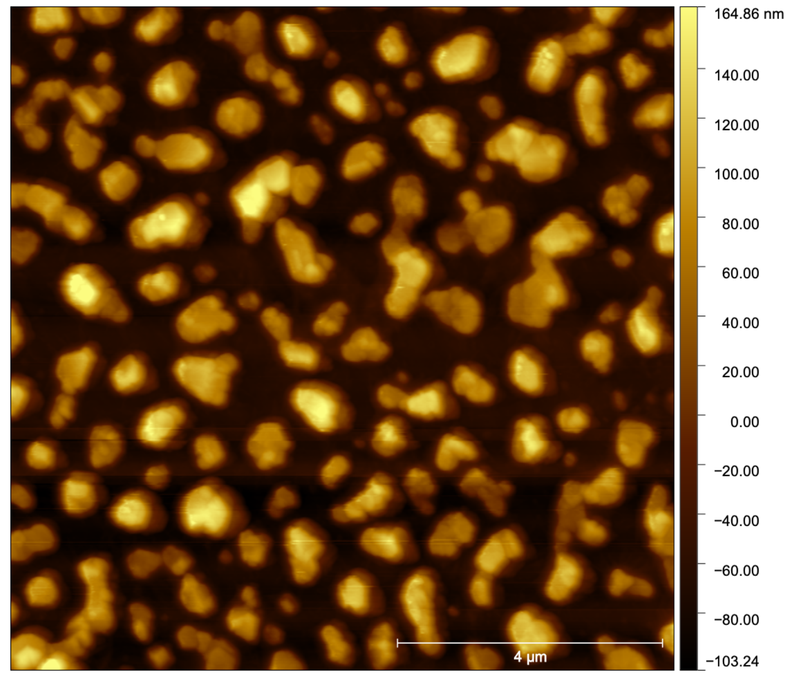


```
Lattice parameters:
 a=5.295,
 b=5.295,
 c=5.372
Angles:
 alpha=90.0,
 beta=90.0,
 gamma=119.99999999999999
```

What is the crystal system of this crystal structure?

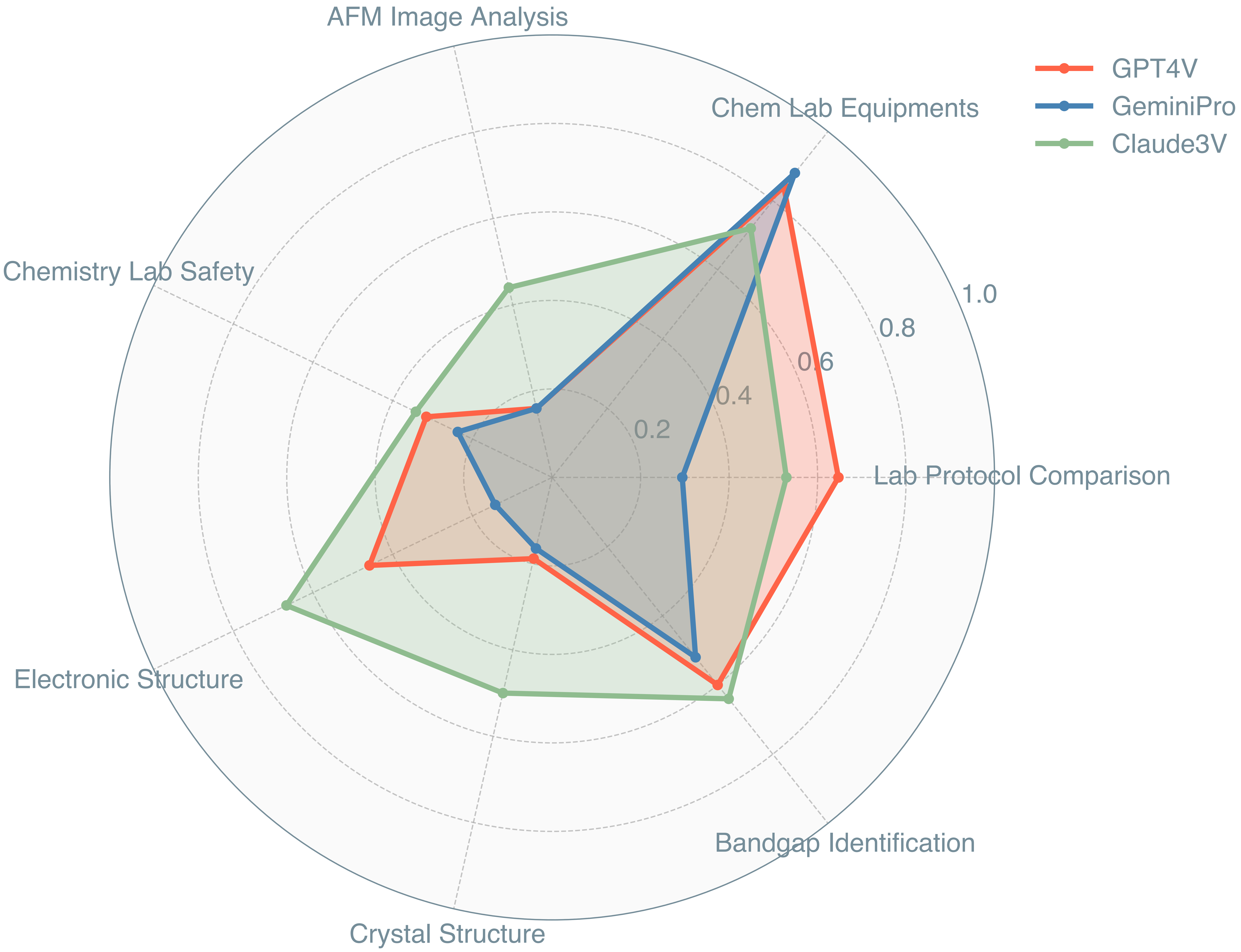
- a) monoclinic
- b) orthorhombic
- c) tetragonal
- d) hexagonal
- e) rhombohedral
- f) cubic

