

Revolutionizing Glass Development through AI and Large Language Models: from Laboratory to Industry

N. M. Anoop Krishnan

Department of Civil
Engineering
School of Artificial Intelligence
(Joint Appt.)



Indian Institute of Technology Delhi





Collaborators

Alumni (PhD)

- Ravinder
- Rajesh Kumar
- S Kasimuthumaniyan
- Sourav Sahoo
- Mohd Zaki

Graduate Students

- Tanu Pittie
- Ashish Yadav
- Shweta Rani Keshri
- Suresh Bishnoi
- Sajid Mannan
- Sheikh Junaid Fayaz
- Hargun Grover
- Dushyant Soni
- Sanchit Bedi
- Indrajeet Mandal
- Kausik Hira

Graduate Students

- Harshit Bisht
- Vinay Kumar
- Karn Tiwari
- Parv Pratap Singh
- Akash Kumar Singh

Collaborators

- Mathieu Bauchy
- Morten Smedskjaer
- Gaurav Sant
- Nitya Nand Gosvami
- Hariprasad Kodamana
- Sumanta Das
- Bu Wang
- Sayan Ranu
- Tanishq Gupta
- Santiago Miret
- Prathosh A P

Collaborators (contd.)

- John Mauro
- Amarnath R Allu
- Yann Le Pape
- Gideon Lyngdoh
- Kai Yang
- Christian Hoover
- Mengyi Wang
- Sudheer Ganishetti
- Han Liu
- Vineeth Venugopal
- Vaibhav Bihani
- Manish Agarwal
- Mausam
- K Annapurna
- Sujith Mangalathu
- Souvik Chakraborty
- Lothar Wondraczek
- Kevin Jablonka

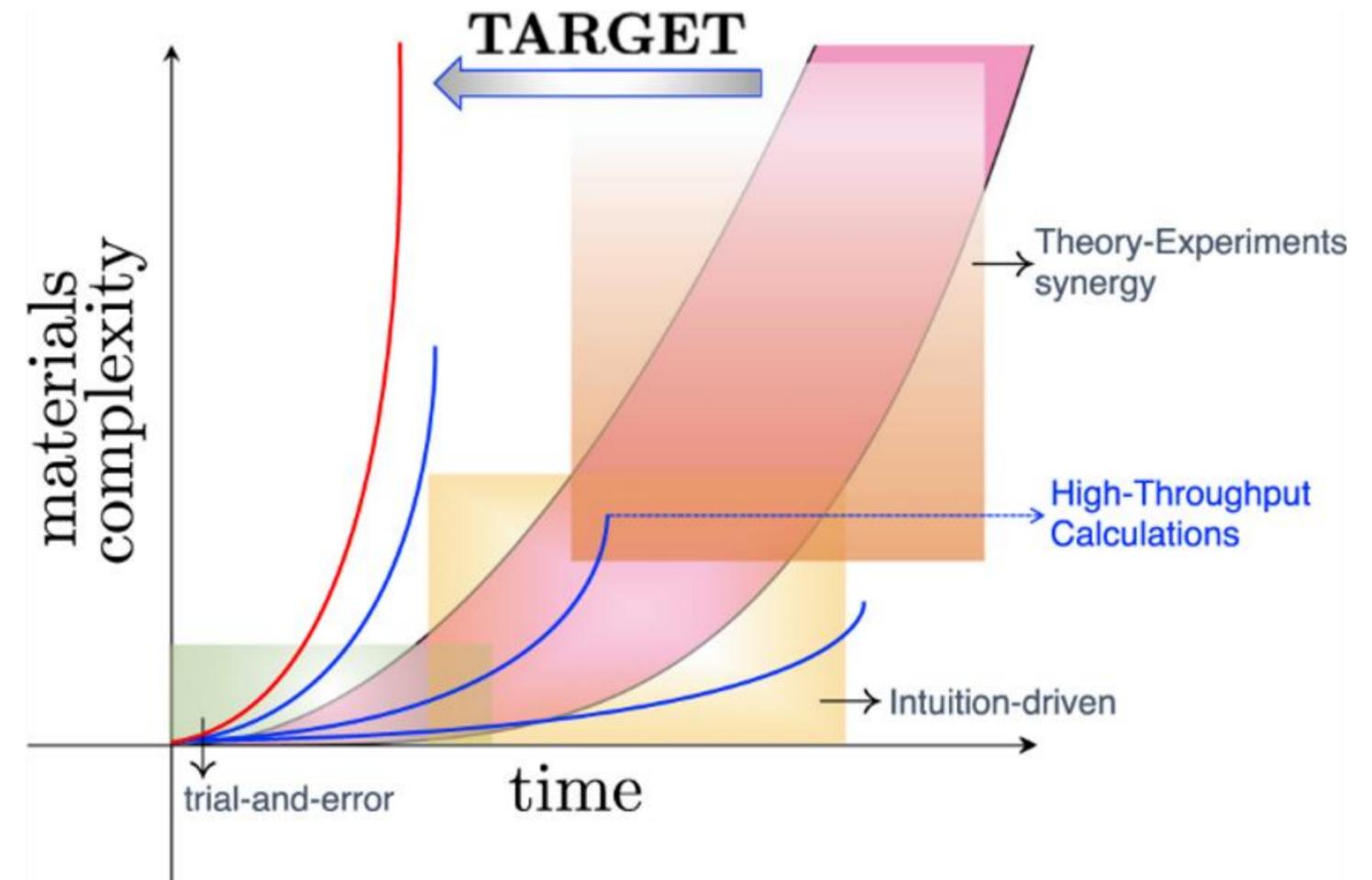
Collaborators...

- Edgar Zanotto
- Amreen Jan
- Pratik Bhaskar
- Sri Harsha Kota
- Stephane Gin
- Isabella Pignatelli
- Naresh Bhatnagar
- Matthieu Micoulaut
- Utkarsh Pratiush
- Laurent Pilon
- Gabriel Falzone
- Dhruvil Sheth
- Sushmit Goyal
- Eswaraiah Varrla
- Daniel Cassar
- Sahil Manchanda
- ...

Glass Discovery

Motivation

- Glasses: back-bone of society
- Glass discovery: empirical, uneconomical with a ***design-to-deploy period of 20-30 years***
- UN sustainable development goals: 10/17 can potentially be addressed through glass innovation: IYG
- Areas: energy, infrastructure, healthcare, CO₂ emission, space sector

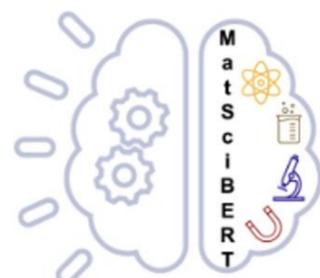
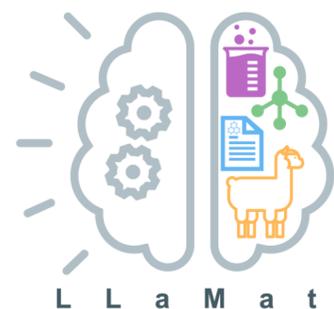


Lookman et al., npj Comp. Mat., 2019

Alternate approach: Data-driven Modeling



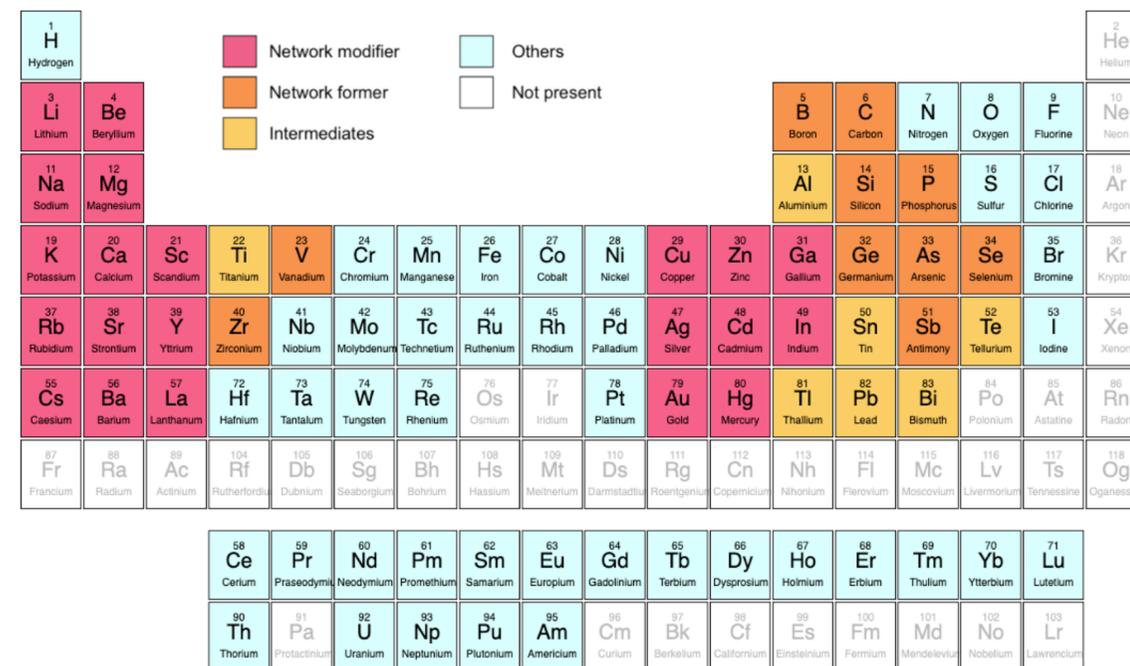
AI for Glass Discovery



MatSciBERT-
Material
science aware
Language
Model

Scientific NLP

Information extraction
and glass prediction

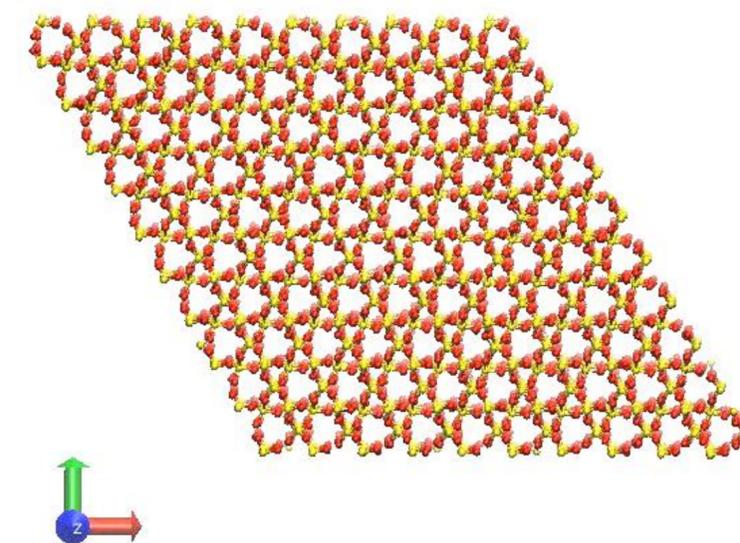


Industrial Scale Manufacturing

AI-driven process
optimization at
industrial scale

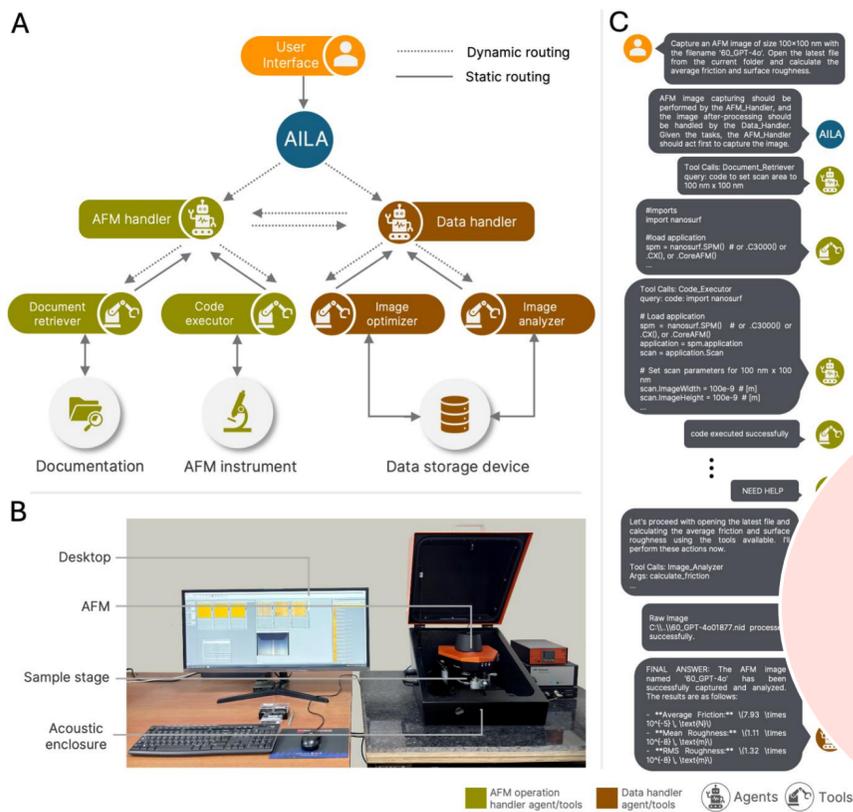
Materials simulation

AI-driven
simulations and
analysis



Self-driving labs

Synthesis and
characterization

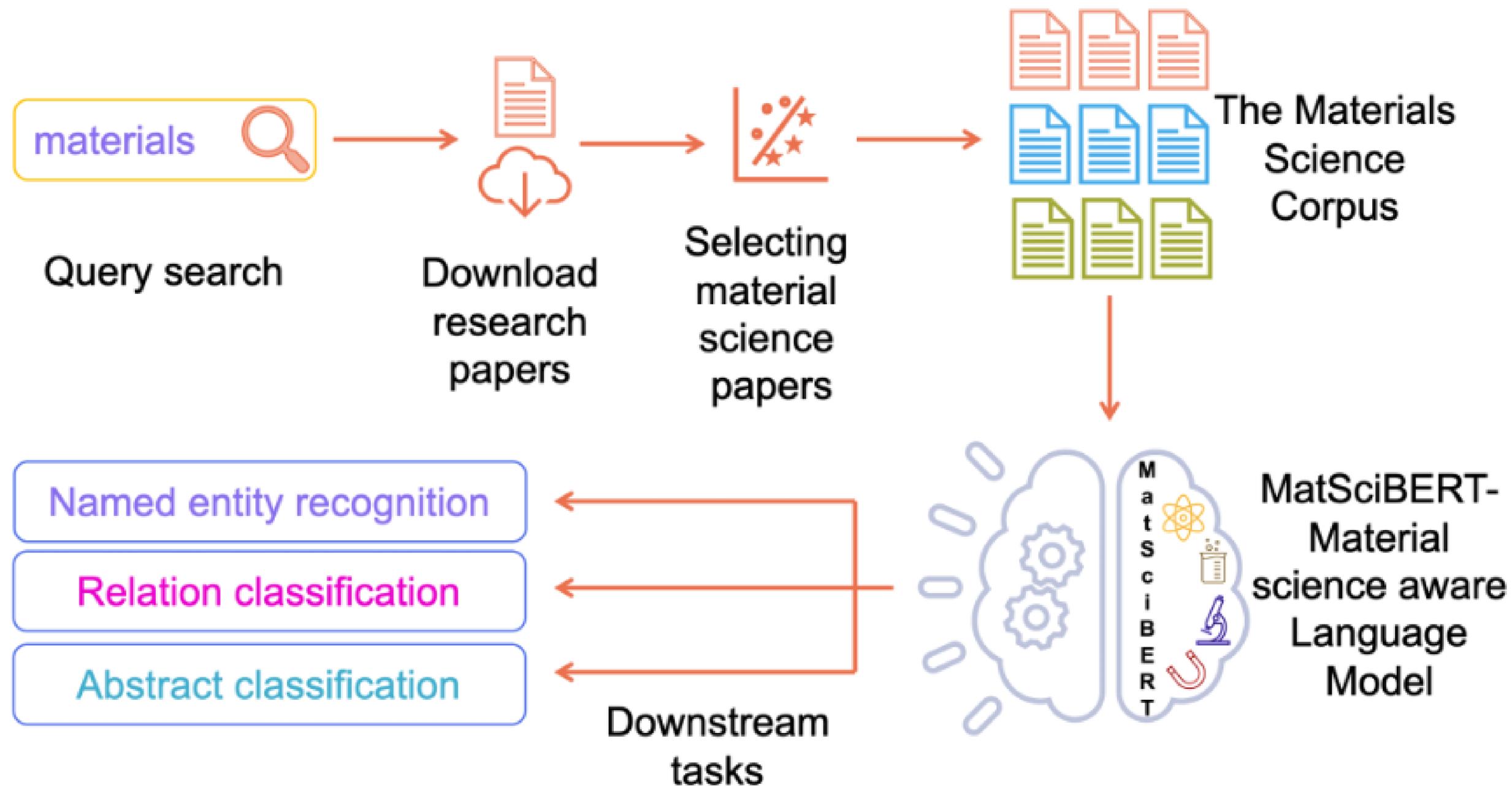




MatSciBERT

First Materials-aware language model

§ BERT based
Language Model / Text
Encoder pre-trained for
15 days on Materials
Science corpus
consisting of **~285M**
words



