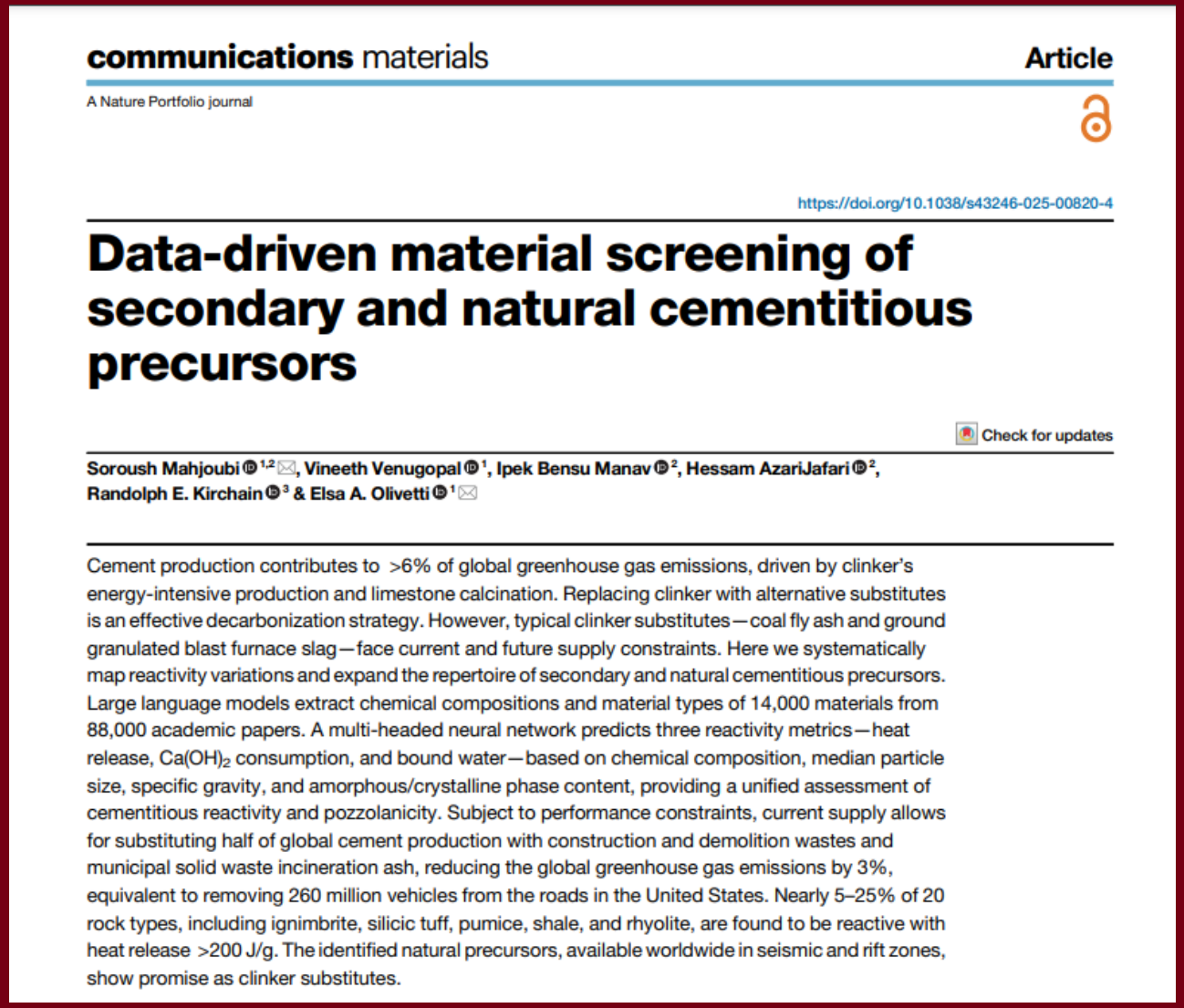
Massachusetts Institute of Technology

Data-driven material screening of secondary and natural cementitious precursors



communications materials

A Nature Portfolio journal

Article

https://doi.org/10.1038/s43246-025-00820-4

Data-driven material screening of secondary and natural cementitious precursors

Check for updates

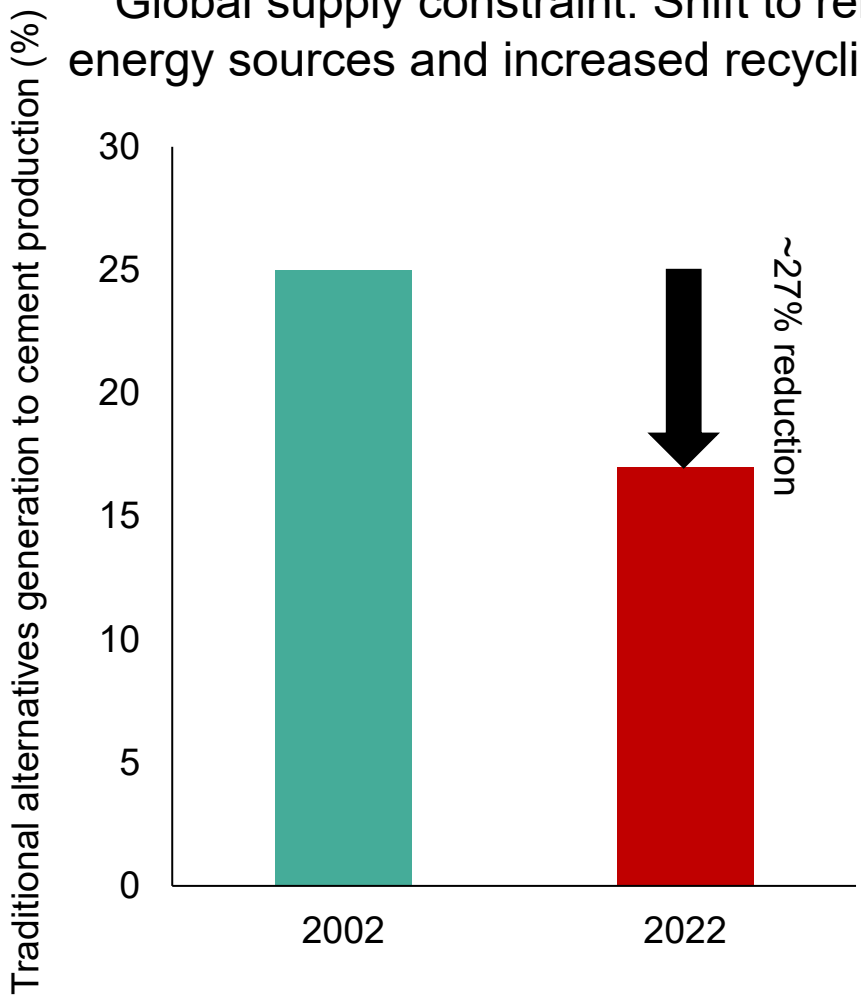
Soroush Mahjoubi[1,2], Vineeth Venugopal[1], Ipek Bensu Manav[2], Hessam AzariJafari[2], Randolph E. Kirchain[3] & Elsa A. Olivetti[1]

Cement production contributes to >6% of global greenhouse gas emissions, driven by clinker's energy-intensive production and limestone calcination. Replacing clinker with alternative substitutes is an effective decarbonization strategy. However, typical clinker substitutes—coal fly ash and ground granulated blast furnace slag—face current and future supply constraints. Here we systematically map reactivity variations and expand the repertoire of secondary and natural cementitious precursors. Large language models extract chemical compositions and material types of 14,000 materials from 88,000 academic papers. A multi-headed neural network predicts three reactivity metrics—heat release, $Ca(OH)_2$ consumption, and bound water—based on chemical composition, median particle size, specific gravity, and amorphous/crystalline phase content, providing a unified assessment of cementitious reactivity and pozzolanicity. Subject to performance constraints, current supply allows for substituting half of global cement production with construction and demolition wastes and municipal solid waste incineration ash, reducing the global greenhouse gas emissions by 3%, equivalent to removing 260 million vehicles from the roads in the United States. Nearly 5–25% of 20 rock types, including ignimbrite, silicic tuff, pumice, shale, and rhyolite, are found to be reactive with heat release >200 J/g. The identified natural precursors, available worldwide in seismic and rift zones, show promise as clinker substitutes.