AFM images

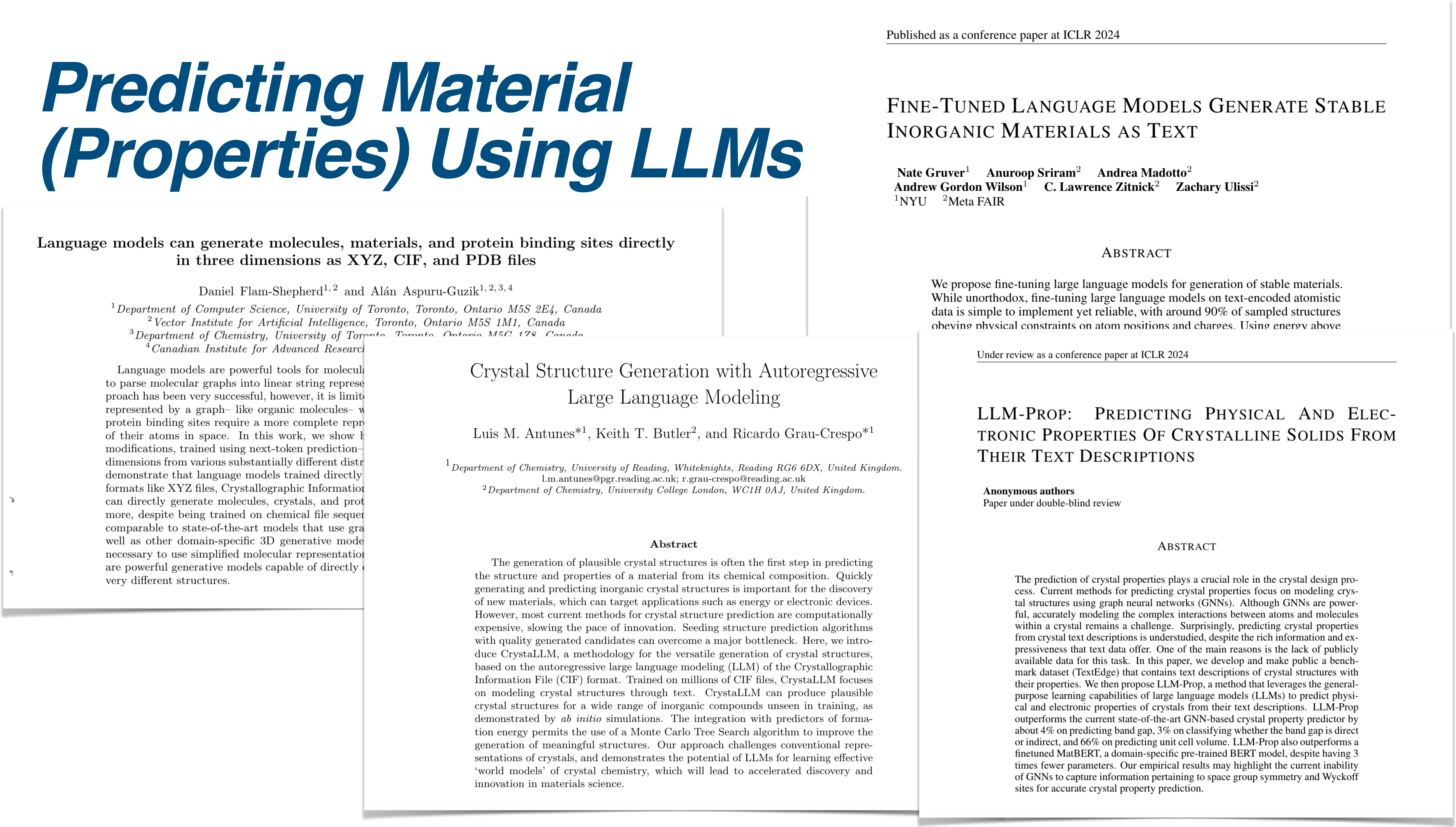


How many gold nanoislands can be observed in the image? Consider only those with a positive height as gold nanoislands.

Testing the Limitations of Multimodal Models



Predicting Material (Properties) Using LLMs



Language models can generate molecules, materials, and protein binding sites directly in three dimensions as XYZ, CIF, and PDB files

Daniel Flam-Shepherd[1, 2] and Alán Aspuru-Guzik[1, 2, 3, 4]
[1] *Department of Computer Science, University of Toronto, Toronto, Ontario M5S 2E4, Canada*
[2] *Vector Institute for Artificial Intelligence, Toronto, Ontario M5S 1M1, Canada*
[3] *Department of Chemistry, University of Toronto, Toronto, Ontario M5G 1Z8, Canada*
[4] *Canadian Institute for Advanced Researc*

Language models are powerful tools for molecul
to parse molecular graphs into linear string represe
proach has been very successful, however, it is limit
represented by a graph– like organic molecules– w
protein binding sites require a more complete repr
of their atoms in space. In this work, we show
modifications, trained using next-token prediction–
dimensions from various substantially different distr
demonstrate that language models trained directly
formats like XYZ files, Crystallographic Informatio
can directly generate molecules, crystals, and prot
more, despite being trained on chemical file seque
comparable to state-of-the-art models that use gra
well as other domain-specific 3D generative mode
necessary to use simplified molecular representatio
are powerful generative models capable of directly
very different structures.