Structure factor database

Image analysis

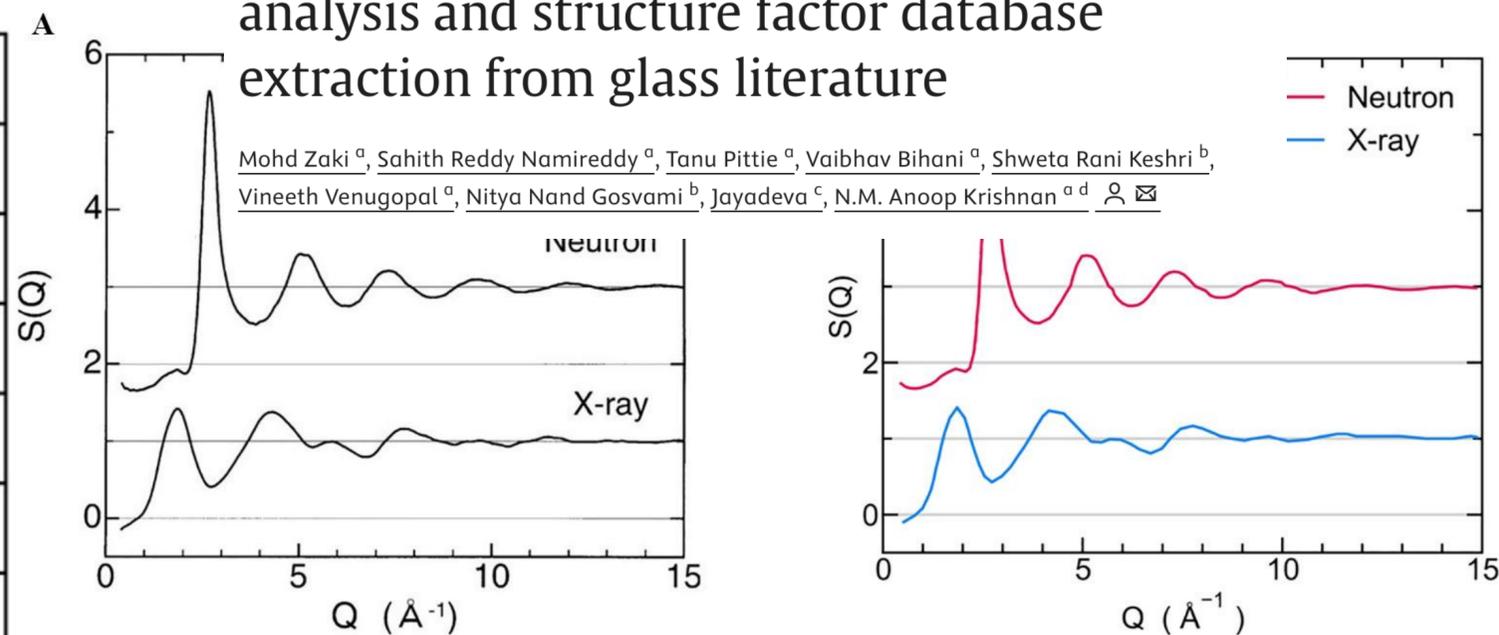


Journal of Non-Crystalline Solids: X
Volume 15, September 2022, 100103

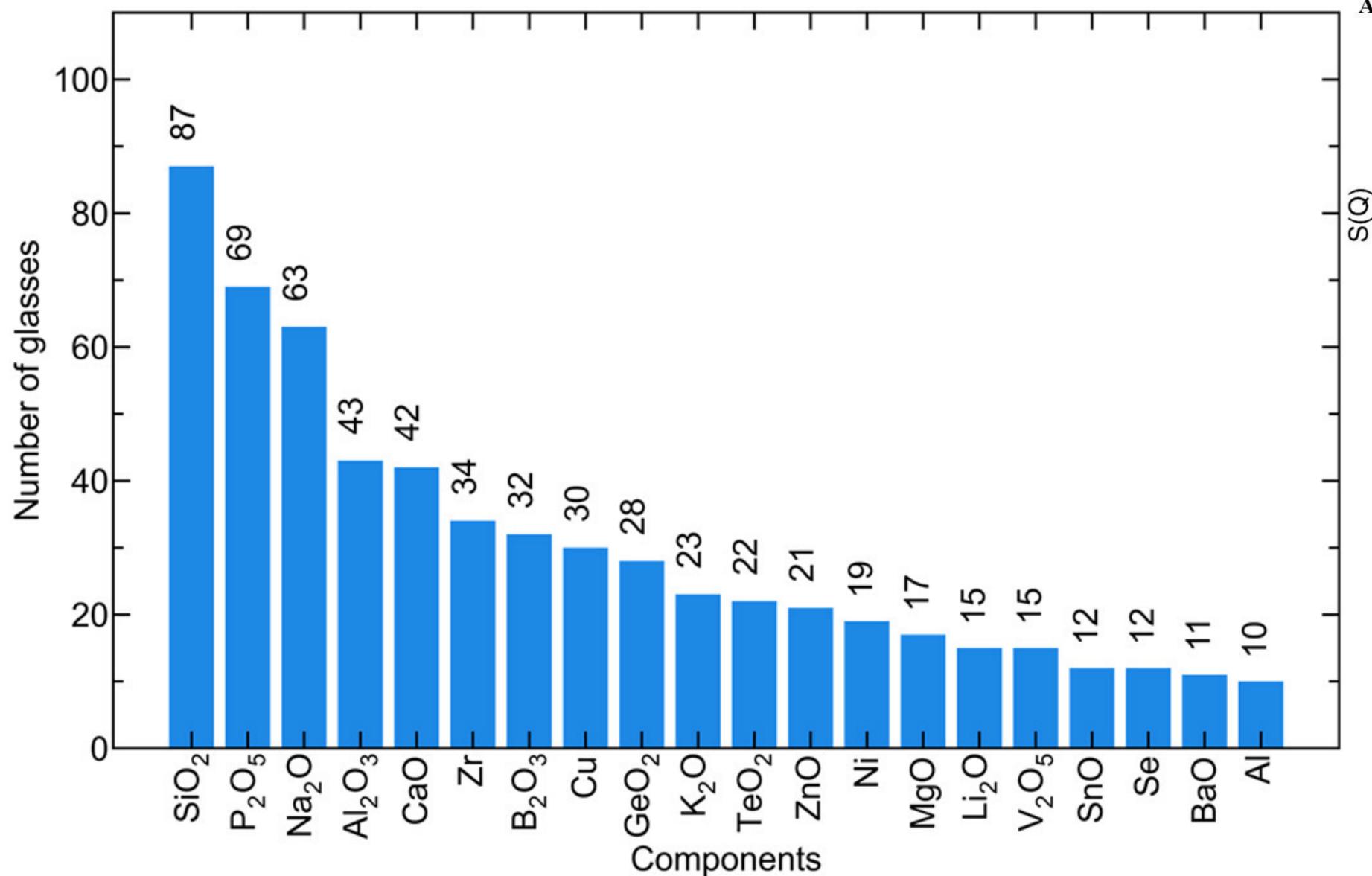


Natural language processing-guided meta-analysis and structure factor database extraction from glass literature

Mohd Zaki^a, Sahith Reddy Namireddy^a, Tanu Pittie^a, Vaibhav Bihani^a, Shweta Rani Keshri^b, Vineeth Venugopal^a, Nitya Nand Gosvami^b, Jayadeva^c, N.M. Anoop Krishnan^{a,d} ✉



- 421 structure factors
- 302 glass compositions
- 22 unary, 142 binary, and 103 ternary glasses
- Both experimental and simulated

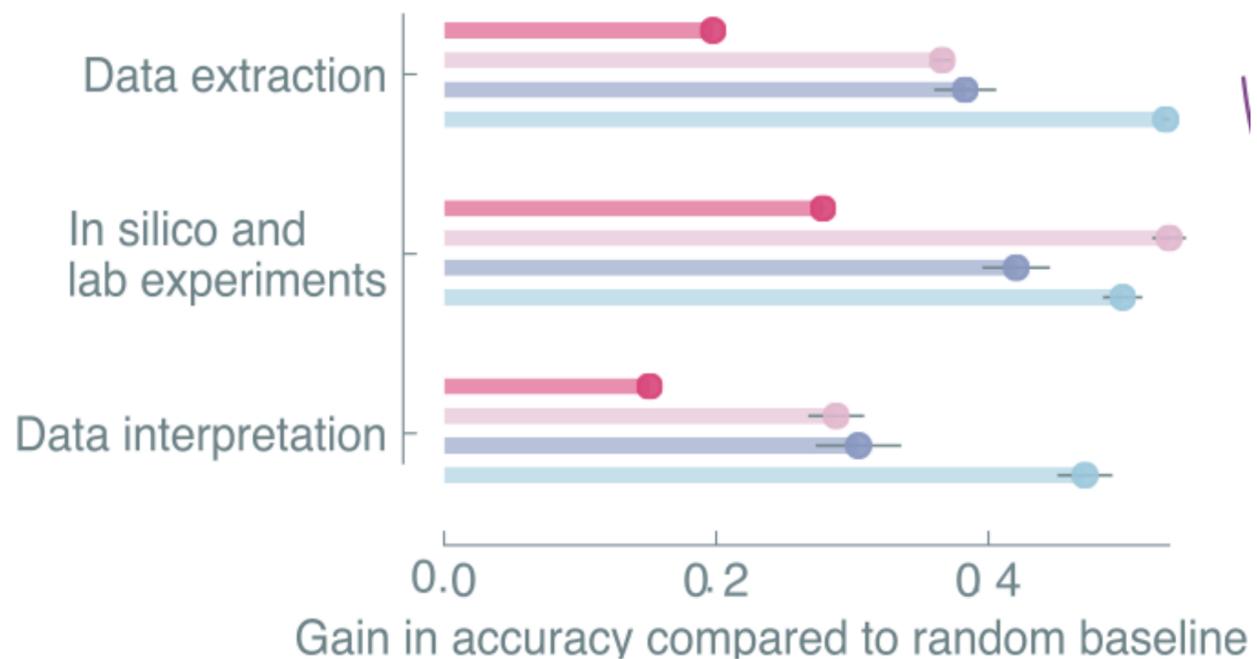


MaCBench

Benchmarking multimodal models

Models

- Llama 3.2 90B Vision
- Gemini-1.5-Pro
- GPT-4o
- Claude-3.5-Sonnet

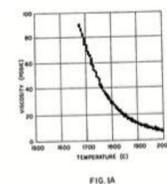


Example questions

Tables
How many material compositions are present in the table?

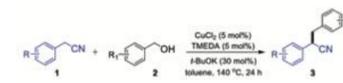
Material ID	Composition	Properties
1
2
3
4
5
6
7
8
9
10

Plots
What is the viscosity (in poise) of the fiber at a temperature of 2073 K?



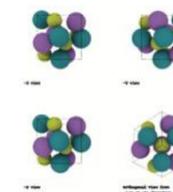
Organic reactions and molecules

Which solvent was used in the reaction?



In silico experiments

What is the volume of the unit cell of this crystal in Å³?

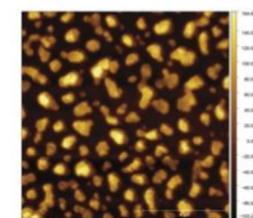


Laboratory experiments

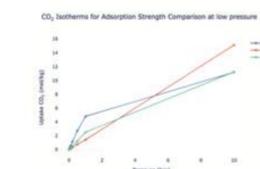
The chemist wants to run a volatile chemical reaction overnight. What statements about the picture are true?



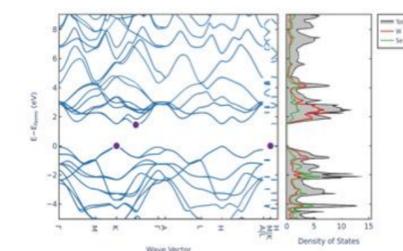
AFM characterization
What is the maximum height of the gold nanoinlands in nanometers?



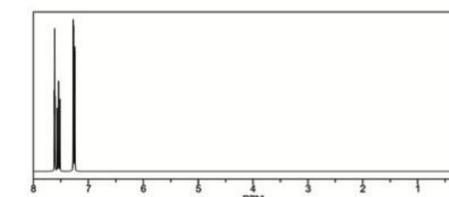
MOF characterization
What is the correct ascending order of adsorption strength at low pressure (0.01 bar) in mol/kg for CO₂ in the given MOF isotherms simulated at 298 K?



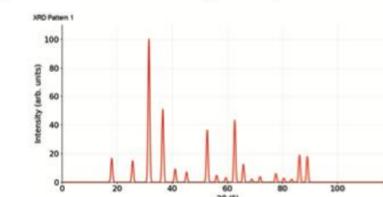
Bandgap analysis
Does the electronic structure in the following image show a bandgap?



Molecule analysis
The measured molecule is an aromatic 6-membered ring with 2 chloride substituents. Based on the NMR spectrum, how are the substituents located to each other?



XRD pattern analysis
What is the crystal structure type of the material based on following XRD pattern?





LLaMat

LLM for Materials

Pretraining



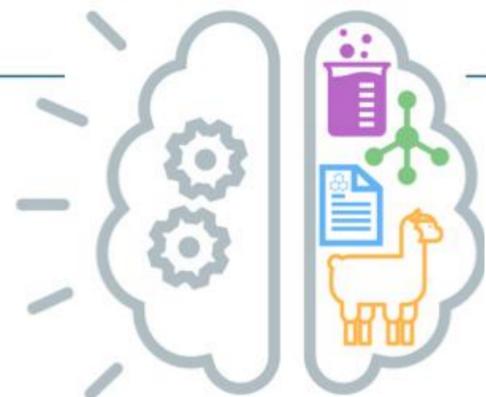
Peer-reviewed MatSci Publications
Materials Science Community Discourse



Crystallographic Information Files

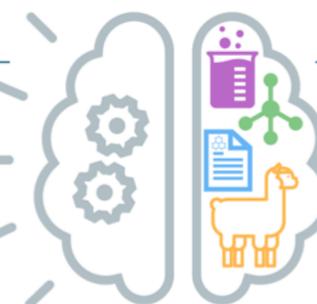
RedPajama (subset)

Publications
Redpajama
MatSci Comm.
CIF



LLaMat

Instruction Finetuning



Finetuning for Crystals

OpenOrca



MatSciNLP



LLaMat



Syntactic Instructions

What is the lattice parameters of...?

What is the volume of the unit cell?

MathQA



MatSciInstruct



Semantic Instructions

Which atoms are present in the crystal?

What is the coordinate of...?

Which element is present with ...?

MatBookQA



MaSQA



LLaMat-Chat

Materials IE co-pilot



MatNLP
MOF

Doping
General



	0	1	2	3	4
0 Sample	Li2O	Na2O	Al2O3	SiO2	
1 C1	1	0	1	4	
2 C2	0.8	0.2	1	4	
3 C3	0.6	0.4	1	4	
4 C4	0.5	0.5	1	4	



IE from tables



IE from text

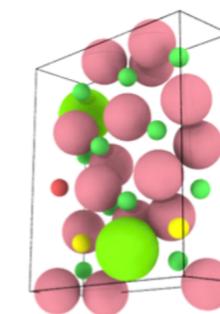


LLaMat-CIF

Crystal Generator



Generate a stable crystal based on...





LLaMat

First Large Language Model for Materials

Dataset	Size	Items
Mat Sci Comm Discourse	14 M tokens	37401
Elsevier	24B tokens	3.42 M (521 journals)
Springer	6B tokens	0.641M (395 journals)
Red Pajama	1B tokens	
CIF Files	~4gb	454,046

MatSci-Instruct NLP Tasks

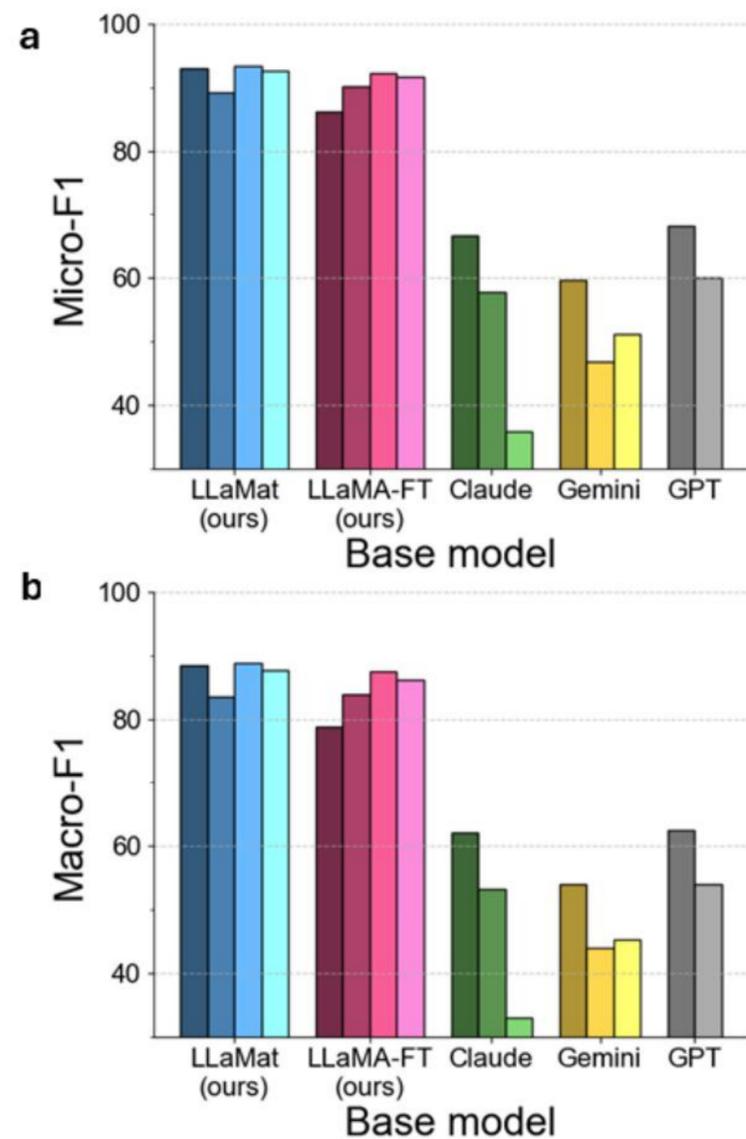
Machine Reading Comprehension	224
Question Answering	224
Open-Ended Generation	214
Classification	203
Information Extraction	201
Relation Extraction	193
Analysis	189
Topic Modeling	188
Writing	180
Commonsense Reasoning	172
Code Interpretation	172
Event Extraction	167
Grammar Correction	165
Clustertin	162
Named Entity Recognition	160
Text Simplification	153
Summarization	149
Sentiment Analysis	149
Part-of-Speech Tagging	146
Editing	145



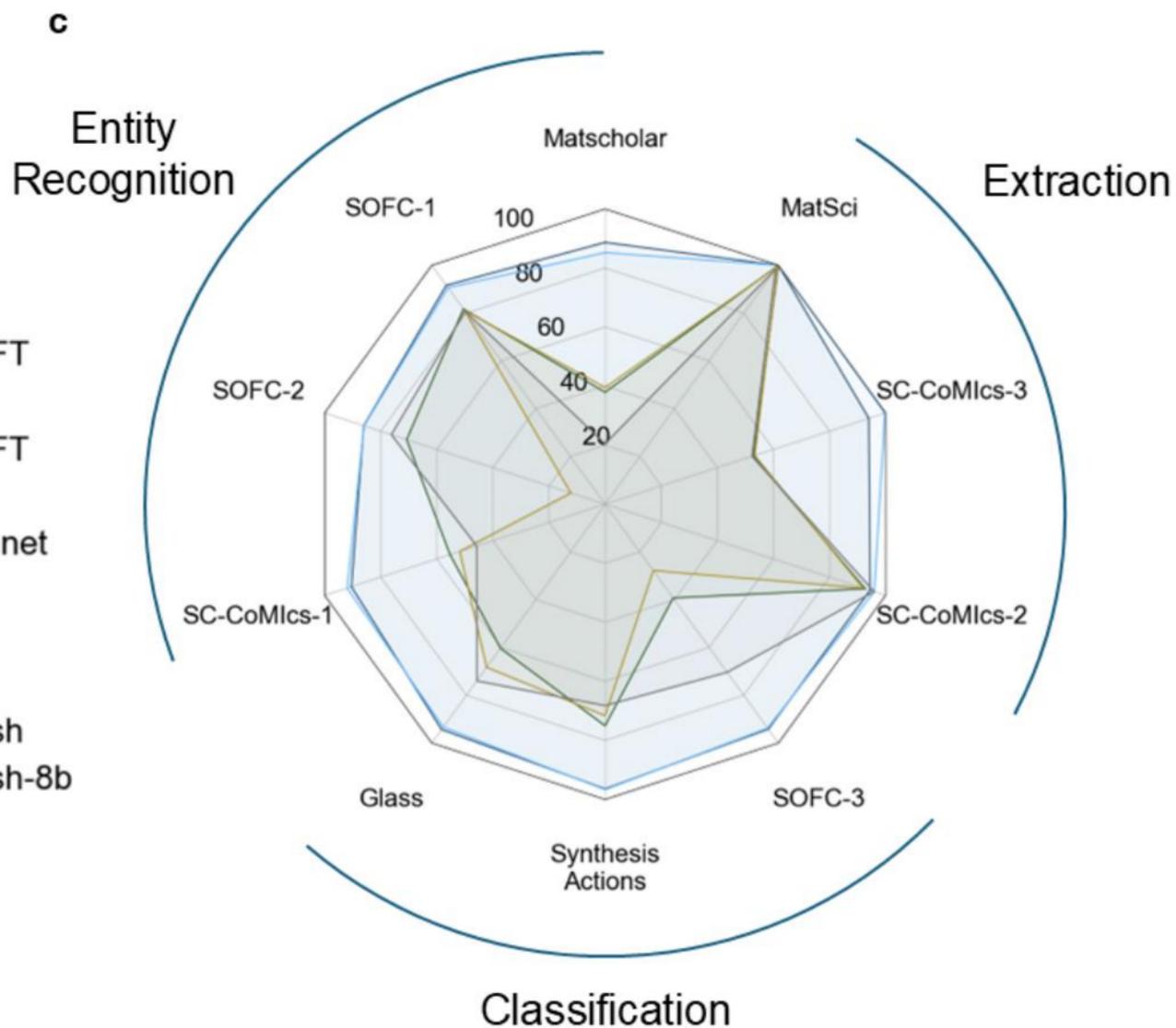
LLaMat

First Large Language Model for Materials

LLaMat significantly outperforms LLMs on MatSci tasks!