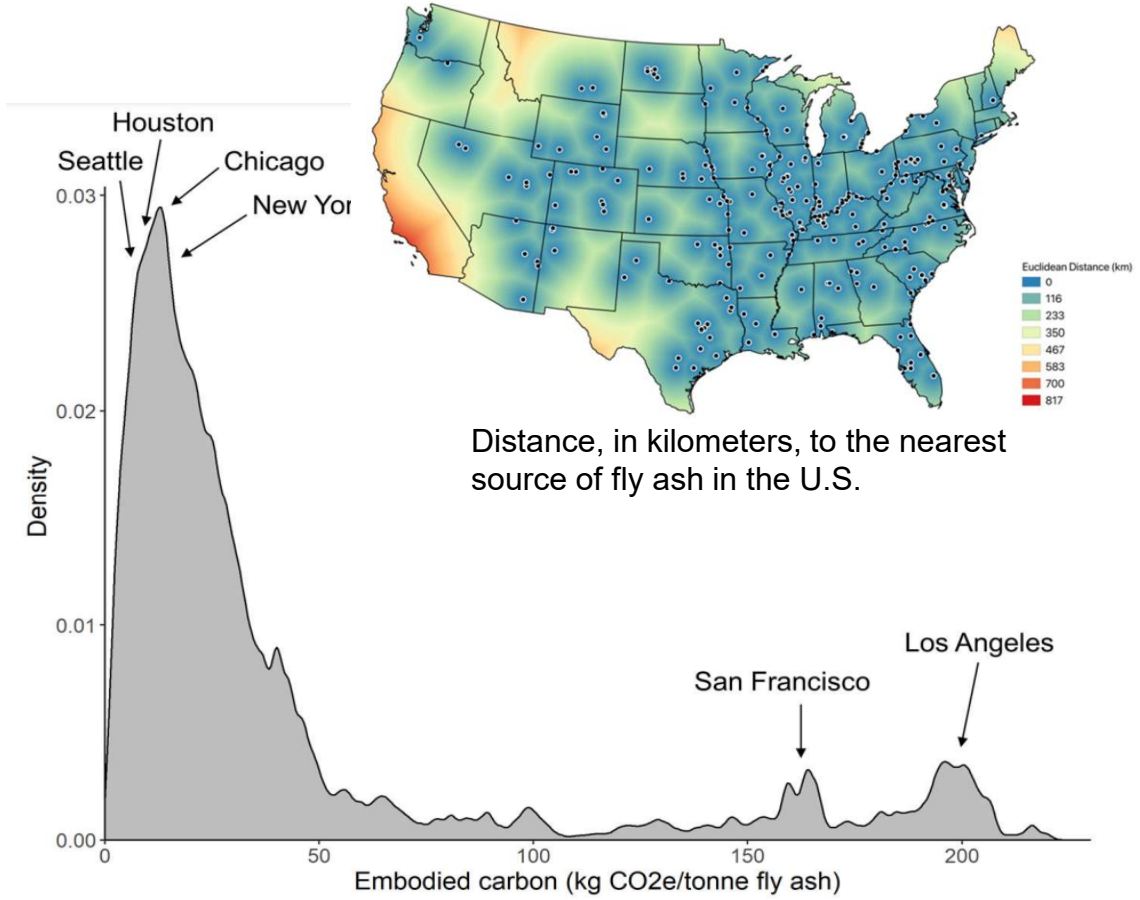
Soroush Mahjoubi (mahjoubi@mit.edu)
Postdoctoral associate at Olivetti Group & Concrete Sustainability Hub

SCMs can be deployed now, but traditional sources (virgin coal fly ash & blast furnace slag) cannot scale much beyond current levels

Global supply constraint: Shift to renewable energy sources and increased recycling of steel