Published as a conference paper at ICLR 2024

FINE-TUNED LANGUAGE MODELS GENERATE STABLE INORGANIC MATERIALS AS TEXT

Nate Gruver[1] **Anuroop Sriram**[2] **Andrea Madotto**[2]
Andrew Gordon Wilson[1] **C. Lawrence Zitnick**[2] **Zachary Ulissi**[2]
[1]NYU [2]Meta FAIR

ABSTRACT

We propose fine-tuning large language models for generation of stable materials. While unorthodox, fine-tuning large language models on text-encoded atomistic data is simple to implement yet reliable, with around 90% of sampled structures obeying physical constraints on atom positions and charges. Using energy above

---

Crystal Structure Generation with Autoregressive Large Language Modeling

Luis M. Antunes*[1], Keith T. Butler[2], and Ricardo Grau-Crespo*[1]

[1] *Department of Chemistry, University of Reading, Whiteknights, Reading RG6 6DX, United Kingdom.*
l.m.antunes@pgr.reading.ac.uk; r.grau-crespo@reading.ac.uk
[2] *Department of Chemistry, University College London, WC1H 0AJ, United Kingdom.*

Abstract

The generation of plausible crystal structures is often the first step in predicting the structure and properties of a material from its chemical composition. Quickly generating and predicting inorganic crystal structures is important for the discovery of new materials, which can target applications such as energy or electronic devices. However, most current methods for crystal structure prediction are computationally expensive, slowing the pace of innovation. Seeding structure prediction algorithms with quality generated candidates can overcome a major bottleneck. Here, we introduce CrystaLLM, a methodology for the versatile generation of crystal structures, based on the autoregressive large language modeling (LLM) of the Crystallographic Information File (CIF) format. Trained on millions of CIF files, CrystaLLM focuses on modeling crystal structures through text. CrystaLLM can produce plausible crystal structures for a wide range of inorganic compounds unseen in training, as demonstrated by *ab initio* simulations. The integration with predictors of formation energy permits the use of a Monte Carlo Tree Search algorithm to improve the generation of meaningful structures. Our approach challenges conventional representations of crystals, and demonstrates the potential of LLMs for learning effective 'world models' of crystal chemistry, which will lead to accelerated discovery and innovation in materials science.

Under review as a conference paper at ICLR 2024

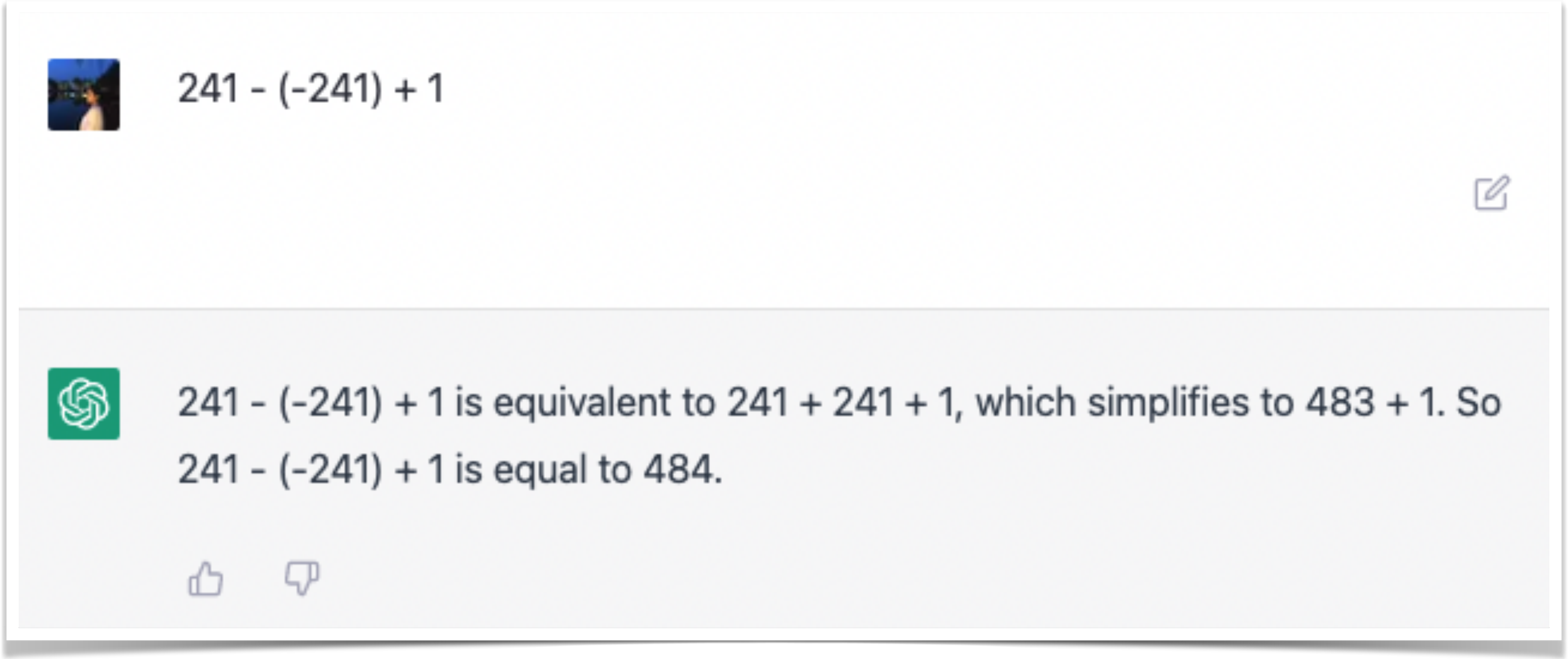
LLM-PROP: PREDICTING PHYSICAL AND ELECTRONIC PROPERTIES OF CRYSTALLINE SOLIDS FROM THEIR TEXT DESCRIPTIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