- LLaMat-3-chat
- LLaMat-3
- LLaMat-2-chat
- LLaMat-2
- LLaMA-3-chat-FT
- LLaMA-3-FT
- LLaMA-2-chat-FT
- LLaMA-2-FT
- Claude-3.5-Sonnet
- Claude-3-Opus
- Claude-3-Haiku
- Gemini-1.5-Pro
- Gemini-1.5-Flash
- Gemini-1.5-Flash-8b
- GPT-4o
- GPT-4



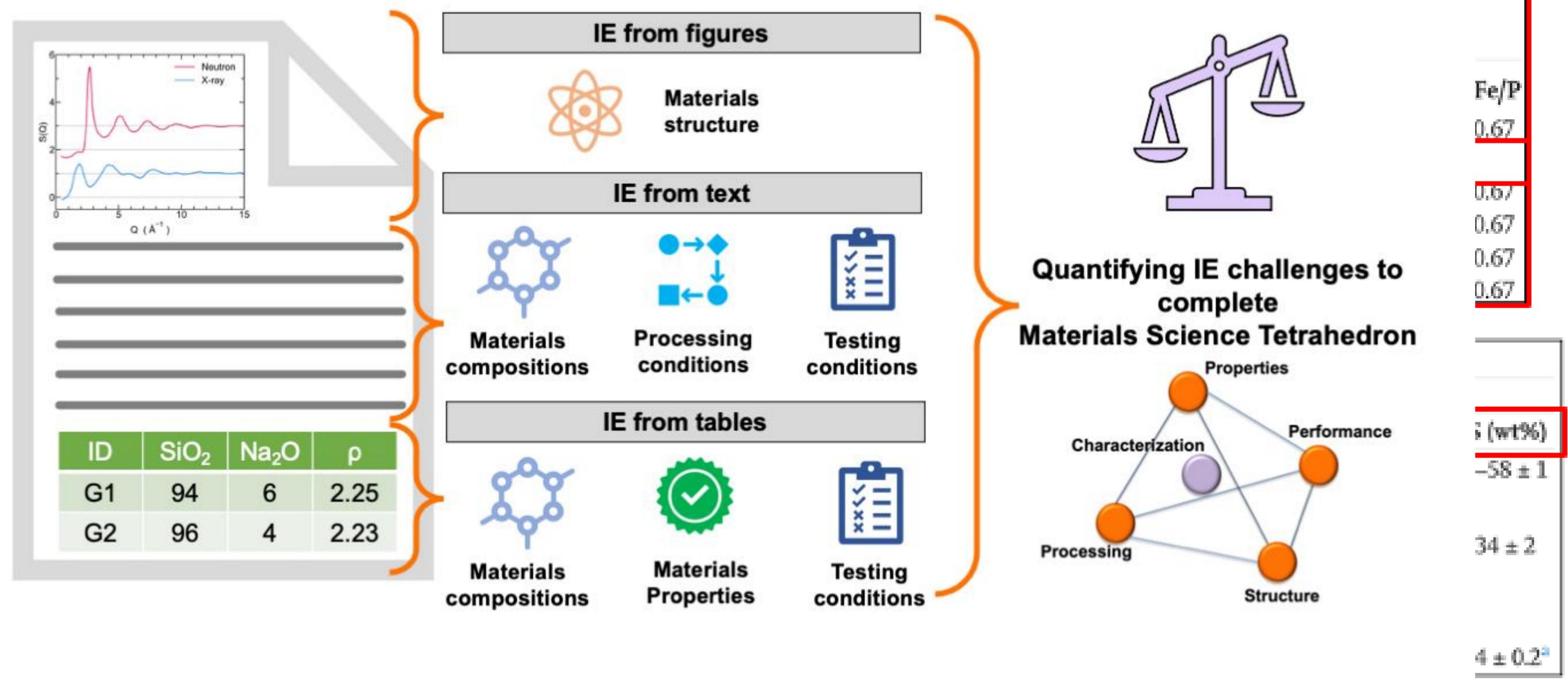


IE from Tables



Information Extraction from Tables

Challenges





Information Extraction from Tables

Challenges

6. **Single-cell compositions:** Entire composition in one table cell (some coefficients may be missing)
7. **Percentages not summing to 100:** Re-normalization is needed. Common case is doping
8. **Percentage contributions as variables:** Like x, y, etc. Contributions can have arithmetic expressions
9. **Partial Info Tables:** Percentages of only some constituents reported in table. Infer remaining composition from table caption or publication text
10. **Other corner cases**

Composition	$\log \sigma_{298} (\text{S cm}^{-1})$
$(80\text{GeS}_2-20\text{Ga}_2\text{S}_3)_{90}-(\text{LiI})_{10}$	-6.34 (5)
$(80\text{GeS}_2-20\text{Ga}_2\text{S}_3)_{90}-(\text{NaCl})_{10}$	-6.92 (5)

Base composition(mol%): $75\text{TeO}_2-20\text{ZnO}-5\text{Na}_2\text{O}$

Glass	Tm_2O_3 (wt%)	Yb_2O_3 (wt%)	Ho_2O_3 (wt%)
Undoped	-	-	-
TZN	-	-	-
TT05	0.5	-	-
TTY20	1	2	-
TTH05	1	-	0.5
TTYH	1	5	0.15

	$(\text{GeSe}_2)_{1-x}(\text{GeBr}_4)_x$		$(\text{GeSe}_2)_{1-x}\text{Br}_x$		
x	0.079	0.167	0.265	0.250	0.429
$p_{\text{Se}}(\%)$	94.1	86.0	76.5	84.1	78.0

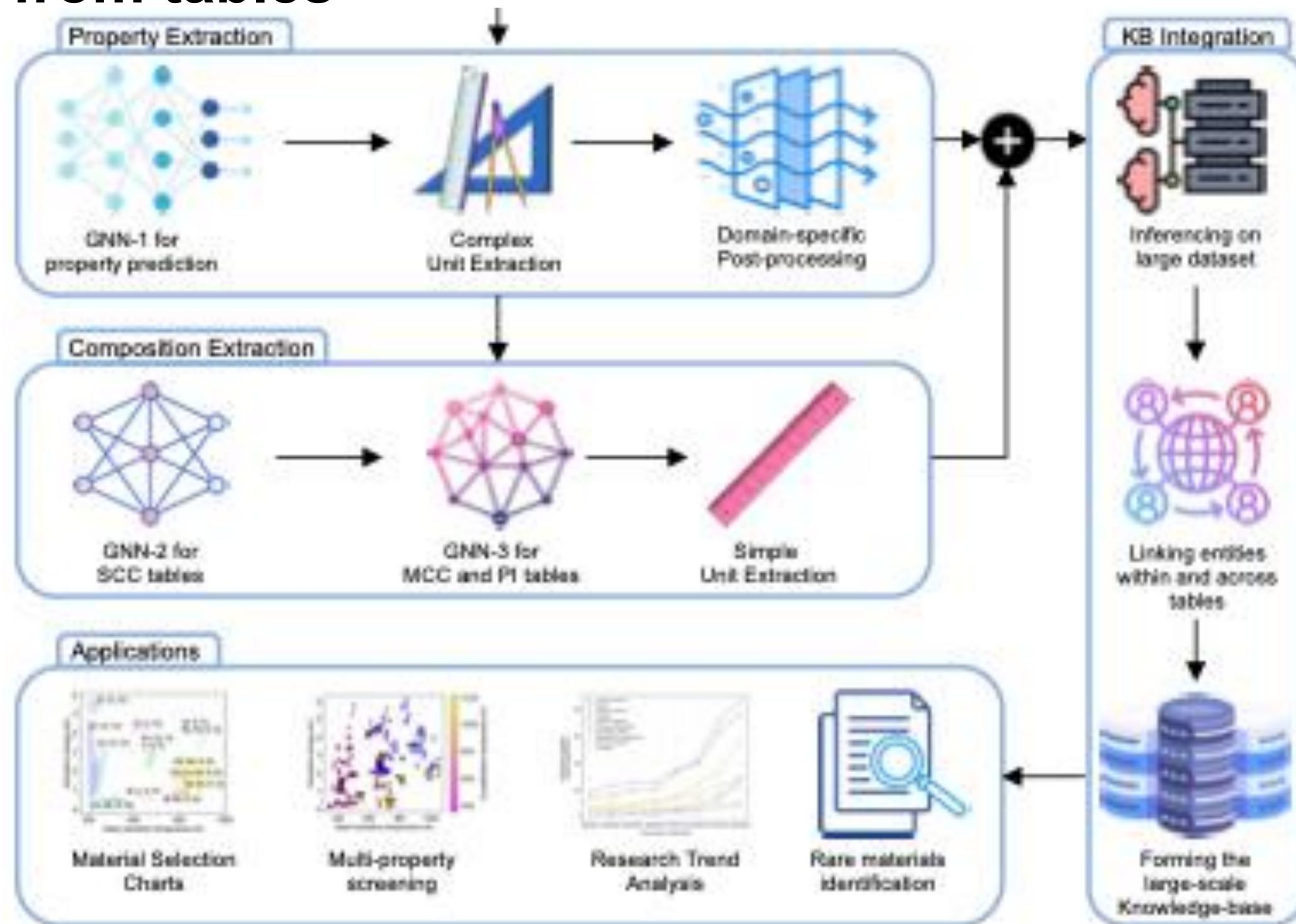
Glass composition in the paper:
 $20\text{La}_2\text{S}_3 - (80 - x)\text{Ga}_2\text{S}_3 - x\text{CsCl}$

Glass	CsCl (mol%)	n (1.5 μm)
GLSC10	10	2.253 ± 0.001
GLSC20	20	2.265 ± 0.001

MatSKRAFT

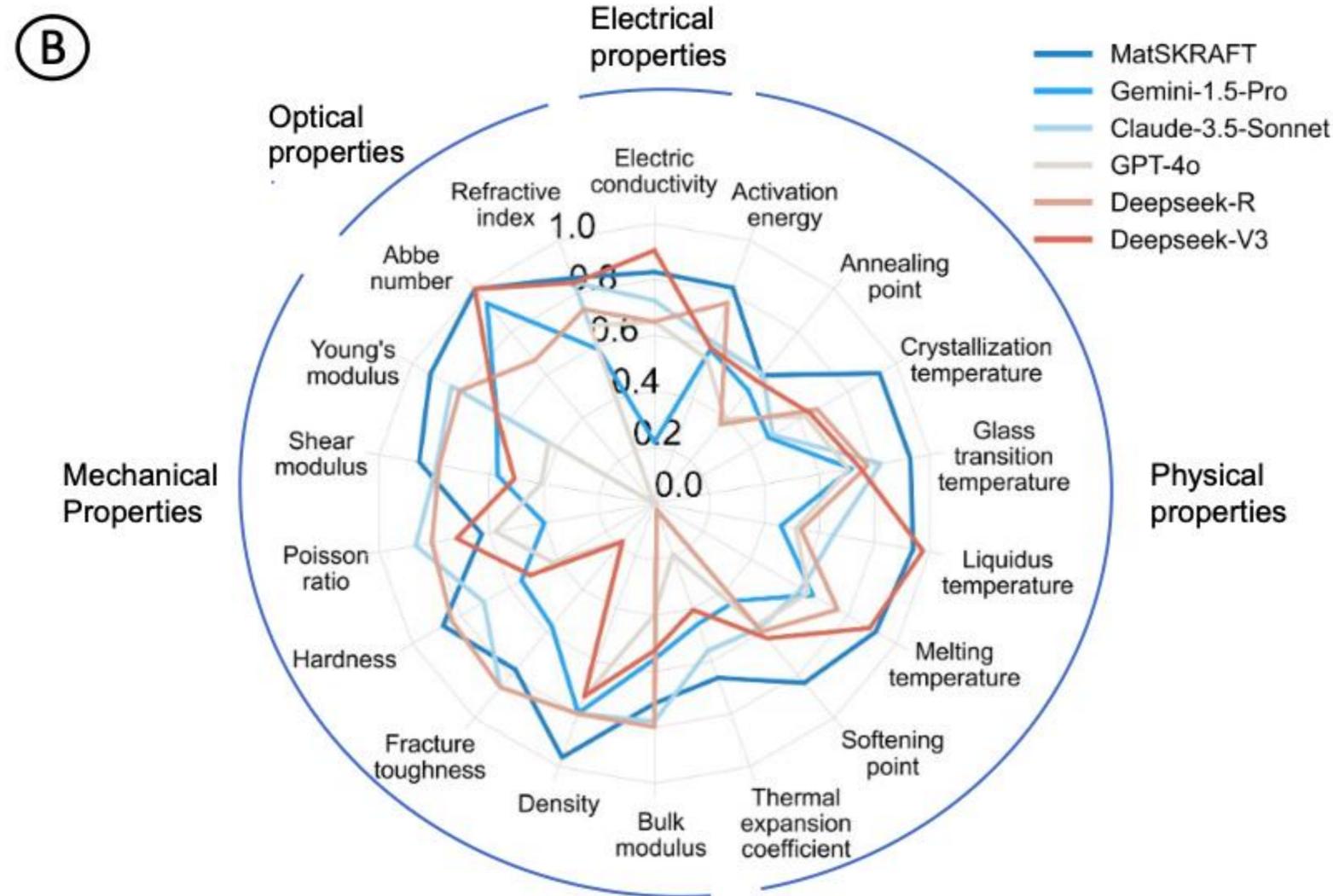
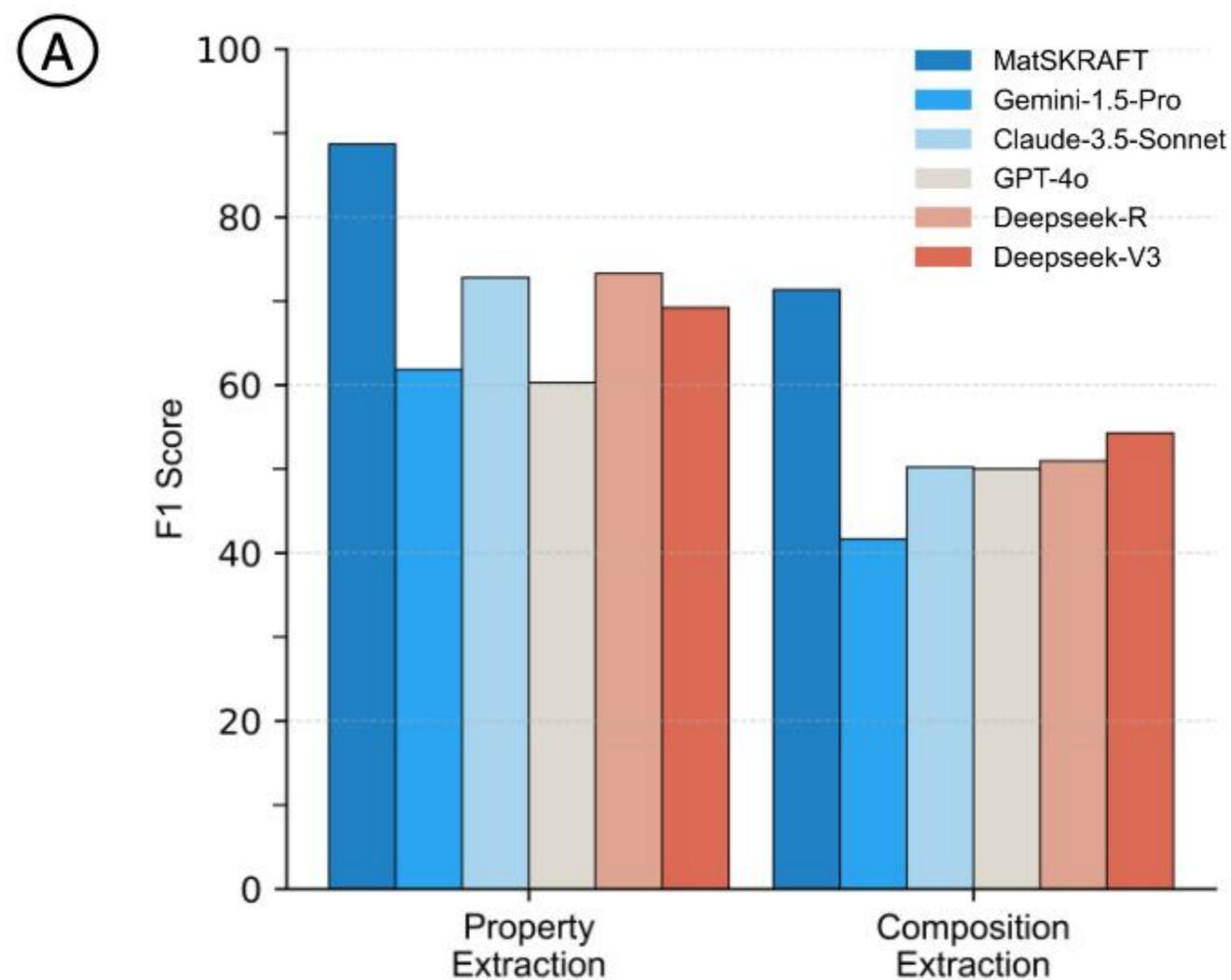
Composition extraction from tables