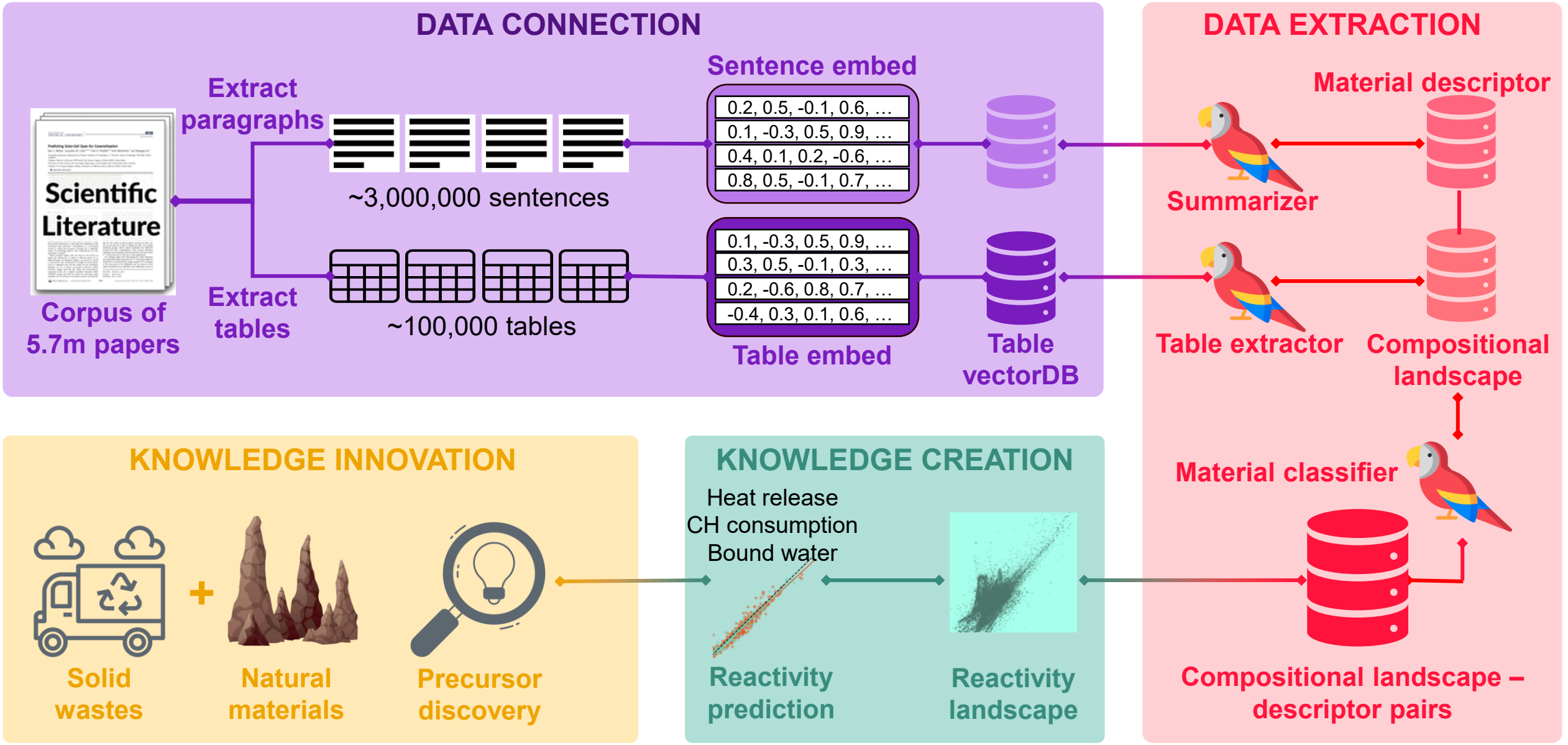


Density function of embodied carbon due to domestic fly ash transportation

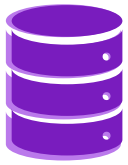


Distance, in kilometers, to the nearest source of fly ash in the U.S.

Framework: From information extraction to material screening



Communicative LLM agents: Extraction and classification



Embedding-based retriever

'Chemical and physical characteristics of the Kutahya (Type F): SiO2,Al2O3,CaO,MgO,SO3,MnO,K2O,P2O5,loss of ignition,58.07,17.09, 63.10,1.20,3.30,0.09,-,0.43,0.27,0.98

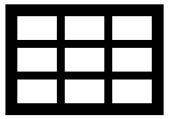
XML table with relevant information

ASTM Class C (high lime, CaO >10%) and Class F (low lime, CaO <10%) fly ashes were obtained from the Soma and Kutahya thermal power plants, Turkey, respectively.

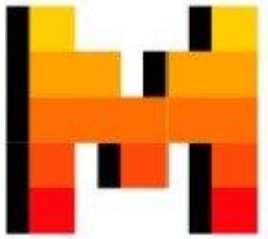
The maximum UCS values were attained in Group 6 for both Class C and F fly-ash mixtures. Zero-slump compositions were found to be more effective in Group 6.

Relevant sentences that describe the material (n = 10)

LLM: Extractor



Structured table of chemical compositions



Summarizer: Mistral 7b pre-trained

Kutahya fly ash (Type F) is a mineral rich in calcium oxide (~10%), obtained from Kutahya thermal power plant in Turkey, with a high Blaine fineness and low insoluble silicon dioxide content compared to Soma fly ash (Type C).