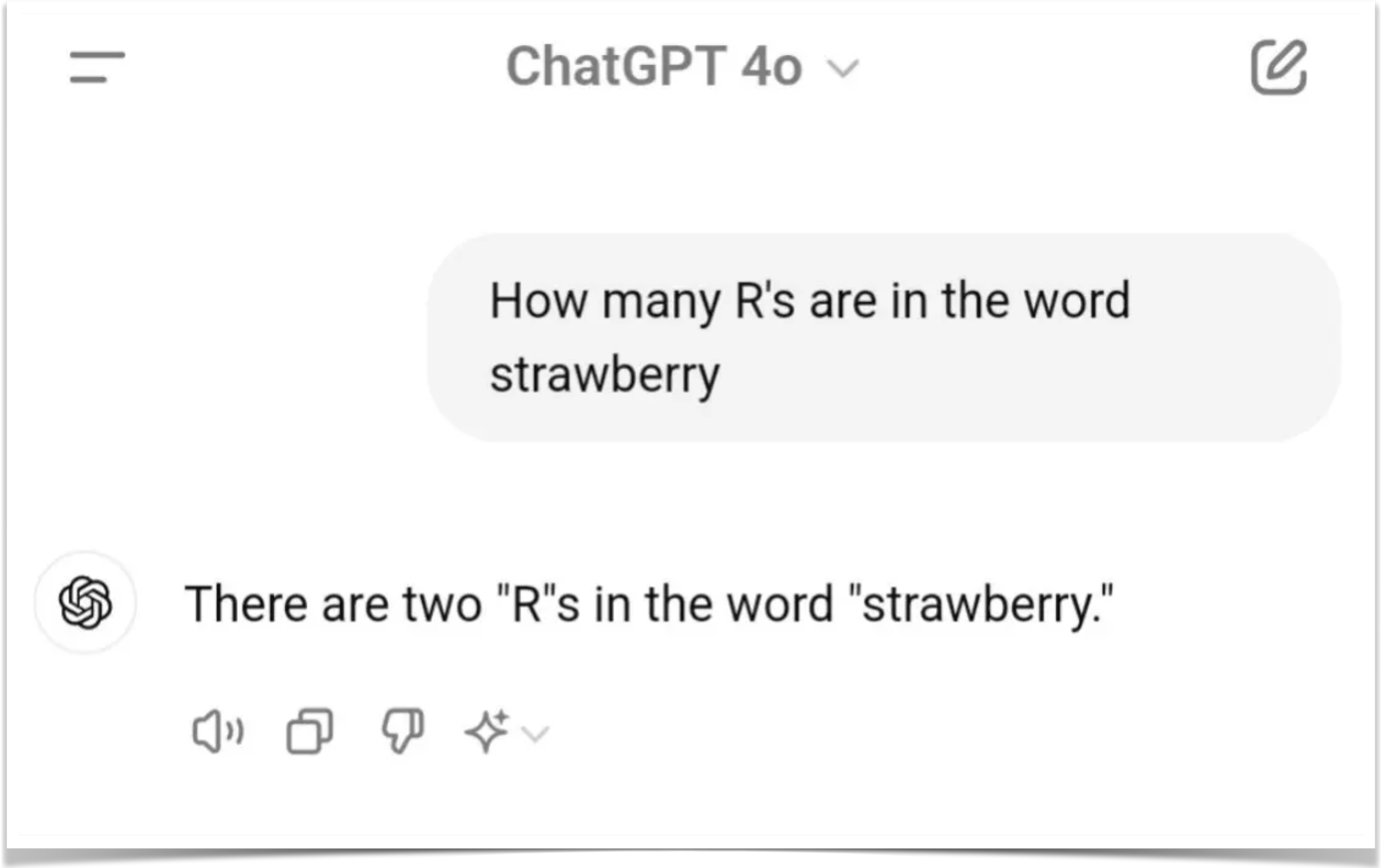
The prediction of crystal properties plays a crucial role in the crystal design process. Current methods for predicting crystal properties focus on modeling crystal structures using graph neural networks (GNNs). Although GNNs are powerful, accurately modeling the complex interactions between atoms and molecules within a crystal remains a challenge. Surprisingly, predicting crystal properties from crystal text descriptions is understudied, despite the rich information and expressiveness that text data offer. One of the main reasons is the lack of publicly available data for this task. In this paper, we develop and make public a benchmark dataset (TextEdge) that contains text descriptions of crystal structures with their properties. We then propose LLM-Prop, a method that leverages the general-purpose learning capabilities of large language models (LLMs) to predict physical and electronic properties of crystals from their text descriptions. LLM-Prop outperforms the current state-of-the-art GNN-based crystal property predictor by about 4% on predicting band gap, 3% on classifying whether the band gap is direct or indirect, and 66% on predicting unit cell volume. LLM-Prop also outperforms a finetuned MatBERT, a domain-specific pre-trained BERT model, despite having 3 times fewer parameters. Our empirical results may highlight the current inability of GNNs to capture information pertaining to space group symmetry and Wyckoff sites for accurate crystal property prediction.

But They Cannot Even Do Math!



241 - (-241) + 1

241 - (-241) + 1 is equivalent to 241 + 241 + 1, which simplifies to 483 + 1. So 241 - (-241) + 1 is equal to 484.

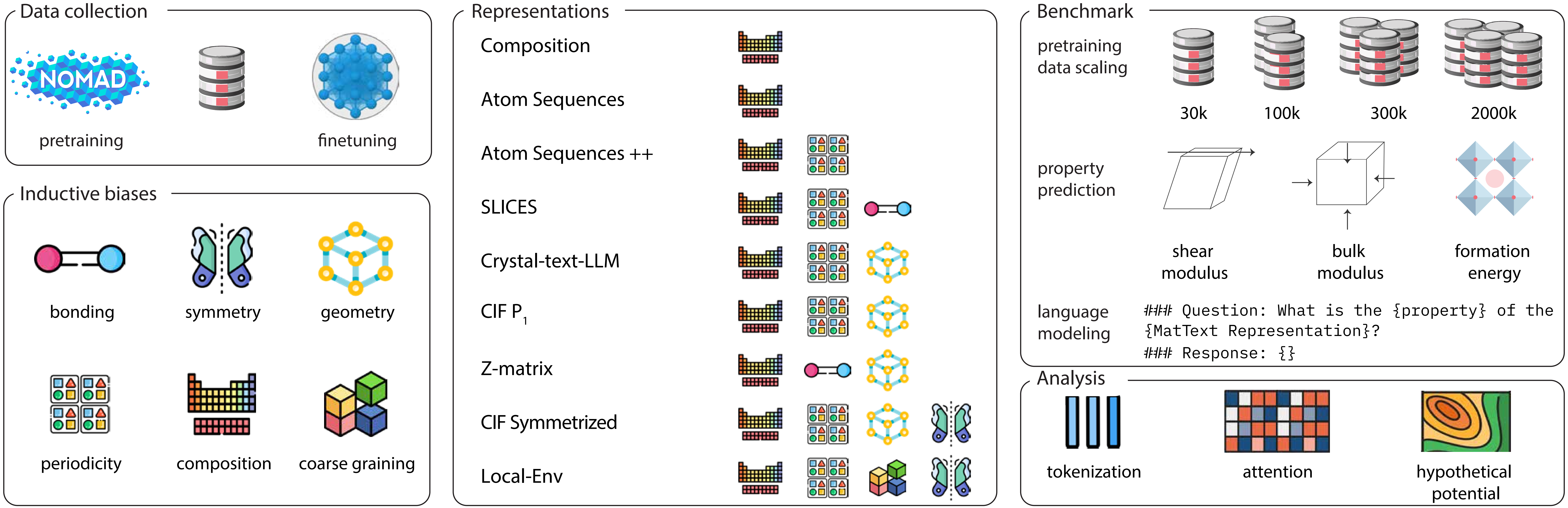


How many R's are in the word strawberry

There are two "R"s in the word "strawberry."