1. Property extraction with units
2. Composition extraction
3. Linking them
4. Applications



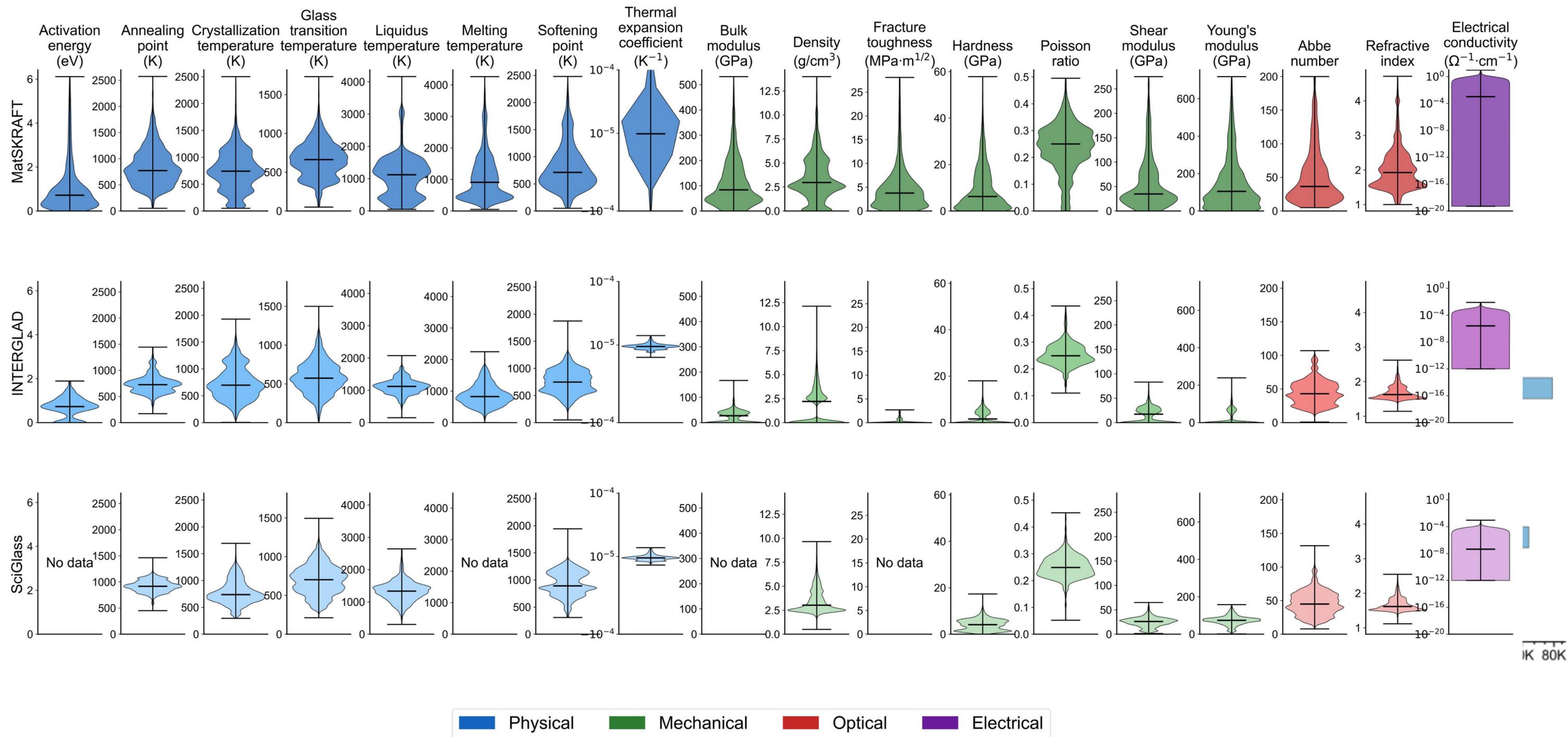
MatSKRAFT

Composition and property extraction from tables

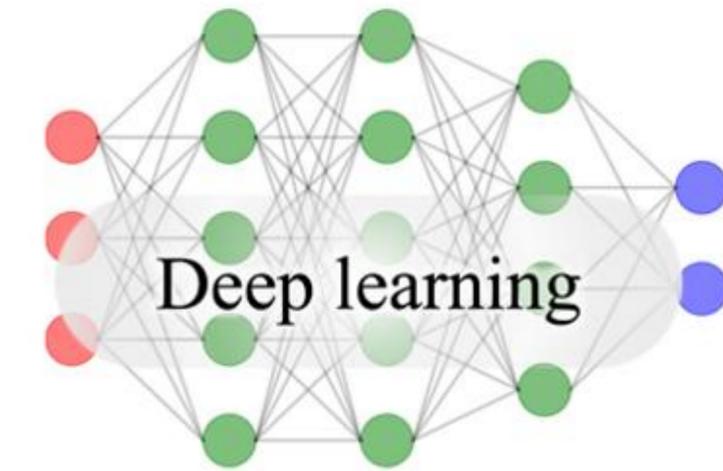
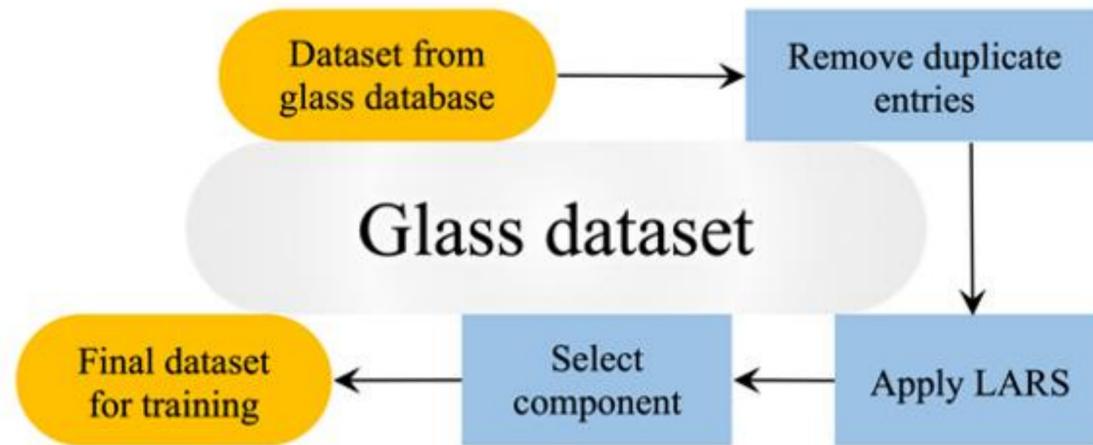


MatSKRAFT

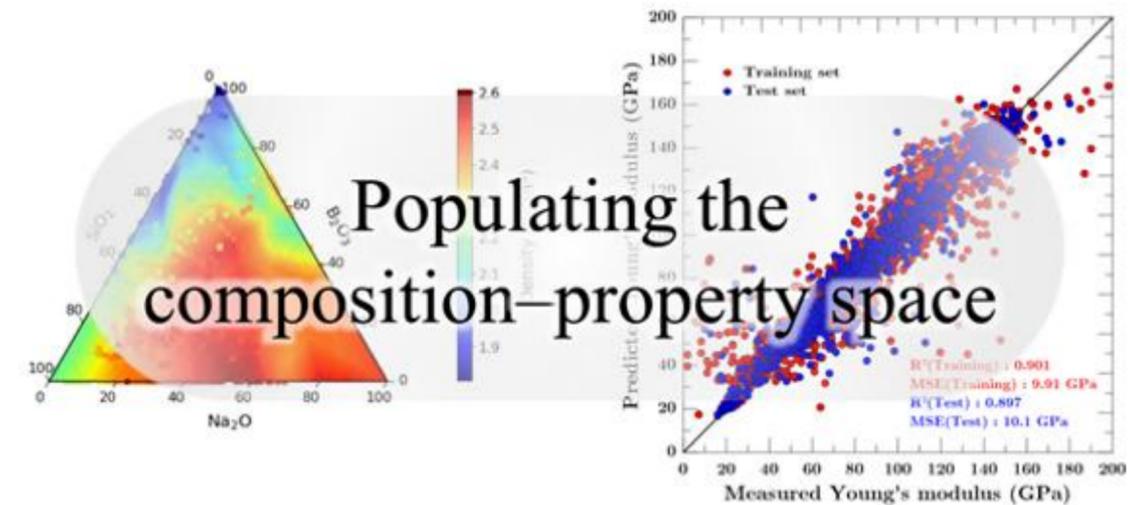
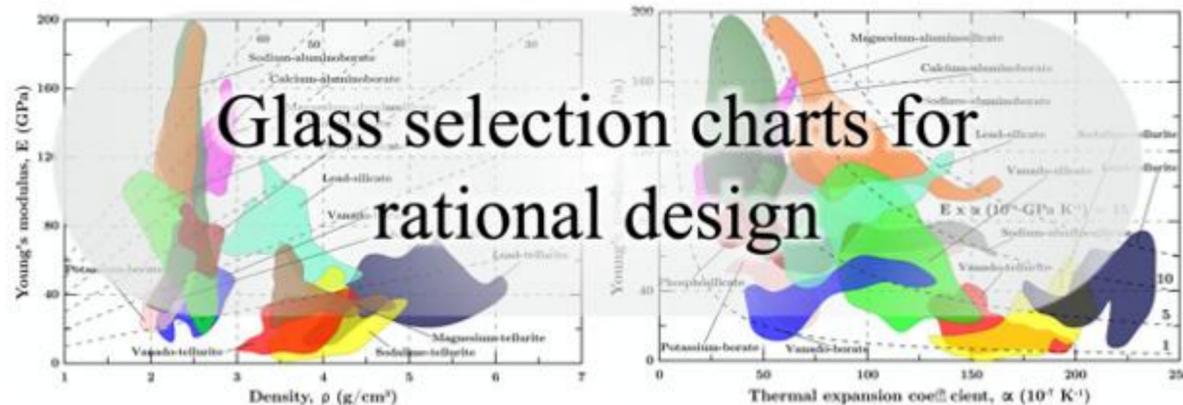
Extracted dataset



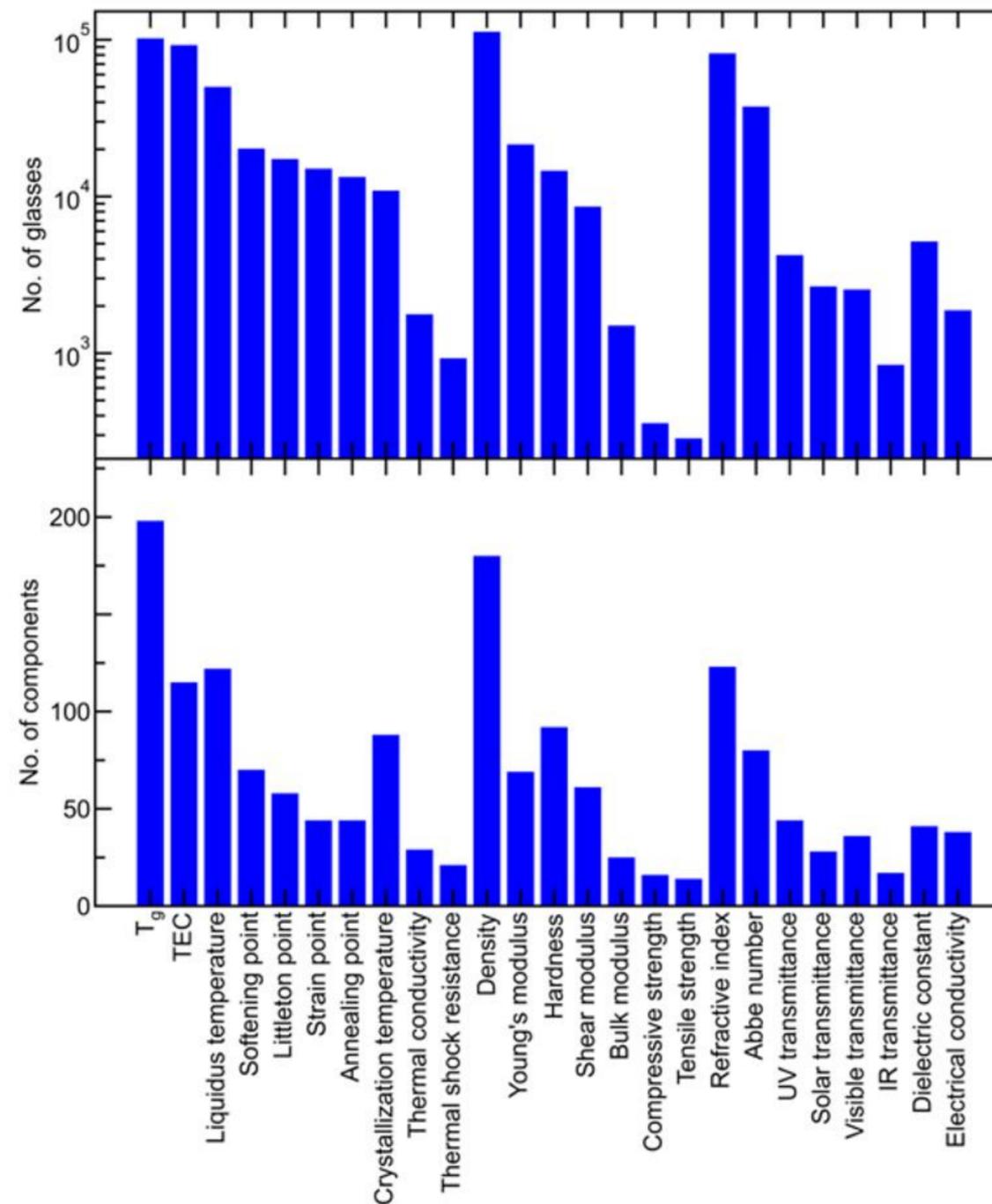
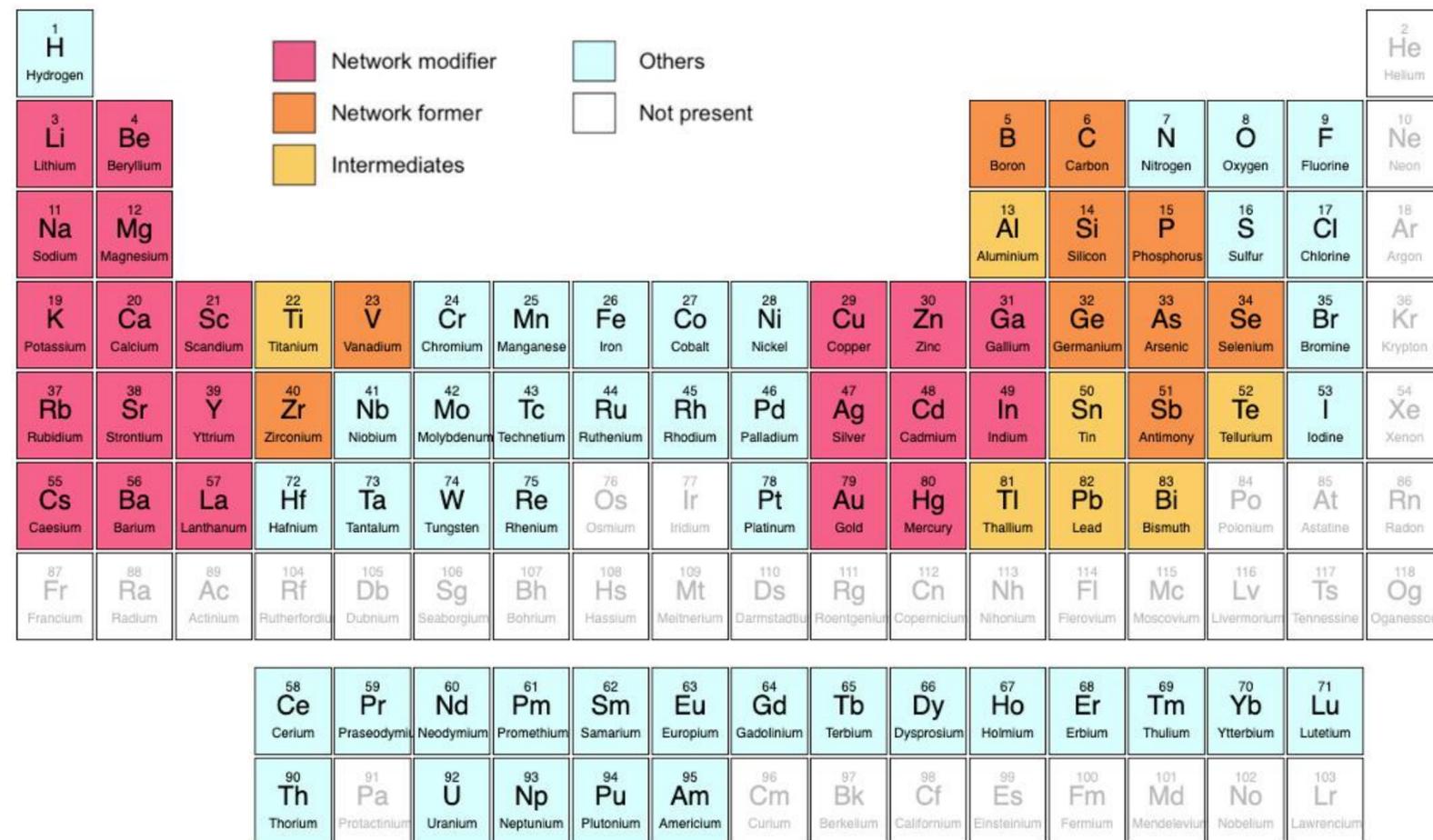
Data-driven modeling in glasses