Descriptor of material

LLM: Classifier

Ash, Fly ash, Type F, Low calcium, high Blaine fineness, low insoluble silicon dioxide content

Categories, subcategories

- 88,000 research papers analyzed for information extraction
- 100,000 tables considered from the papers for extracting information
- ~3 mil sentences extracted from the papers

Generation: From XML to specific, structured format

- LLMs considered for comparative study: Fine-tuned (FT), 8-bit quantized fine-tuned (QFT), and pre-trained (PT)
- Evaluation of retrieval-to-generation is conducted in three levels:
  (1) Relevance: Detection of relevance,
  (2) Materials: Detection of materials,
  (3) Compositions tuples: Extraction of numerical chemical compositions

Quality of generation: metrics

Model name	Type	Params	Performance metrics								Average end-to-end time (s)
			Relevance - F1		Materials - F1		Compositions - F1		Compositions - RMSE		
			Train	Test	Train	Test	Train	Test	Train	Test	
ChatGPT 4	PT	1.5T	0.84	0.94	0.52	0.81	0.48	0.72	20.11	9.69	25.74*
ChatGPT 3.5	PT	175B	0.36	0.84	0.86	0.57	0.96	0.90	2.55	3.96	**6.35***
ChatGPT 3.5	FT	175B	0.92	**0.95**	**1.00**	0.83	**0.99**	**0.93**	**1.49**	2.81	6.87*
Mistral v0.2	PT	7B	0.28	0.84	0.88	0.69	0.84	0.78	7.04	9.89	11.42
Mistral v0.2	QFT	7B	**0.94**	0.91	0.92	**0.85**	0.95	0.89	3.24	3.38	13.58

Material description using Mistral 7b pretrained

'Chemical and physical characteristics of the Kutahya: SiO2,Al2O3,Fe2O3,CaO,MgO,SO3,TiO2,MnO,K2O,Na2O,loss of ignition,58.07,17.09,10.27,63.10,1.20,3.30,0.09,-,0.43,0.27,0.98

XML table

Fly Ash. Two types of ASTM Class C and F fly ash from Soma and Kütahya Thermal Power Plants in Turkey were used.

The pozzolanic activity indices of Soma and Kütahya fly ashes are 109.5 % and 74.5 % respectively according to ASTM C 311..
…

Embedding-based retrieval: 10 relevant sentences in the DOI

Mistral 7b pre-trained

Kutahya fly ash (Type F) is a mineral rich in calcium oxide (~10%) with a high Blaine fineness and low insoluble silicon dioxide content compared to Soma fly ash (Type C).