https://community.openai.com/t/chatgpt-simple-math-calculation-mistake/62780
https://prompt.16x.engineer/blog/why-chatgpt-cant-count-rs-in-strawberry

MATText



Data collection

pretraining

finetuning

Inductive biases

bonding

symmetry

geometry

periodicity

composition

coarse graining

Representations

Composition

Atom Sequences

Atom Sequences ++

SLICES

Crystal-text-LLM

CIF P_1

Z-matrix

CIF Symmetrized

Local-Env

Benchmark

pretraining data scaling

30k

100k

300k

2000k

property prediction

shear modulus

bulk modulus

formation energy

language modeling

```
### Question: What is the {property} of the
{MatText Representation}?
### Response: {}
```

Analysis

tokenization

attention

hypothetical potential

Santiago Miret
(Intel Labs)

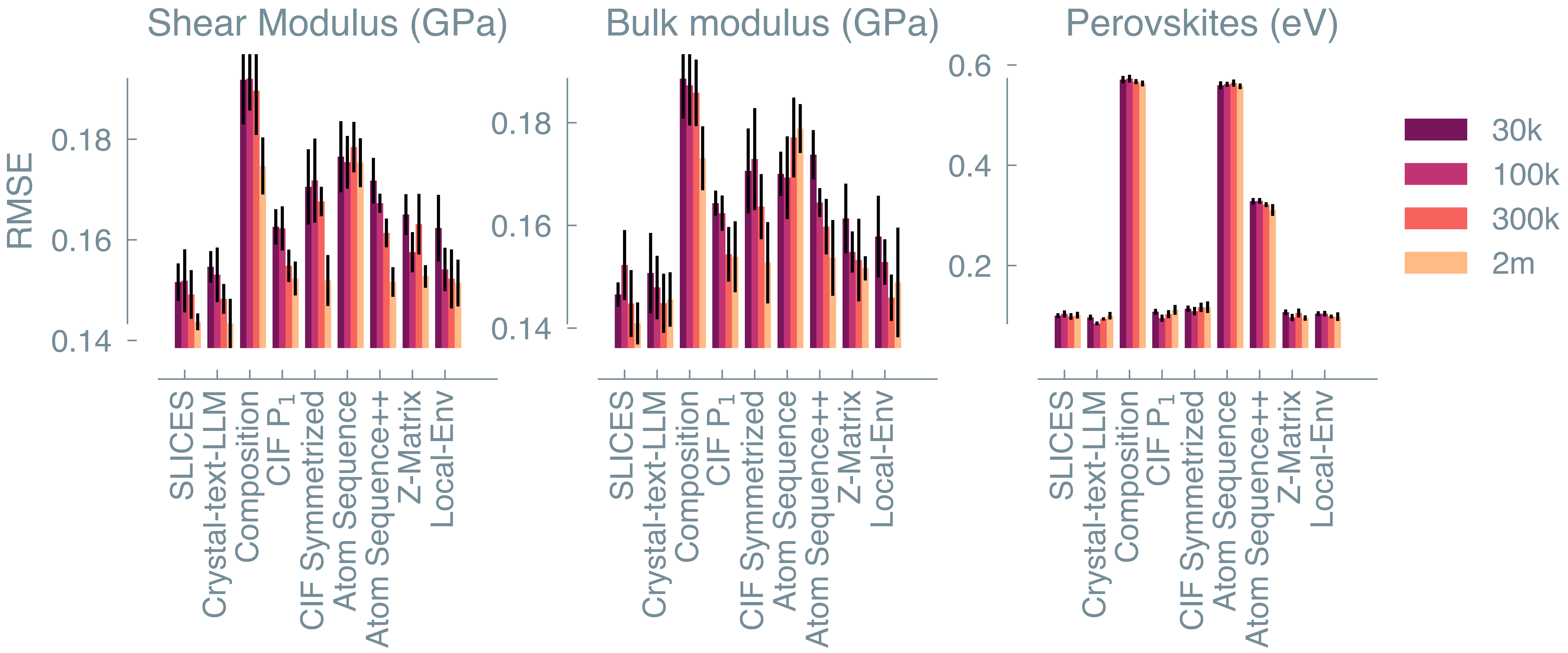
Nawaf Alampara

arXiv 2406.17295
AI4Mat 2024 Spotlight

Encoding Inductive Biases

	Stoichiometry	Bonding	Geometry	Symmetry	Periodicity	Coarse Graining
Composition	✓					
SLICES	✓	✓				
CIF P_1	✓		✓		✓	
Crystal-text-LLM	✓		✓			
			Novel Representations			
Atom Sequences	✓					
Atom Sequences++	✓				✓	
CIF Symmetrized	✓		✓	✓	✓	
Z-Matrix	✓	✓	✓			
Local-Env	✓	✓		✓		✓

Geometry and Scale Do Not Help!