Input Hidden layers Output

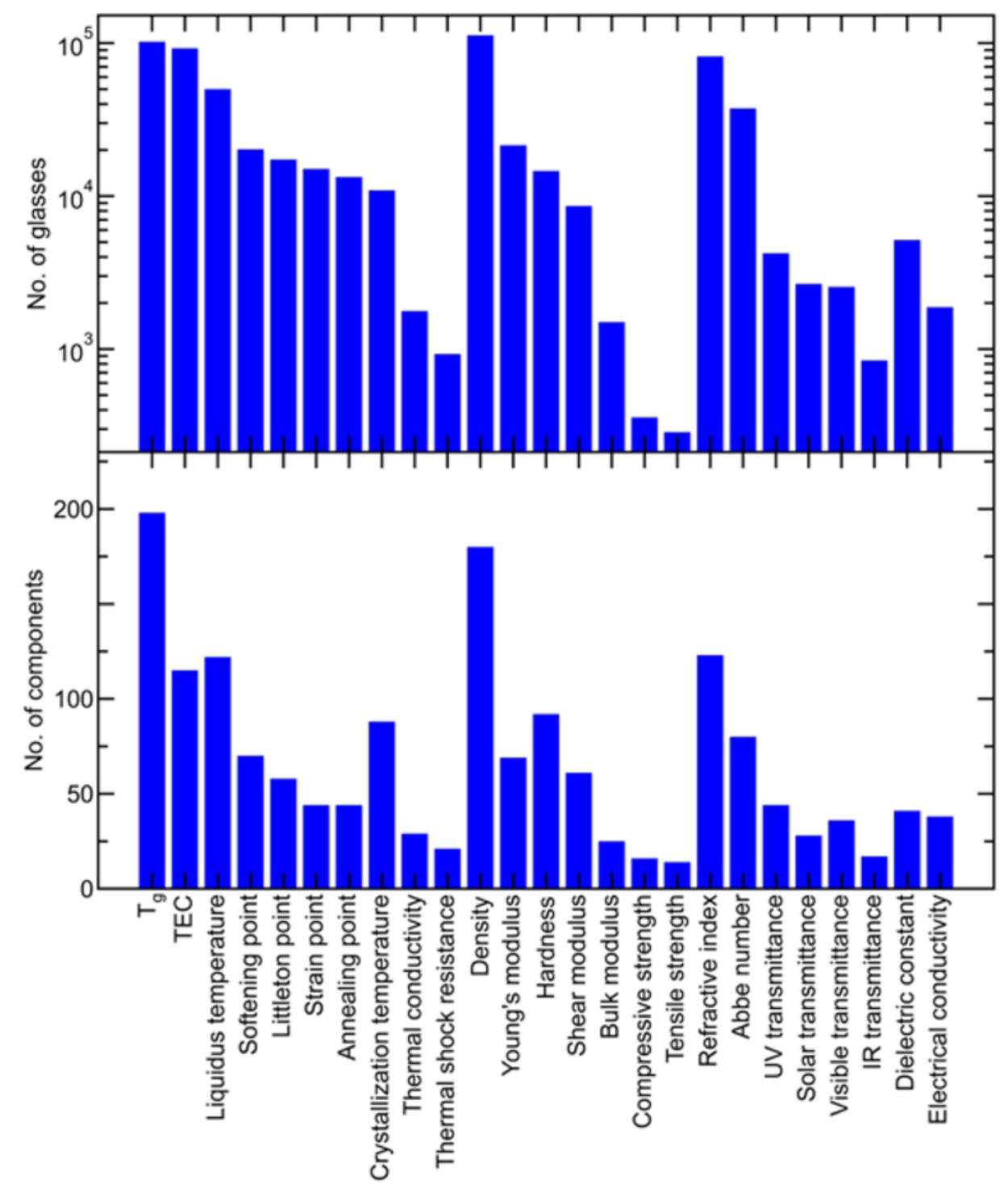
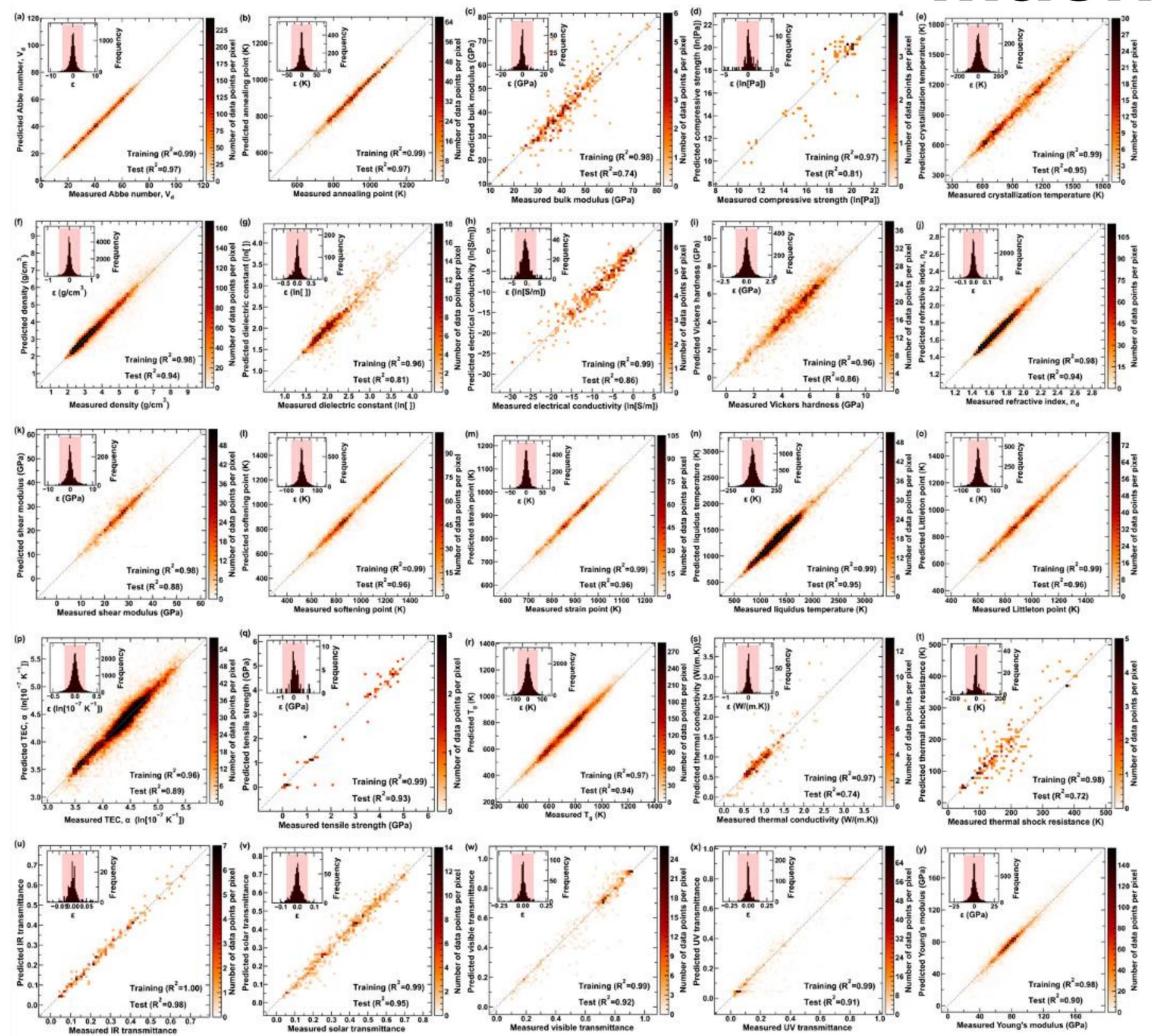


Dataset visualization

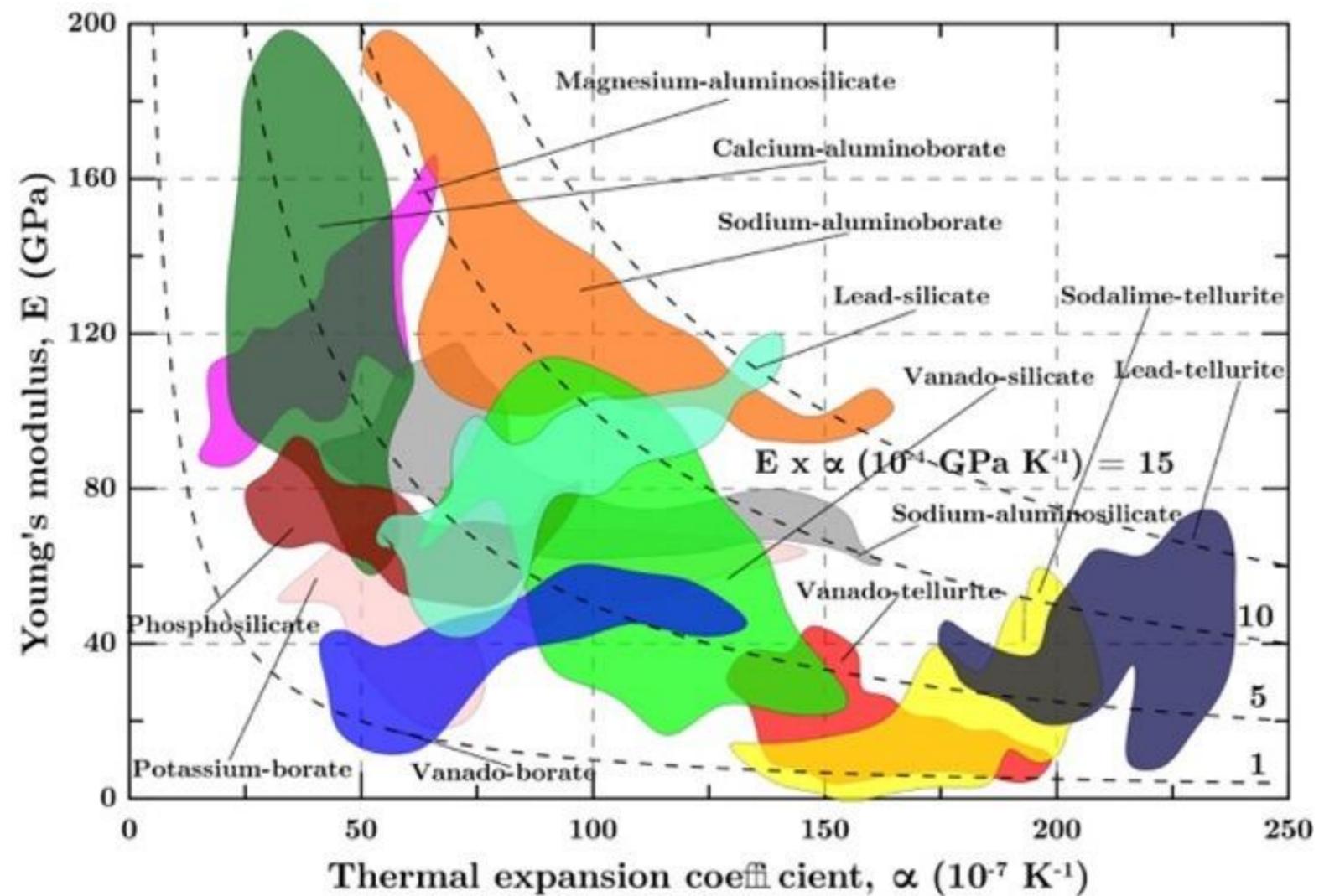
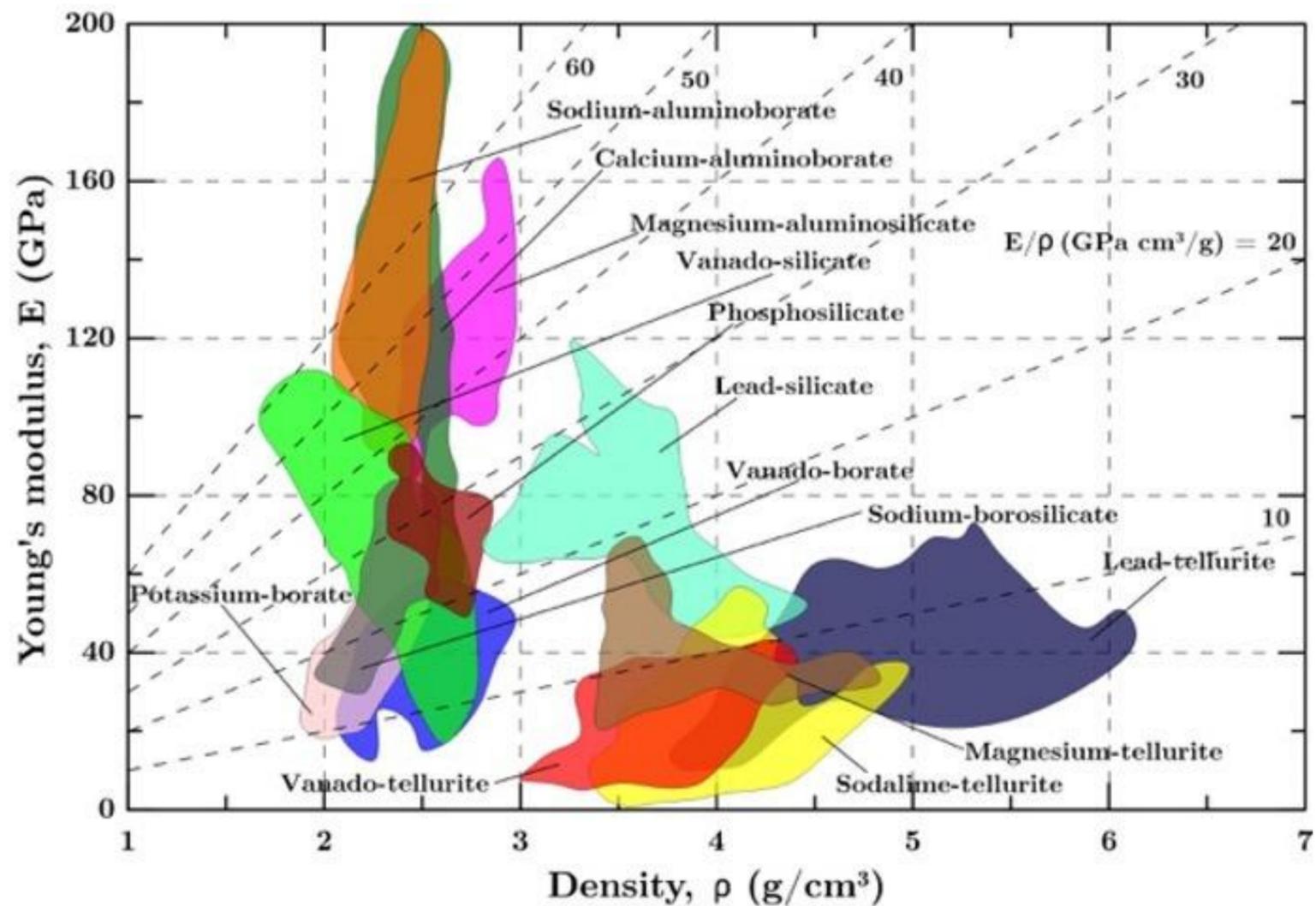


25 properties; 84 elements; 224 components; 350,000 glass compositions

Machine learning models

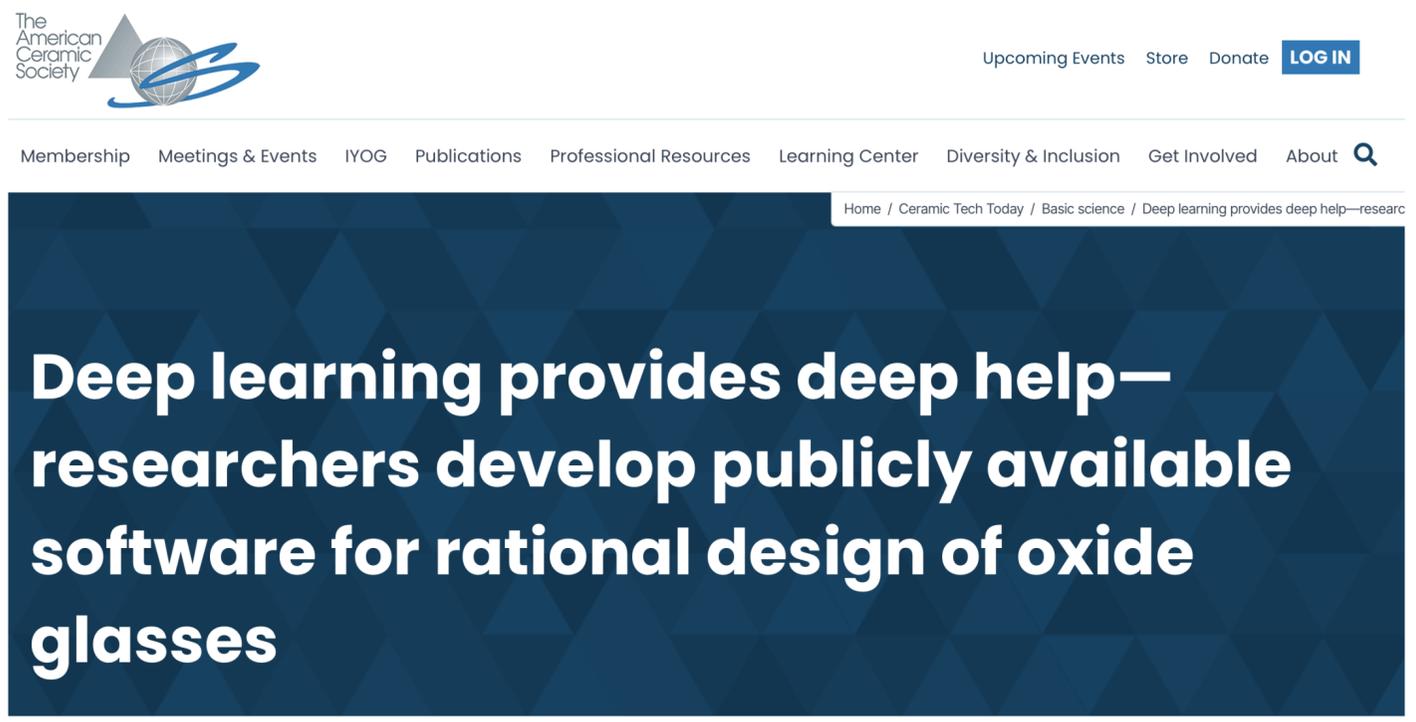


Material selection chart



Python for Glass Genomics

<http://pyggi.substantial.ai>



- **ISRO: Glasses for space application**
- **BRNS: Glasses for nuclear waste immobilization**
- **Sisecam: Glasses for commercial applications**
- **Battery electrolyte glasses**





Python for Glass Genomics

Python for Glass Genomics (PyGGi) is a package for accelerating innovations in the field of glasses and ceramics. It uses state-of-the-art machine learning algorithms and computational techniques to enable the users in designing novel glasses and understanding the composition-property relationships.