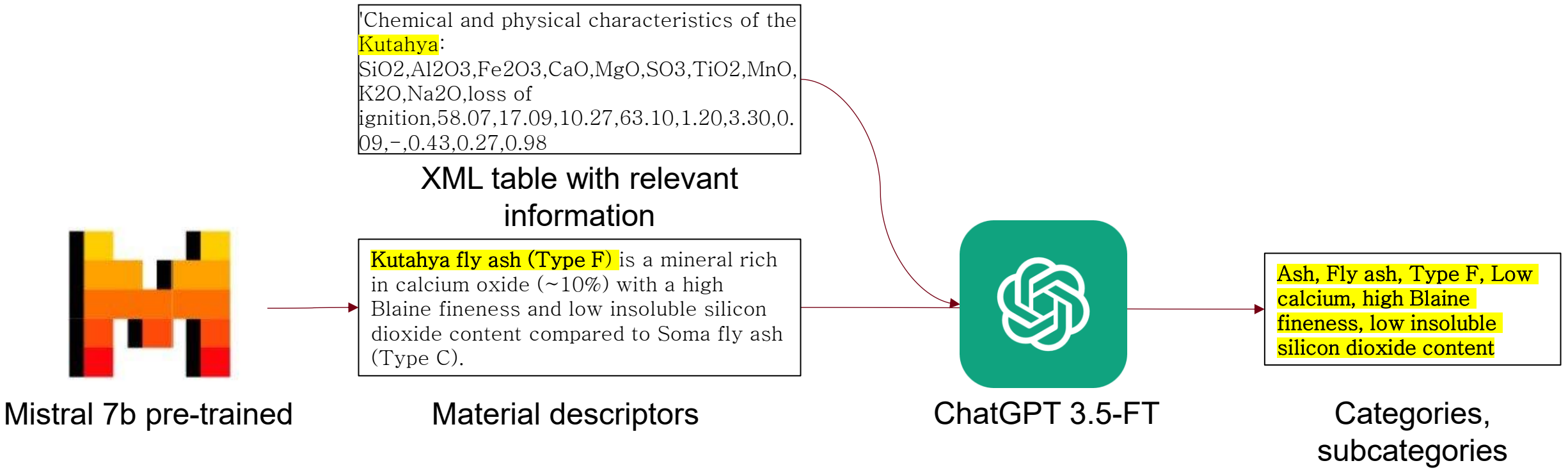
Material descriptors

Extracted material descriptors

Material name	Response
L	Limestone filler (L) is a calcium-rich material with 26.09% CaO content.
F	Fly ash (F) is a mineral constituent with a chemical composition of 51.97% SiO2, primarily derived from the combustion of coal in power plants.
OPC	OPC, or Ordinary Portland Cement, is a binding material with a primary chemical composition of approximately 67.5% calcium oxide.
S	Blast furnace slag (S) is a mineral byproduct from iron production containing approximately 45.5% calcium oxide (CaO).
Acc	The material named "Acc" is a set accelerator in concrete compositions with a solid content of 60%, primarily containing magnesium oxide at 0.70%.
GBFS	GBFS, or Ground Granulated Blast Furnace Slag, is a byproduct from the production of iron in blast furnaces, having a major chemical composition of SiO2 (33.8%), CaO (42.6%), MgO (5.3%), Al2O3 (13.7%), Na2O (0.1%), K2O (0.4%), SO3 (0.8%), Other (1.5%), LOI (1.8%), totaling 100%.
Portugal	Portugal is a country with a cement composition having approximately 62% CaO, 5.35% Al2O3, 3.42% Fe2O3, 20.26% SiO2, 1.88% MgO, 0.11% Na2O, 0.98% K2O, 2.68% SO3, and specific surface area of around 330 m2 kg-1, along with significant amounts of C3S, C2S, C4AF, C3A, Pozzolanic materials, and free lime.
SC	SC is a cement type with a lower calcium content (47.4% vs 63.6% in PC) and higher specific surface area (496 m2/kg vs 352 m2/kg), resulting in faster initial and final setting times but lower compressive strengths at early ages compared to PC.

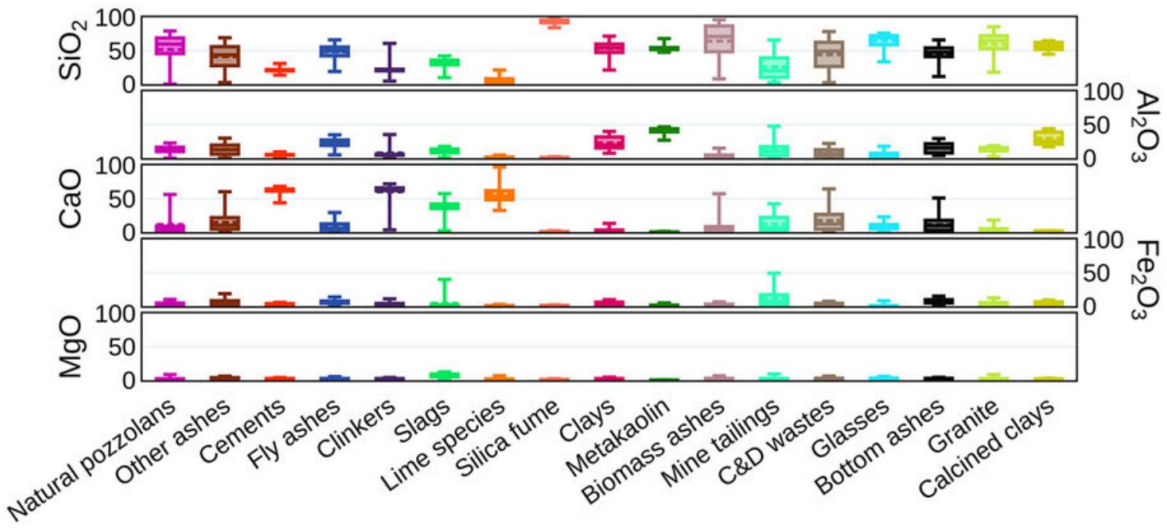
Material classification: Categories & beyond with communicative LLM agents