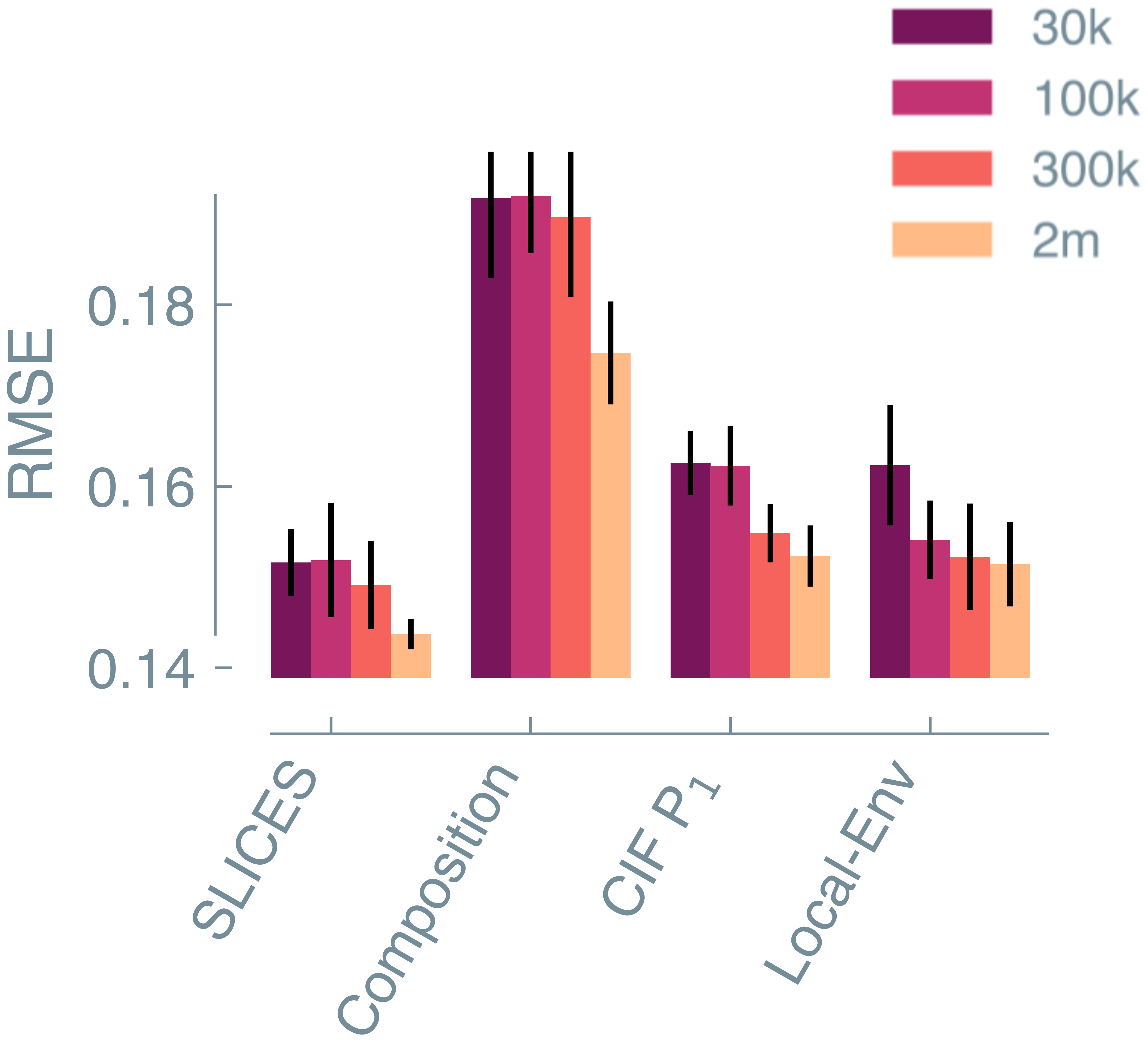


Geometry and Scale Do Not Help!

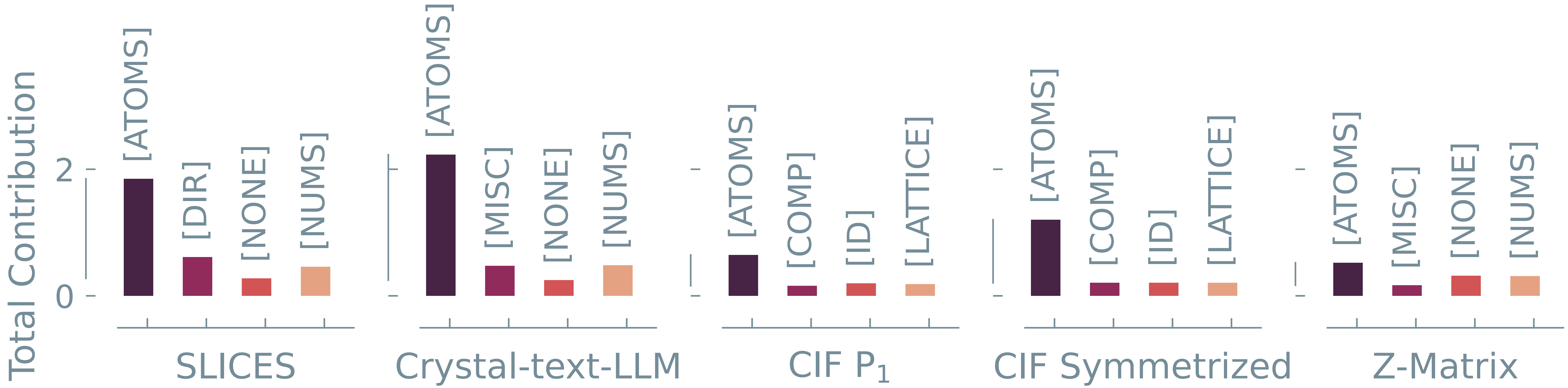
There is some small improvement with scale