[Get Started](#)
[About PyGGi](#)

Overview

 <p>PyGGi Bank Explore composition-property database of glasses collected from literature</p>	 <p>PyGGi Seer Predict glass properties as a function of the glass composition</p>	 <p>PyGGi Zen Discover new glass compositions based on target properties and compositional constraints</p>	 <p>PyGGi Pictionary Explore published works and images from glass literature</p>
<p>4 Packages</p>	<p>9 Properties</p>	<p>34 Oxide Components</p>	<p>50000+ Glasses</p>

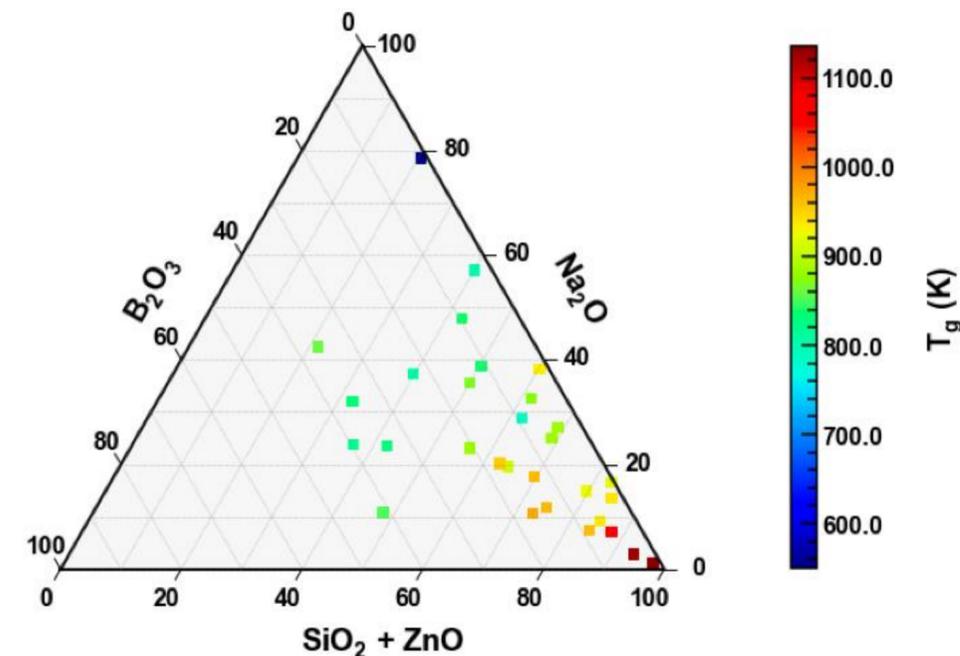
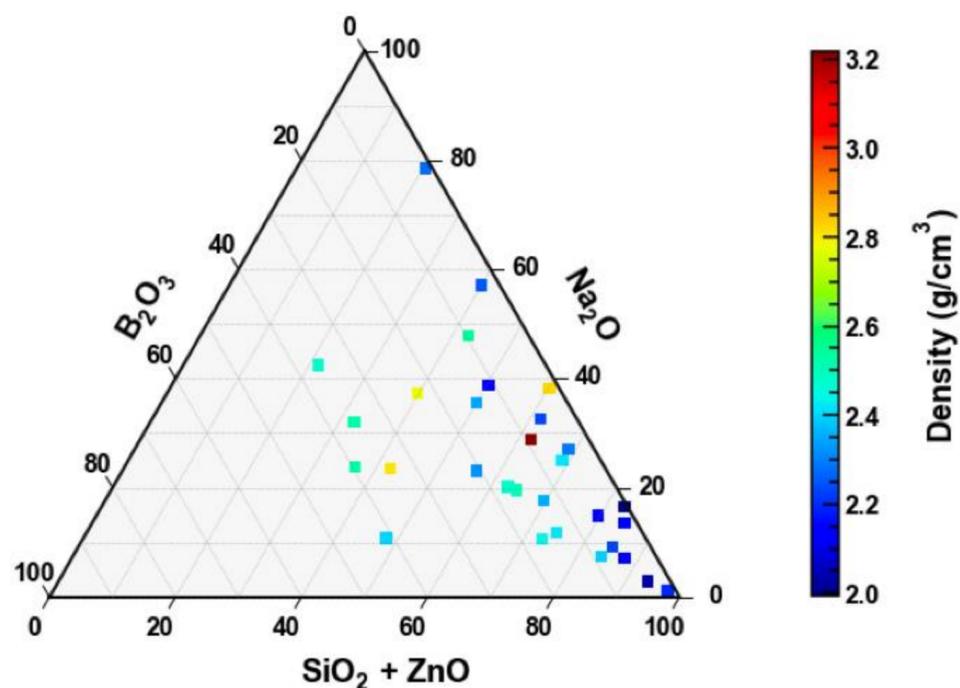
Constraints for nuclear waste glass

Composition constraints:

- $\text{SiO}_2 + \text{B}_2\text{O}_3 \sim 60-70\%$ (wt) is preferred with $\text{SiO}_2 > 45\%$ (wt) with $\text{B}_2\text{O}_3 > 10$ wt%
- $\text{Na}_2\text{O} > 15\%$ wt
- $\text{Na}_2\text{O}/\text{B}_2\text{O}_3 > 1$, preferably 1.25
- $\text{ZnO} \sim 2.5 - 10$ wt% with a never exceed to 15% wt%

Property constraints:

- T_g value $\Rightarrow 550$ °C
- Density = 2-3 g/cm³
- Electrical conductivity = 50 S/cm
- Viscosity > 50 poise at pouring temperature
- Dissolution rate $< 10^{-5}$ to 10^{-6} g/cm²/day)



Optimized glass compositions

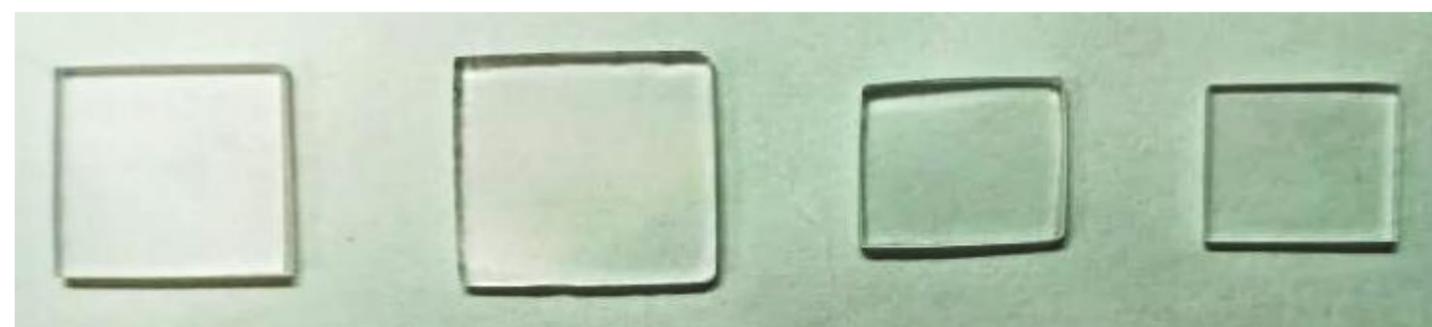
❖ Series of glasses obtained after compositional optimization using machine learned model

Na₂O (mol%)	SiO₂ (mol%)	B₂O₃ (mol%)	ZnO (mol%)	TEC (ln[10⁻⁷ K⁻¹])	Thermal Conductivity (W/(m.K))	Density (g/cm³)	T_g (K)	Viscosity (poise)
21	52	18	9	101.90	0.77	2.62	808.87	2.6831
24	52	18	6	105.57	0.78	2.59	785.74	3.0581
27	52	18	3	114.34	0.77	2.51	776.94	3.5894
30	52	18	0	121.98	0.75	2.52	779.55	1.4328

Glass preparation

Sample Name	Na ₂ O	SiO ₂	B ₂ O ₃	ZnO
NBSZ-9	21	52	18	9
NBSZ-6	24	52	18	6
NBSZ-3	27	52	18	3
NBSZ-0	30	52	18	0

Temperature (°C)	Time (h)
0-300	3
300	4
300-800	5
800	4
800-1200	2
1200	1



NBSZ-9

NBSZ-6

NBSZ-3

NBSZ-0

Comparison between ML and experiments

Na ₂ O (mol%)	SiO ₂ (mol%)	B ₂ O ₃ (mol%)	ZnO (mol%)	TEC (ln[10 ⁻⁷ K ⁻¹])	Thermal Conductivity (W/(m.K))	Density predicted (g/cm ³)	Density measured (g/cm ³)	T _g (K)	T _g measured (K)
21	52	18	9	101.90	0.77	2.62	2.61	809	803
24	52	18	6	105.57	0.78	2.59	2.60	786	806
27	52	18	3	114.34	0.77	2.51	2.56	777	793
30	52	18	0	121.98	0.75	2.52	2.52	780	785

How did we prepare these glasses?

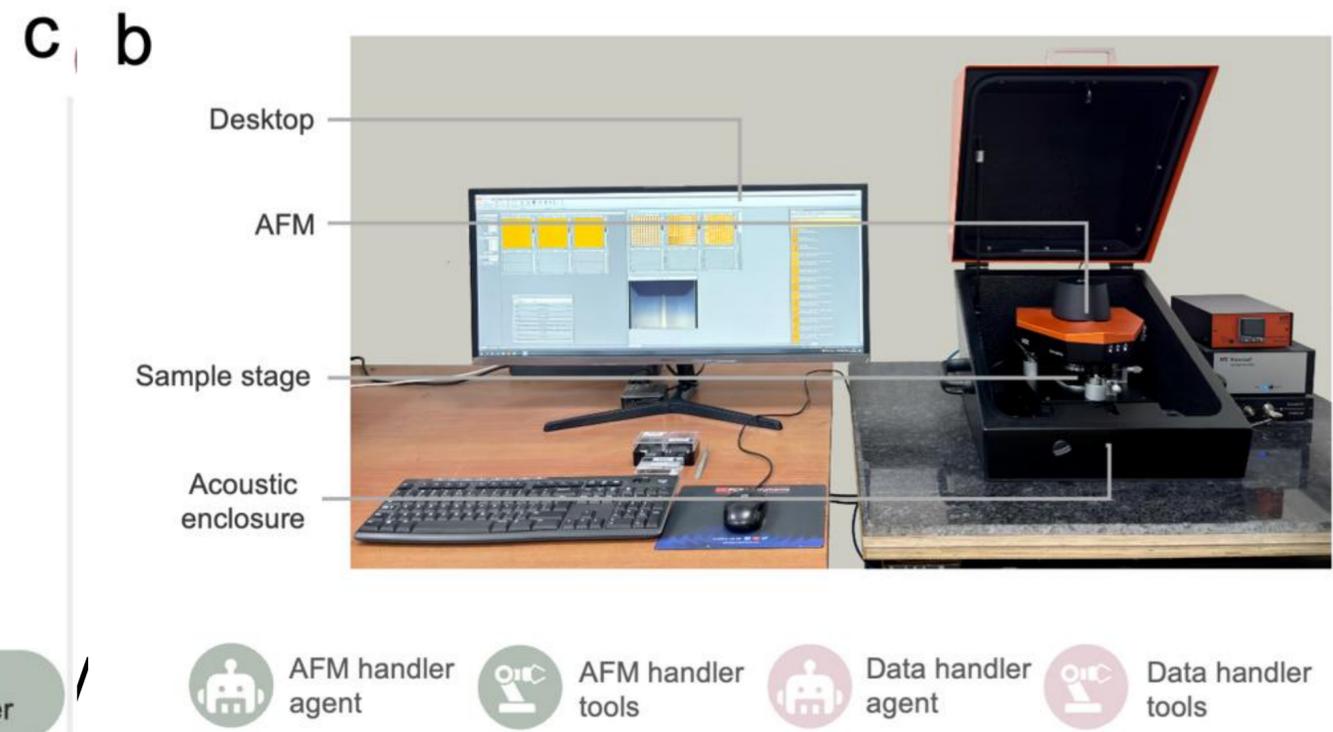
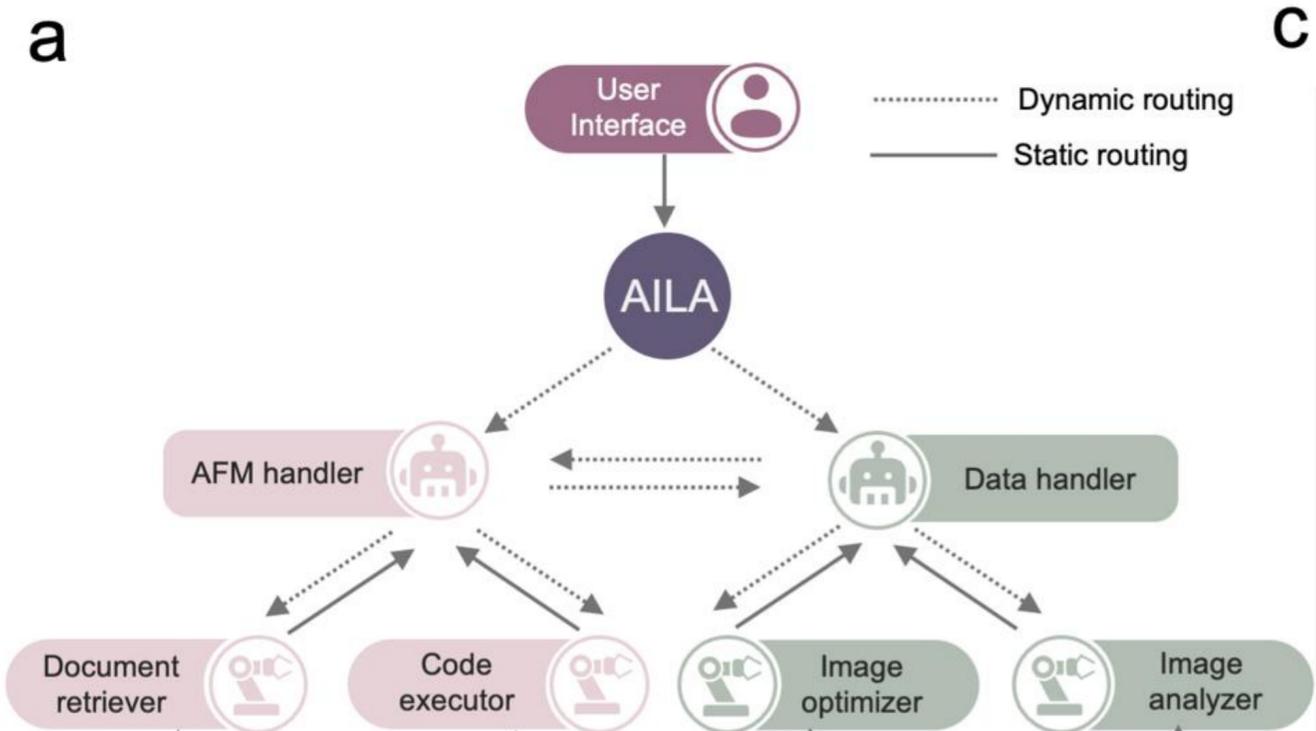


Towards Automated Experiments



Automated Atomic Force Microscopy

Can Agentic LLMs drive experiments?



Single Instruction Question

Set the image capture parameters in the AFM software as follows: Image size: 100x100 nm, Pgain (Proportional gain): 100.

Multi-Instruction Question

Set file name to 'image-1' and then capture a 600x600 nm image with Pgain=100, Igain=10.

Mathematical Reasoning Question

Capture an AFM image with a size of 100x100 μm, configuring the time/line and total lines in settings to ensure the scan completes within 60 seconds.

c

Capture an AFM image of size 100x100 nm with the filename '60_GPT-4o'. Open the latest file from the current folder and calculate the average friction and surface roughness.

AFM image capturing should be performed by the AFM_Handler, and the image after-processing should be handled by the Data_Handler. Given the tasks, the AFM_Handler should act first to capture the image.

Tool Calls: Document_Retriever query: code to set scan area to 100 nm x 100 nm

```
#imports
import nanosurf

#load application
spm = nanosurf.SPM() # or .C3000() or .CX() or .CoreAFM()
...
```

Tool Calls: Code_Executor query: code: import nanosurf

```
# Load application
spm = nanosurf.SPM() # or .C3000() or .CX() or .CoreAFM()
application = spm.application
scan = application.Scan

# Set scan parameters for 100 nm x 100 nm
scan.ImageWidth = 100e-9 # [m]
scan.ImageHeight = 100e-9 # [m]
...
```

code executed successfully

NEED HELP

Let's proceed with opening the latest file and calculating the average friction and surface roughness using the tools available. I'll perform these actions now.

Tool Calls: Image_Analyzer Args: calculate_friction

Raw Image C:\...\60_GPT-4o1877.nid processed successfully.

FINAL ANSWER: The AFM image named '60_GPT-4o' has been successfully captured and analyzed. The results are as follows:

- **Average Friction:** (7.93 times 10⁻⁵)
- **Mean Roughness:** (1.11 times 10⁻⁸)
- **RMS Roughness:** (1.32 times 10⁻⁸)



Automated Atomic Force Microscopy

Can LLMs drive experiments?

The screenshot displays a JupyterLab environment used for automating Atomic Force Microscopy (AFM) experiments. The interface is divided into several key sections:

- Parameter Control (Left):** A sidebar containing various control panels such as "Parameters", "Z-Controller", and "Mode Properties", used to configure the AFM system.
- Image Acquisition (Center):** A large window showing the AFM scan area with various axes (Z-Axis, Deflection, Friction force) and their respective line fits.
- Monitoring (Bottom Left):** A "Monitoring" window showing real-time data plots for "Normal Deflection", "Lateral Friction", "Mean Lock-In Amplitude", "2nd Lock-In Amplitude", "2nd Lock-In V", "Tip Voltage", and "Tip Current".
- Code Editor (Right):** A Python notebook containing a workflow graph and code. The code includes:

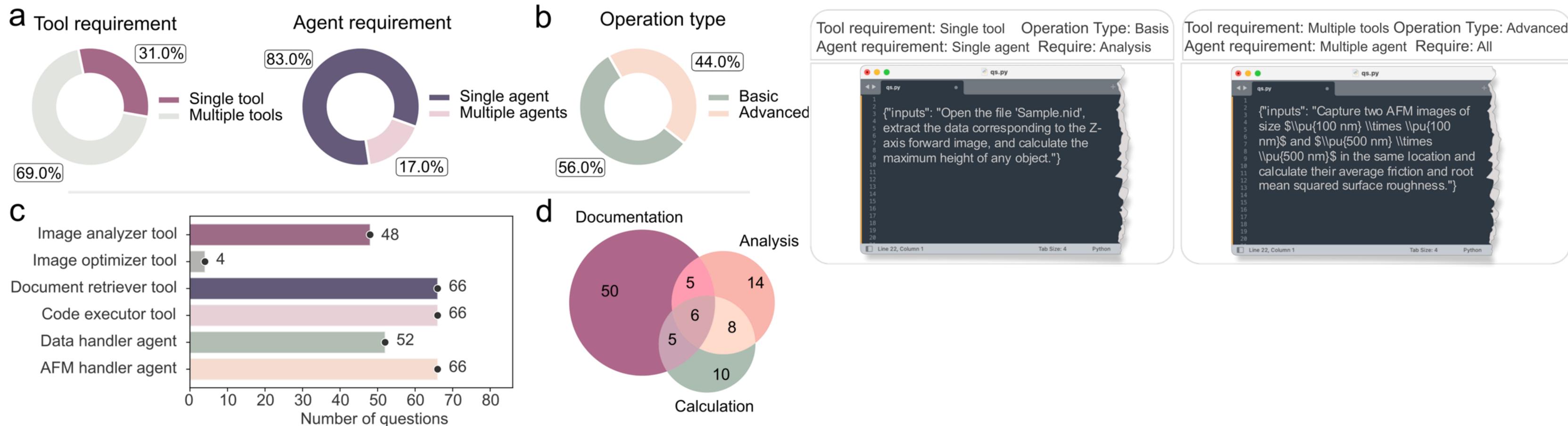
```
workflow.add_edge("AFM_Handler_Tools", "AFM_Handler")  
workflow.add_edge("Data_Handler_Tools", "Data_Handler")  
workflow.add_edge(START, "AILA")  
graph = workflow.compile()  
  
[17]: from IPython.display import Image, display  
display(Image(graph.get_graph(xray=True).draw_mermaid_png()))
```

The graph below the code shows a flow from `__start__` to `ChatPromptTemplate`, `AzureChatOpenAI`, `JsonOutputFunctionsParser`, and then to `AFM_Handler` and `Data_Handler`, which lead to their respective tool nodes and finally to `__end__`.
- Gallery (Far Right):** A "Gallery" window showing a list of scan results with columns for scan ID, size, time, and measure date.