- A ground-truth dataset is manually generated using 300 descriptors
- Materials are categized into 19 categories based on the descriptors using fine-tuned ChatGPT 3.5
- Manual evaluation: Multiclass accuracy 97%

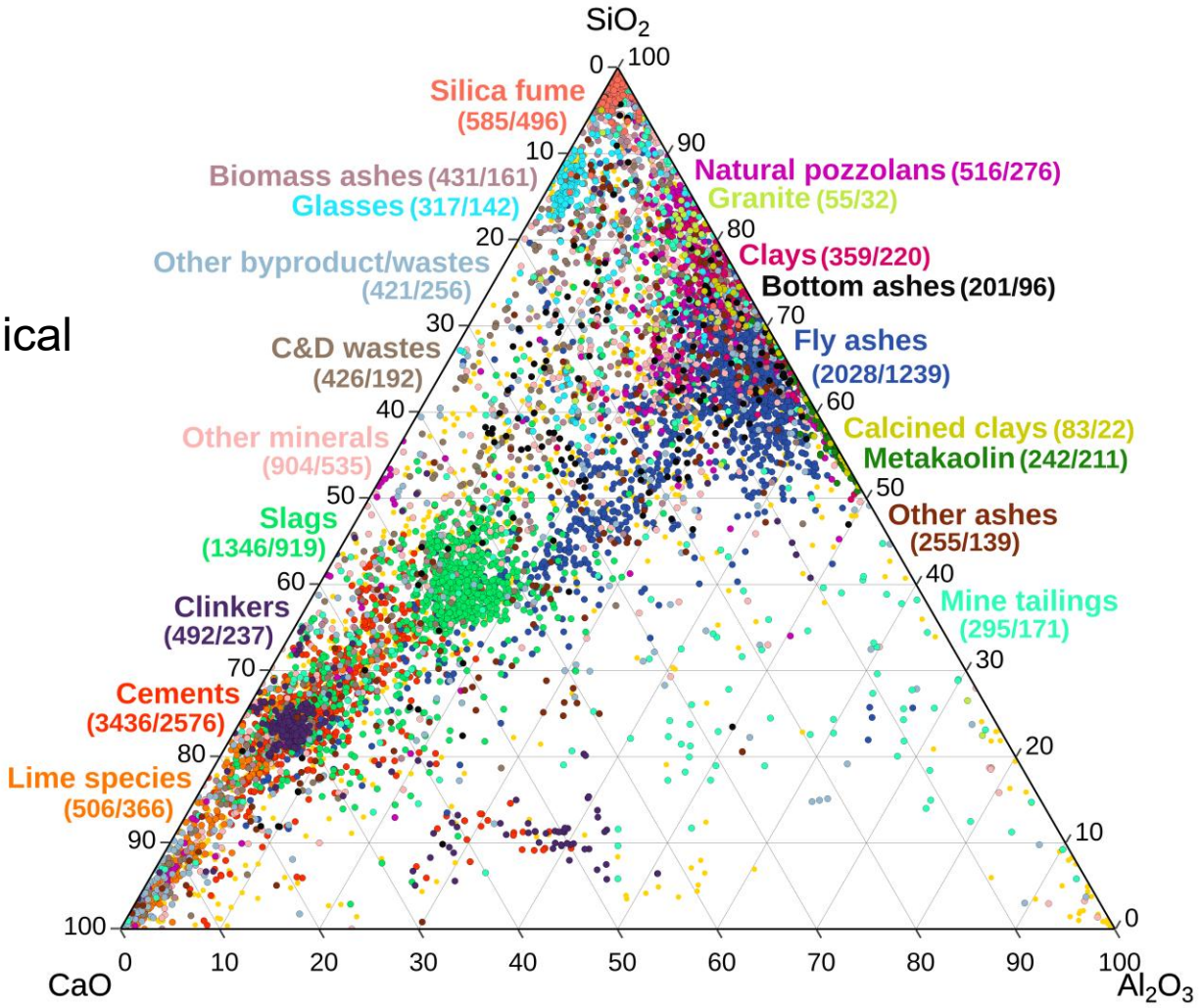


Compositional variations of materials and ternary diagram

- >17,000 materials were extracted from 6,312 journal papers
- Observed substantial variability in the materials' chemical compositions, especially for natural pozzolans and biomass ashes