But does not beat baselines

Geometric information is not effectively leveraged



Little Attention Paid to Numbers

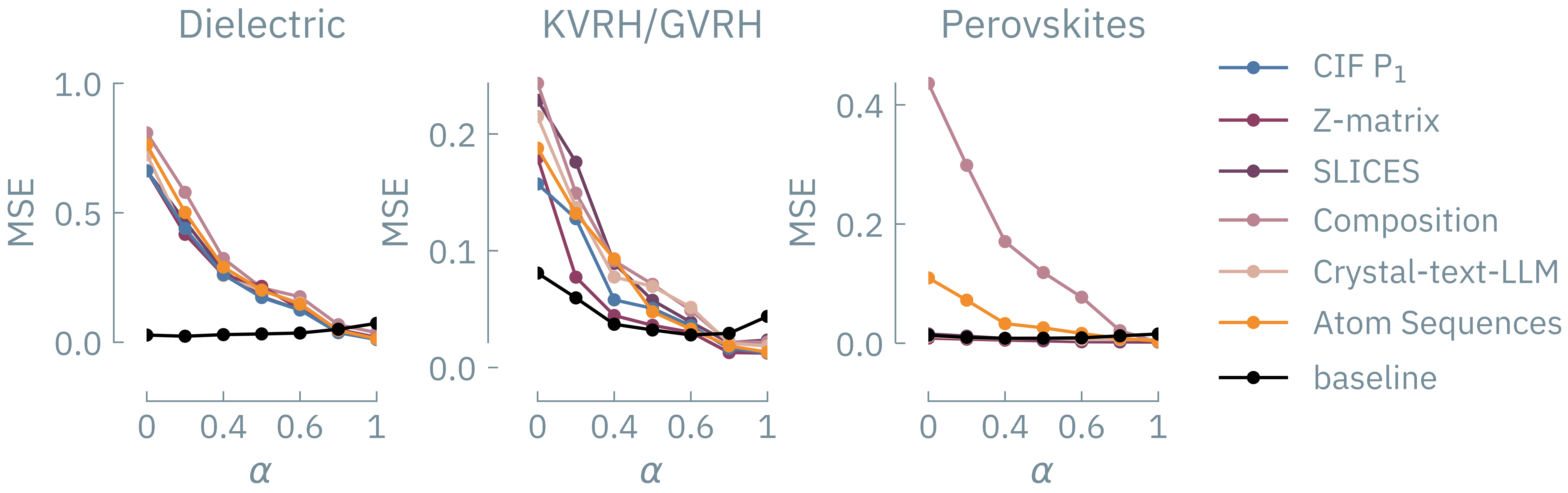


Overall contribution over layers per token class

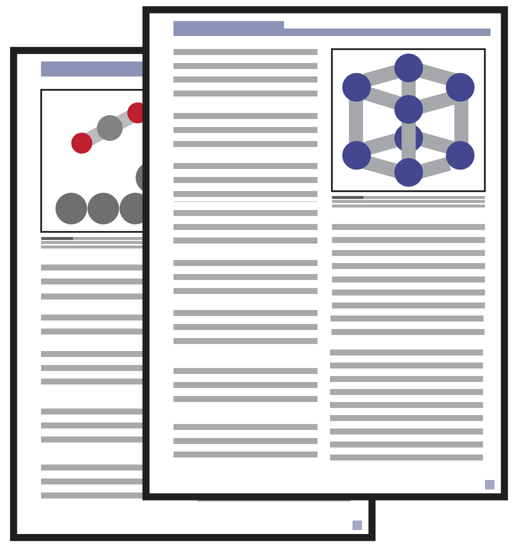
Most attention is on atoms

Separating Data From Model Effects

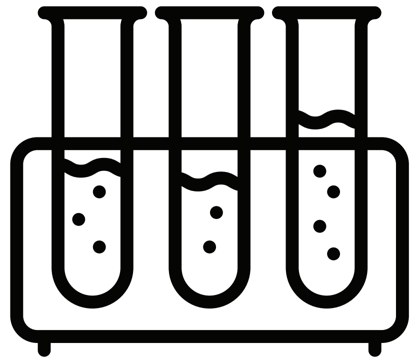
$$E = \alpha E_{\text{comp}} + (1-\alpha)E_{\text{pos}} = \underbrace{\sum_{k=1}^{k} w_k n_k}_{E_{\text{comp}}} + \underbrace{\sum_{i=1}^{N} V\left(\mathbf{r}_i\right) + \sum_{i=1}^{N}\sum_{j\in\mathcal{N}(i)} V\left(\left|\mathbf{r}_i-\mathbf{r}_j\right|\right)}_{E_{\text{pos}}}, \quad \alpha \in (0,1)$$



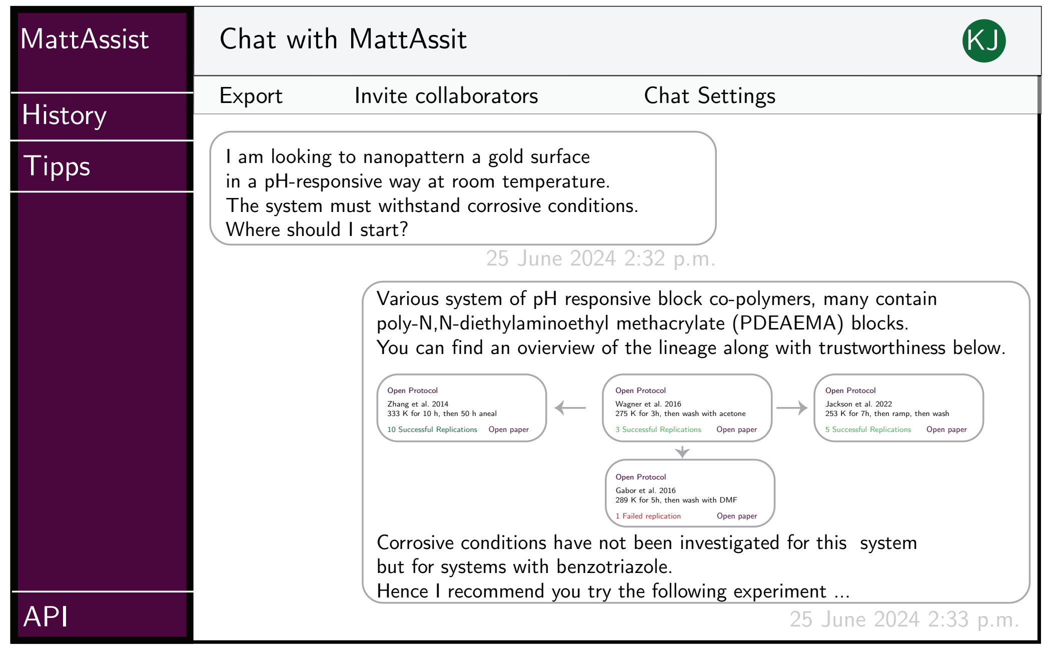
Approach and Techniques



data capture
and mining



collaborations
with experimental
groups and engineers



I am looking to nanopattern a gold surface
in a pH-responsive way at room temperature.
The system must withstand corrosive conditions.
Where should I start?