Automated Atomic Force Microscopy

AFMBench

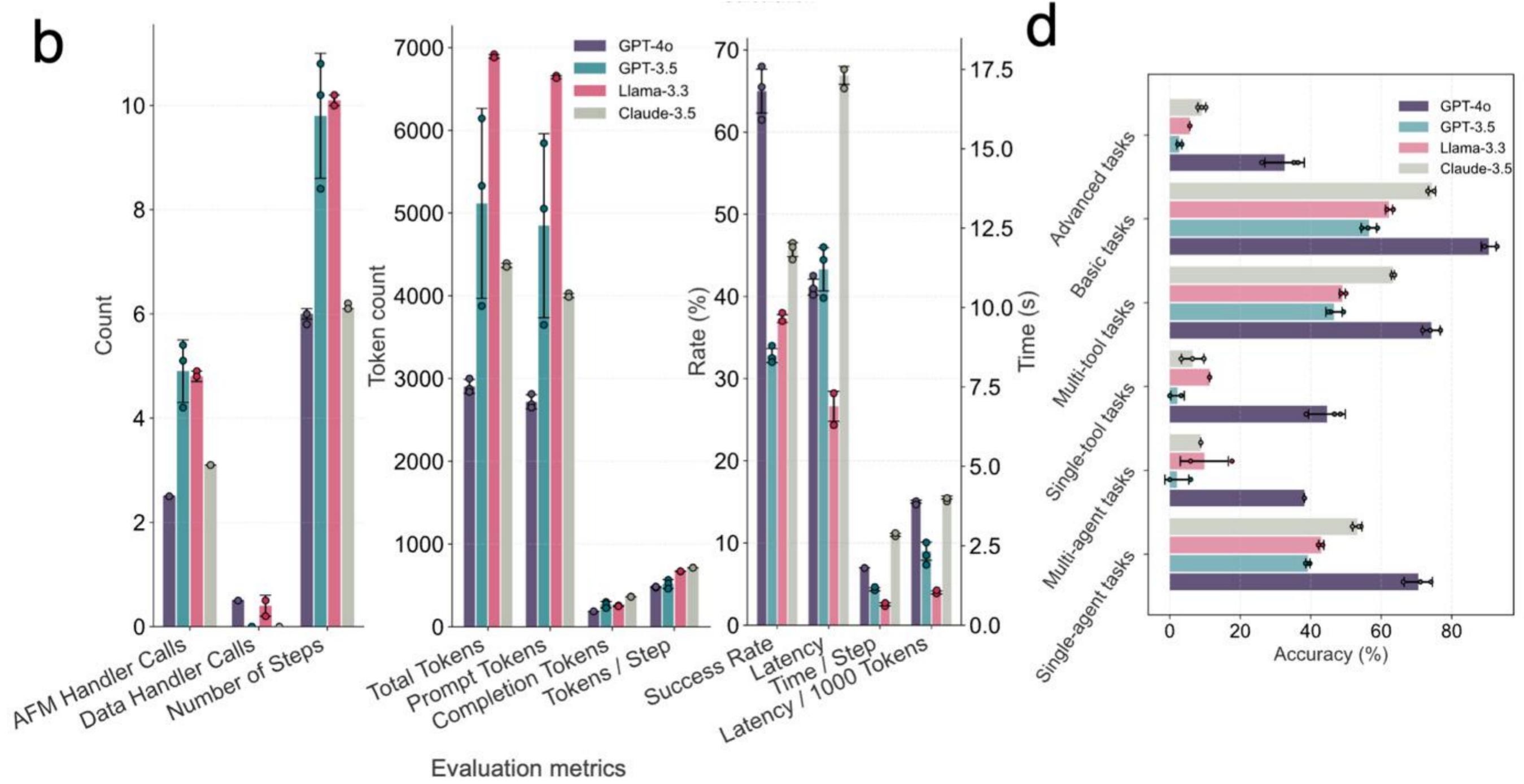


AFMBench: 100 different experiments; 5 challenging open-ended ones



Automated Atomic Force Microscopy

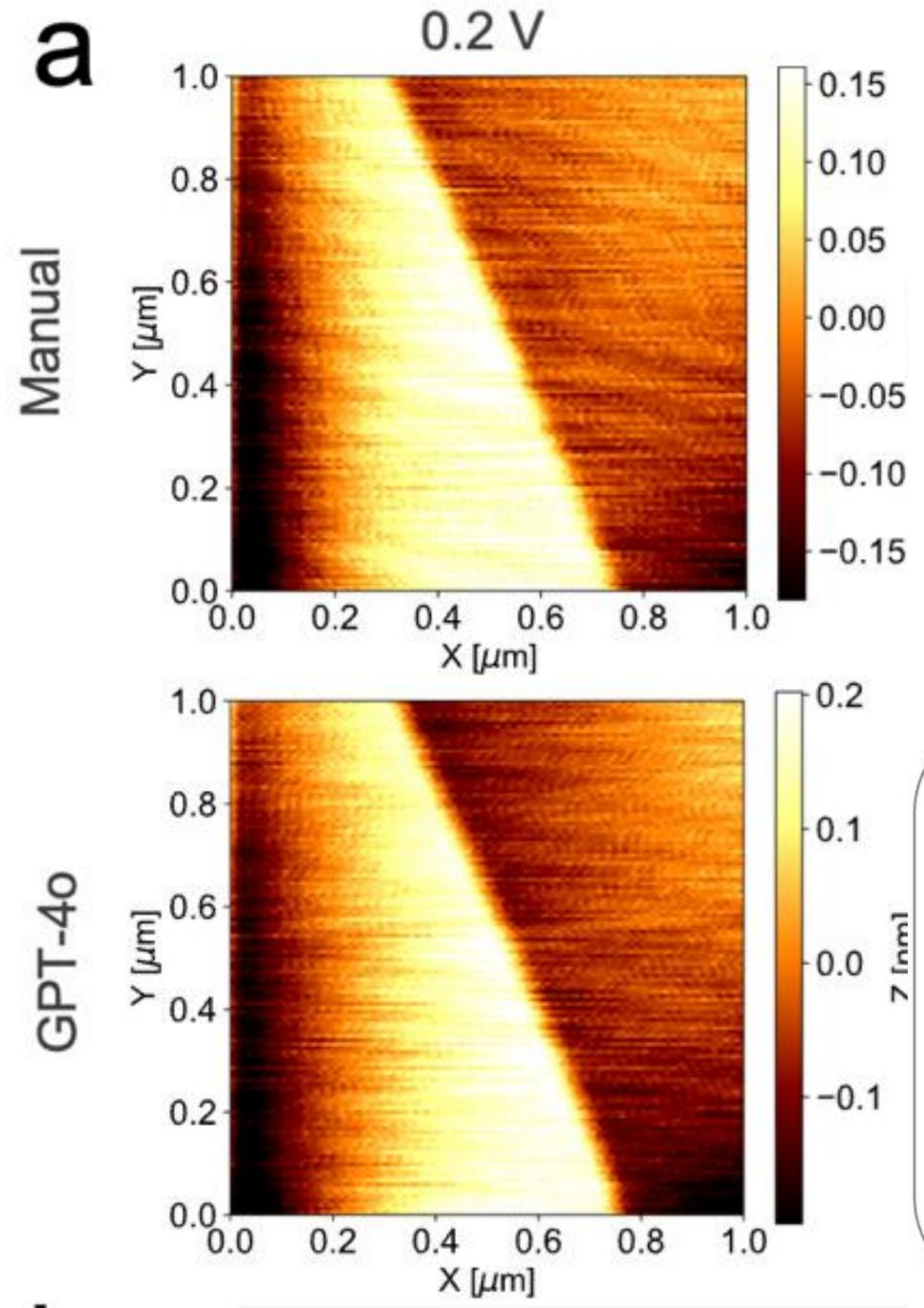
AILA Performance



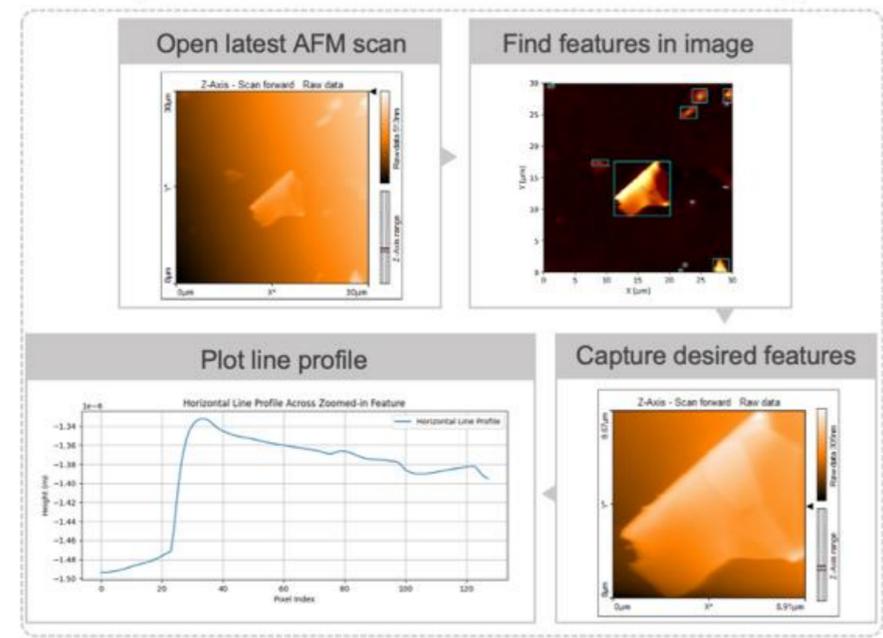


Automated Atomic Force Microscopy

AILA Performance



b Visualize the most recently saved AFM image and identify the largest feature present. Capture a zoomed-in image of this feature. Generate a horizontal line profile across the image to analyze the thickness of the feature and save the resulting line profile image as 'line_profile.png'. Based on the measured thickness, determine the number of graphene layers present in the feature.



The zoomed-in image of the largest feature has been successfully captured and analyzed. The horizontal line profile has been generated and saved as 'line_profile.png'.

Analysis of the Feature

- **Maximum Height Difference:** The line profile reveals a maximum height difference of approximately 1.61×10^{-7} meters (161 nm).

Determination of Graphene Layers

To estimate the number of graphene layers, consider that a single layer of graphene has a thickness of approximately 0.34 nm. The thickness of the detected feature is much larger than a single graphene layer, indicating the presence of multiple layers:

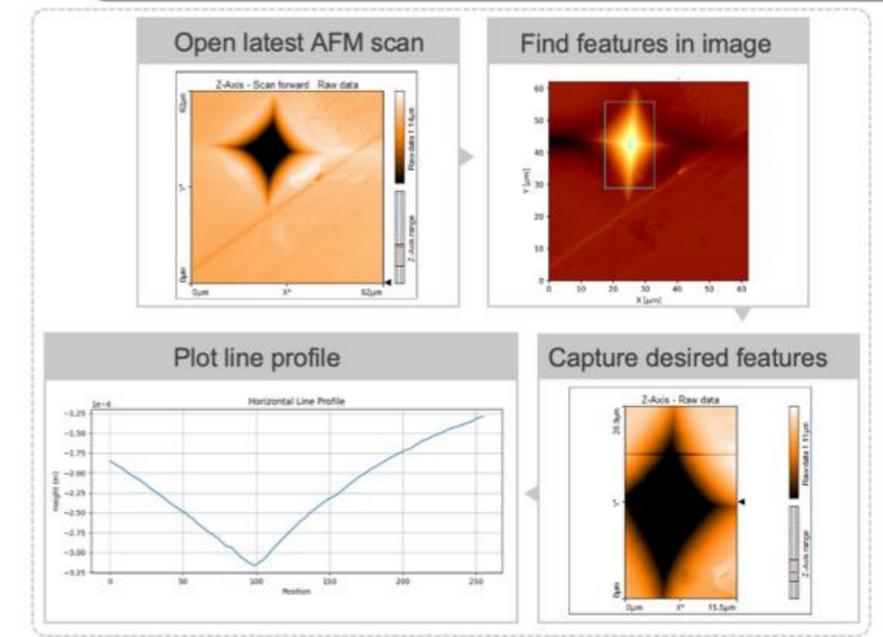
$$\text{Number of Layers} = \frac{161 \text{ nm}}{0.34 \text{ nm/layer}} \approx 473 \text{ layers}$$

Summary

The measured vertical thickness of the largest feature corresponds to approximately 473 layers of graphene.

AILA

c Visualize the most recently saved experimental image and identify the indentation mark. Take a zoomed-in image centered on the indentation area. Generate a horizontal line profile across the zoomed-in image to analyze its geometry and depth and save the line profile image as 'line_profile.png'. Based on the indentation shape and the line profile, determine whether a Vickers or conospherical indenter was used.



The task of generating the zoomed-in image and its horizontal line profile has been successfully accomplished. The line profile image has been saved as 'line_profile.png'. Now, let's analyze the line profile to determine the type of indenter used.

Line Profile Analysis

- 1. Line Profile Characteristics:**
 - Geometry and Depth: The shape of the line profile shows a symmetrical pattern with a sharp peak at the center. This suggests a pyramidal shape typical of a Vickers indentation.
 - 2. Indenter Type Determination:**
 - Vickers Indenter Characteristics:** It creates a diamond-shaped indentation with a sharp peak and distinct edges when viewed as a line profile.
 - Conospherical Indenter Characteristics:** It creates a rounded indentation with smooth slopes and no distinct peak.
- Given the sharp peak and defined edges seen in the horizontal line profile, it is more consistent with a Vickers indenter rather than a conospherical indenter.
- Final Verdict:**
- Indenter Used:** Based on the visual and line profile analysis, a **Vickers indenter** was used for this indentation experiment.

AILA



Towards Industrial Scale

Automation at Plant Level?

Cement manufacturing

- Cement manufacturing: an extremely complex process
- ~8% global CO₂ emissions
- Develop a digital twin for process optimization and quality control in collaboration 5 cement companies round the globe

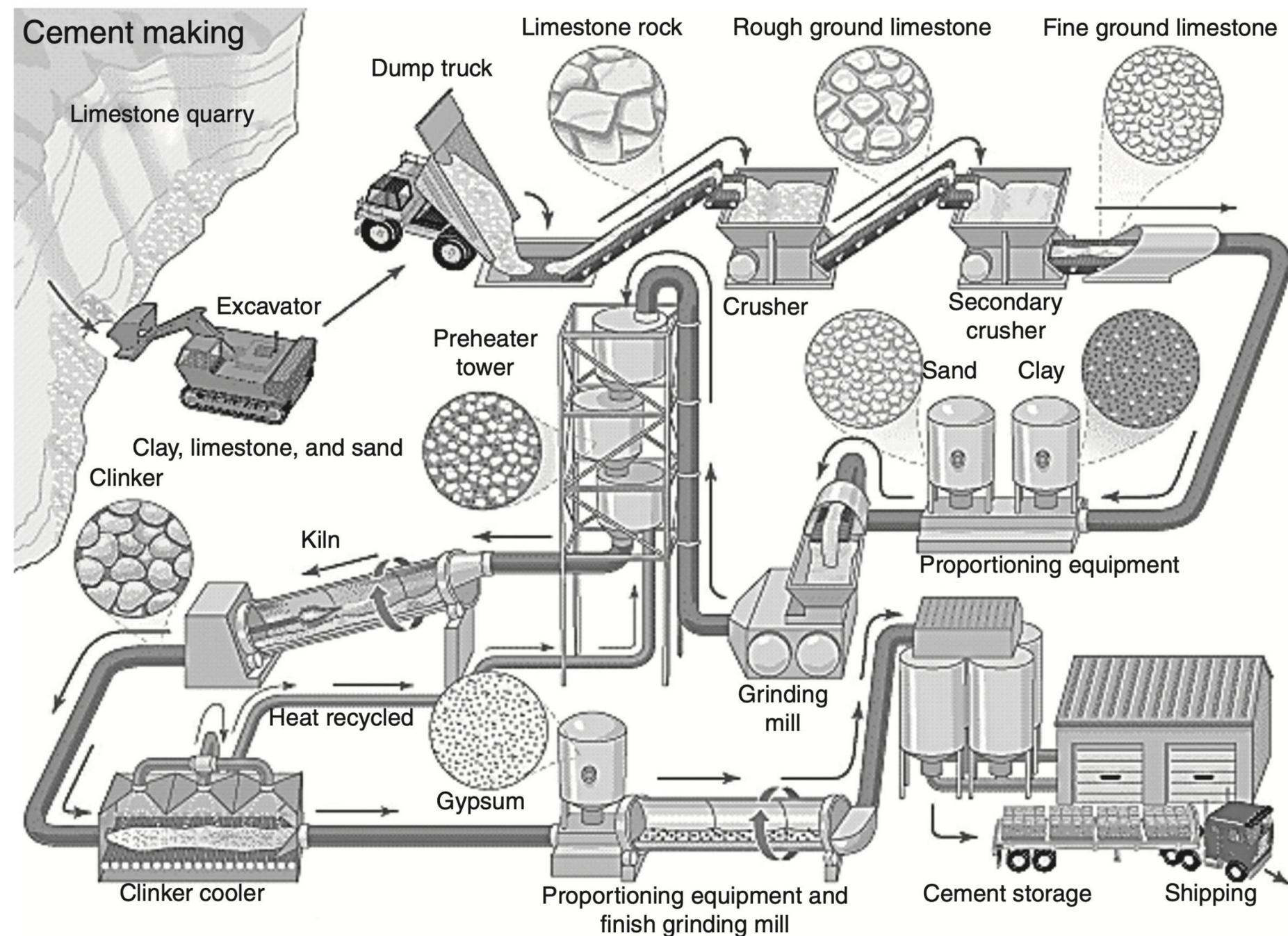
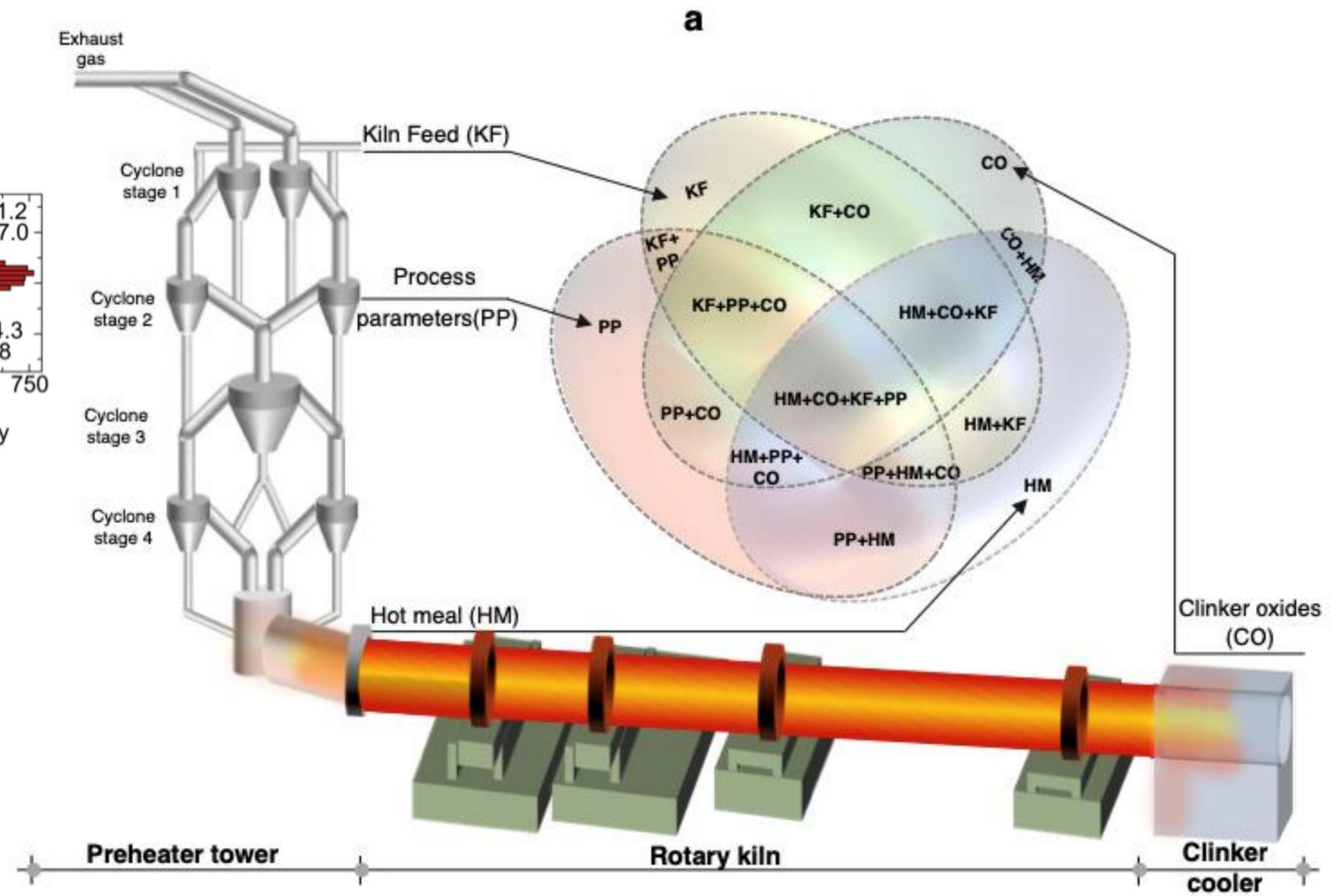
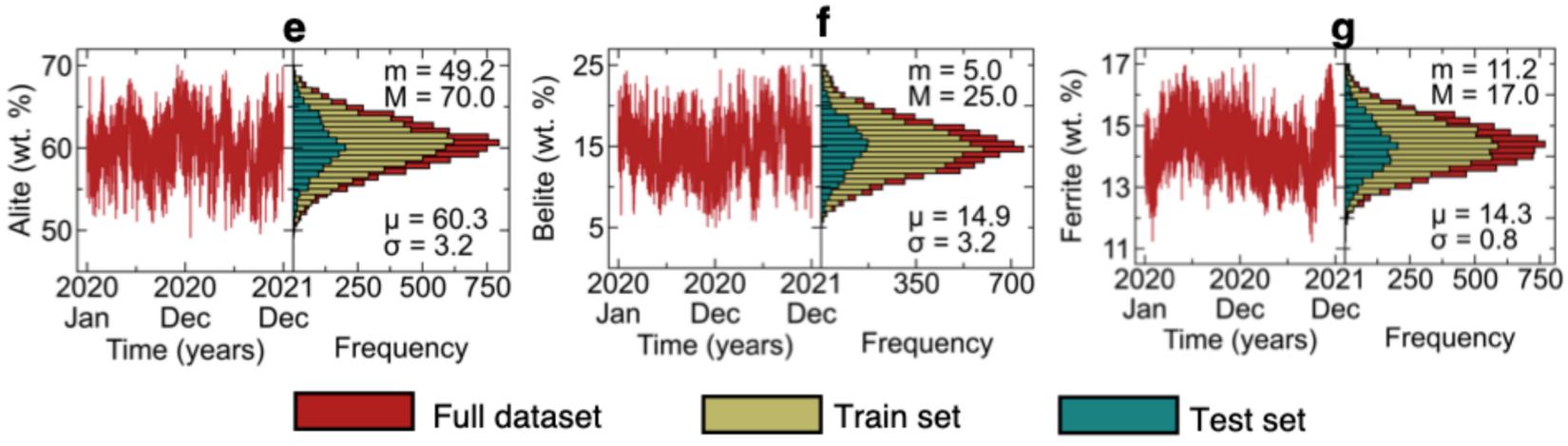
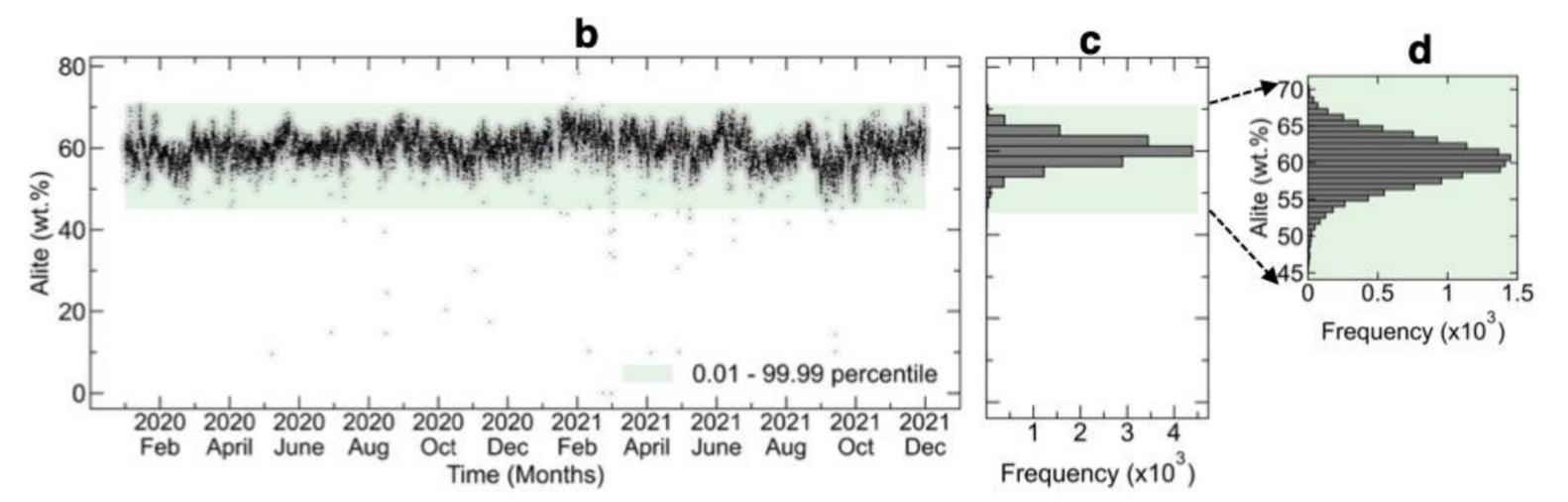


Image Source: Encyclopædia Britannica, Inc.

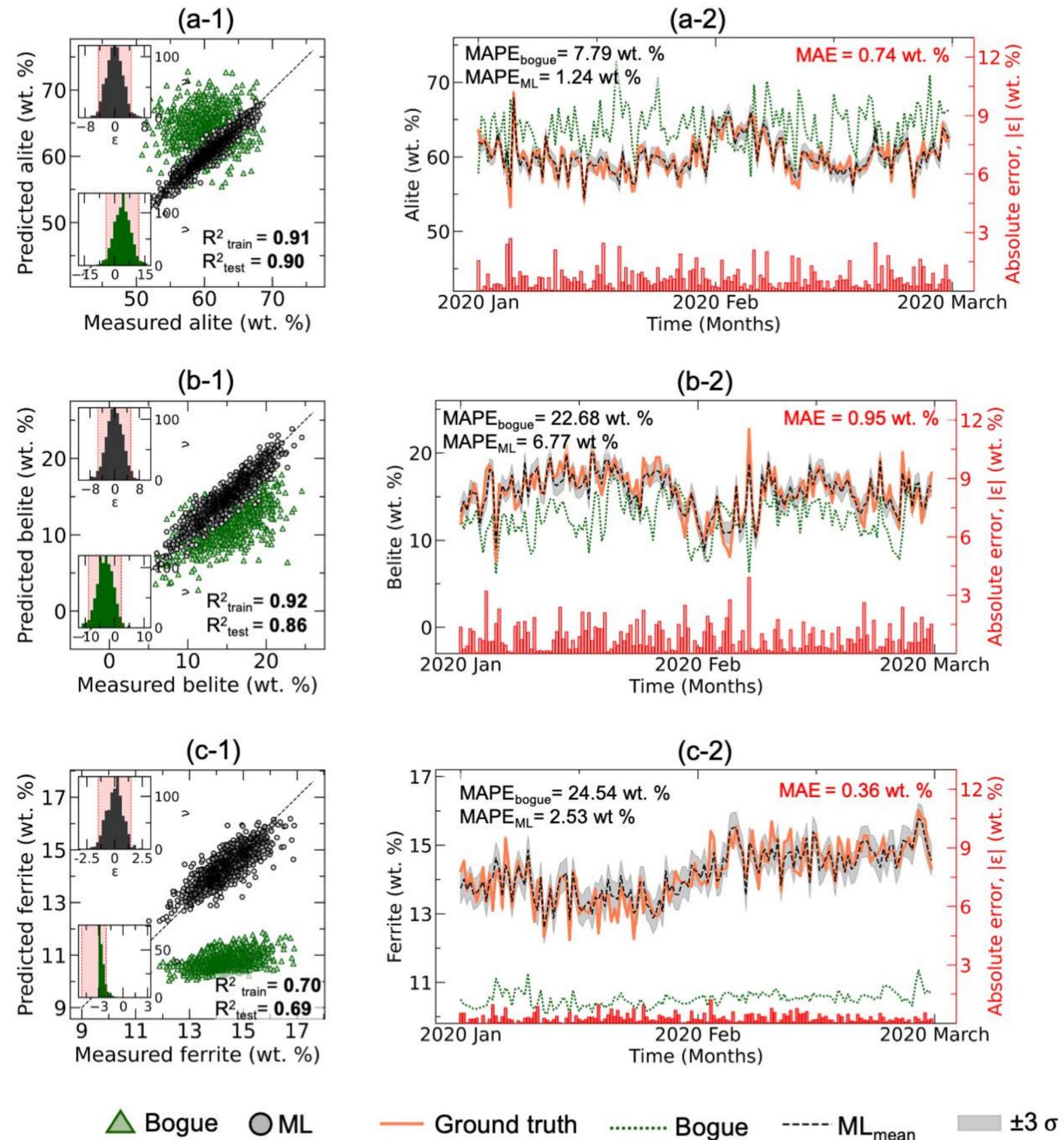
Model features



- Input features include raw materials, fuel, process parameters
- Goal: Clinker mineralogy
- Major challenge: Data cleaning and alignment



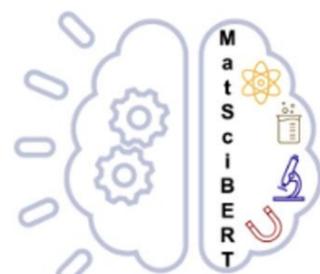
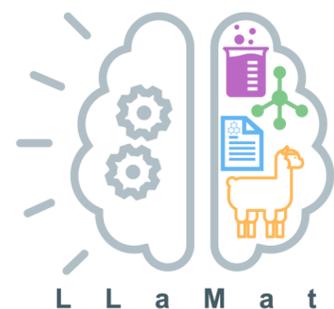
Model performance



- Accurately forecast clinker compositions
- Allows direct control of the plant for optimization
- Performs better than the models employed in the plant

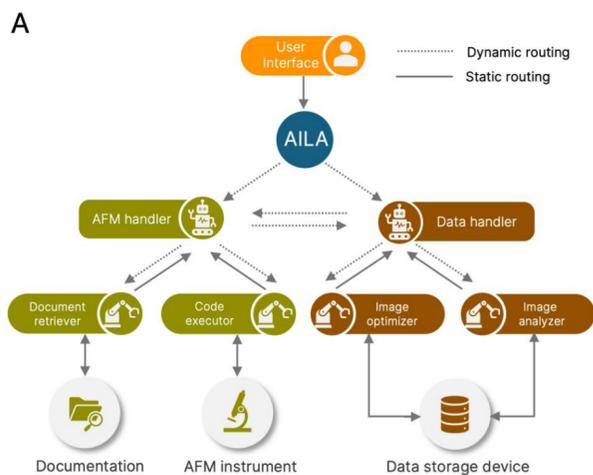
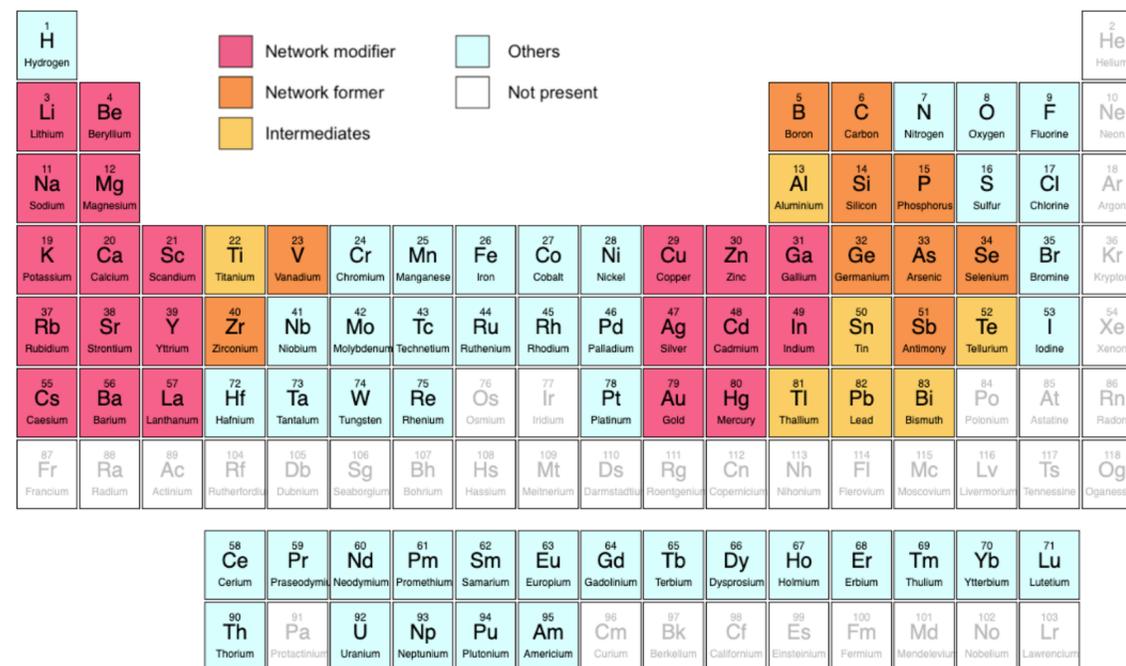


AI in Materials and Scientific Discovery



MatSciBERT-
Material
science aware
Language
Model

Scientific NLP
Information extraction
and materials
prediction



Industrial Scale
Manufacturing

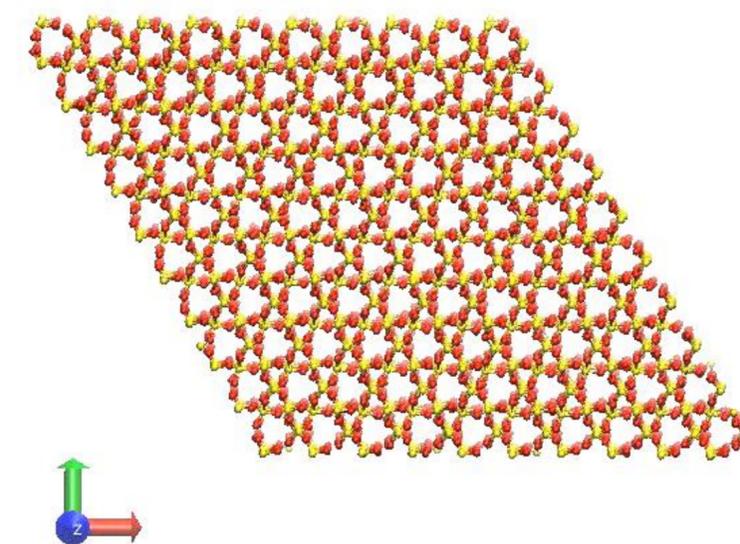
AI-driven process
optimization at
industrial scale

Self-driving
labs

Synthesis and
characterization

Materials
simulation

AI-driven
simulations and
analysis





Alexander von
HUMBOLDT
STIFTUNG



Google DeepMind



Undergraduate and doctoral researchers at M3RG, IIT Delhi

Thank you!



High Performance Computing at IITD

Supercomputer @ IITD : The Central Hybrid Supercomputing Cluster at Indian Institute of Technology Delhi



Machine Intelligence for Materials Science

N. M. Anoop Krishnan · Hariprasad Kodamana · Ravinder Bhattoo

Machine Learning for Materials Discovery

Numerical Recipes and Practical Applications

Focusing on the fundamentals of machine learning, this book covers broad areas of data-driven modeling, ranging from simple regression to advanced machine learning and optimization methods for applications in materials modeling and discovery. The book explains complex mathematical concepts in a lucid manner to ensure that readers from different materials domains are able to use these techniques successfully. A unique feature of this book is its hands-on aspect—each method presented herein is accompanied by a code that implements the method in open-source platforms such as Python. This book is thus aimed at graduate students, researchers, and engineers to enable the use of data-driven methods for understanding and accelerating the discovery of novel materials.

Krishnan · Kodamana · Bhattoo



Machine Learning for Materials Discovery

Machine Intelligence for Materials Science

N. M. Anoop Krishnan
Hariprasad Kodamana
Ravinder Bhattoo

Machine Learning for Materials Discovery

Numerical Recipes and Practical Applications



ISBN 978-3-031-44621-4



springer.com