25 June 2024 2:32 p.m.

Various system of pH responsive block co-polymers, many contain
poly-N,N-diethylaminoethyl methacrylate (PDEAEMA) blocks.
You can find an ovierview of the lineage along with trustworthiness below.

Open Protocol
Zhang et al. 2014
333 K for 10 h, then 50 h aneal
10 Successful Replications Open paper

Open Protocol
Wagner et al. 2016
275 K for 3h, then wash with acetone
3 Successful Replications Open paper

Open Protocol
Jackson et al. 2022
253 K for 7h, then ramp, then wash
5 Successful Replications Open paper

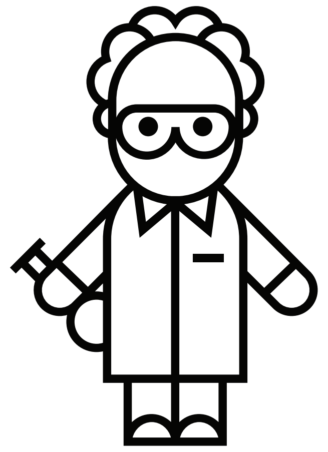
Open Protocol
Gabor et al. 2016
289 K for 5h, then wash with DMF
1 Failed replication Open paper

Corrosive conditions have not been investigated for this system
but for systems with benzotriazole.
Hence I recommend you try the following experiment ...

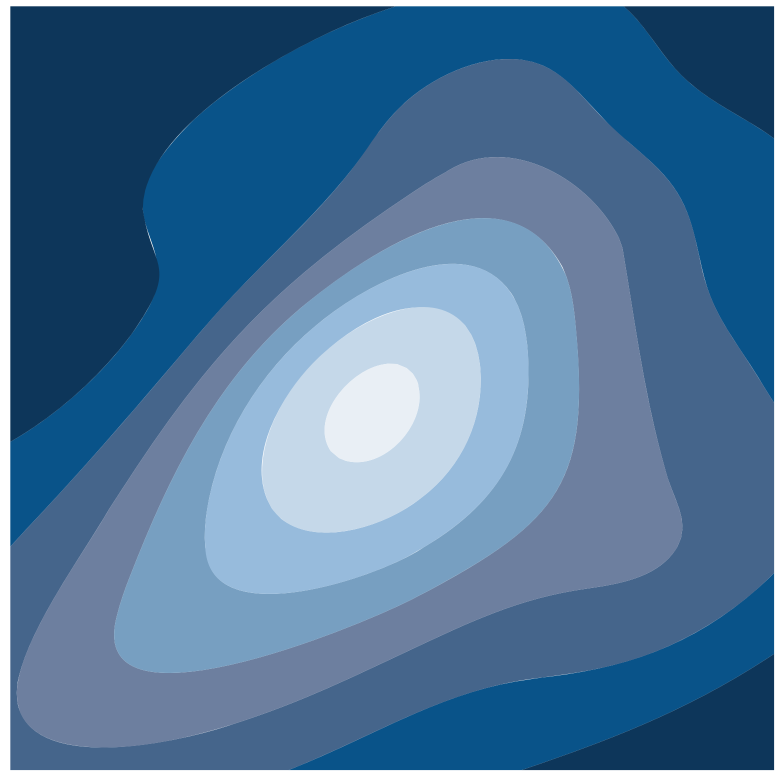
25 June 2024 2:33 p.m.

driven with autonomous agents

data
management



human-in-the-loop
active learning



generative modeling

I’m hiring!

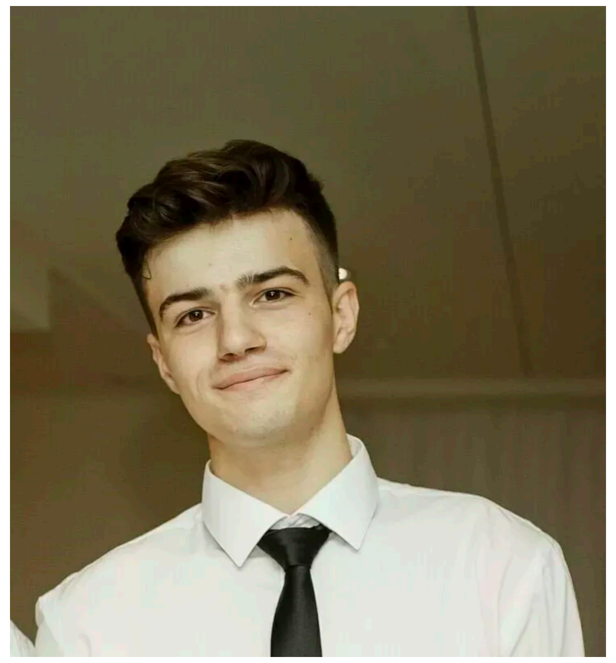
For Ph.D., PostDoc, MSc., internship, etc. in ML for materials/ chemistry (and RDM tools) contact join@lamalab.org

FRIEDRICH-SCHILLER-UNIVERSITÄT JENA

HZB Helmholtz Zentrum Berlin

Carl Zeiss Stiftung

Current Ph.D. Students


Adrian Mirza

Nawaf Alampara

Sreekanth Kunchapu

Martiño Ríos García

Mara Schilling-Wilhelmi

Ali Asghar (Sadra) Aghajani

Anagha Aneesh
(Fullbright Scholar)

Msc. Students

Reza Aliakbari

Abdelrahman Ibrahim

HIPOLE
Helmholtz Institute for Polymers in Energy Applications

CEEC JENA

JCSM
Jena Center for Soft Matter

HZB Helmholtz Zentrum Berlin

Carl Zeiss Stiftung

intel

MERCK

FAIRmat

Develop Materials That Work in the Real World With our Digital Assistants