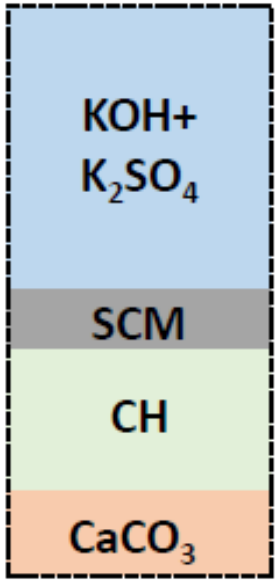
Compositional variations of the 19 pre-defined materials



Ternary diagram of the three major oxides

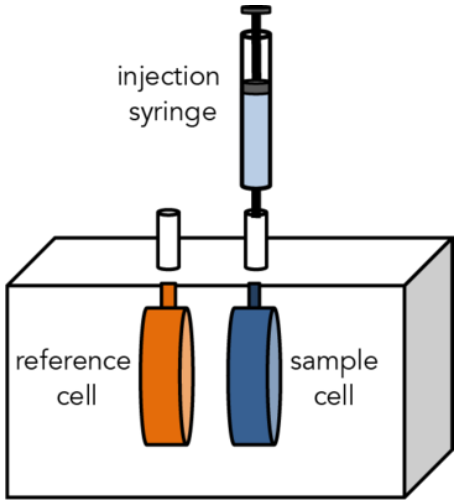
Reactivity prediction based on R3 test (ASTM C1897-20)

- Objective: Measures heat release, calcium hydroxide (CH) consumptions, and bound water in a SCM-calcium hydroxide system over days at a controlled temperature
- Significance: classify materials into inert and reactive

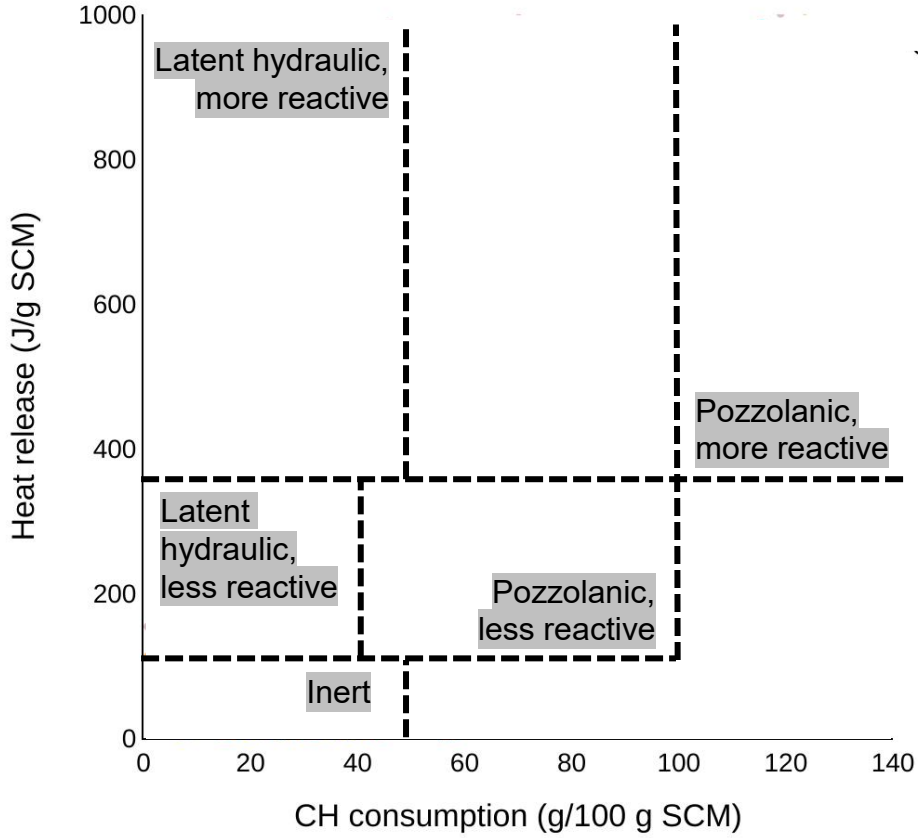


SCM paste

Thermogravimetric analysis to measure the CH consumption

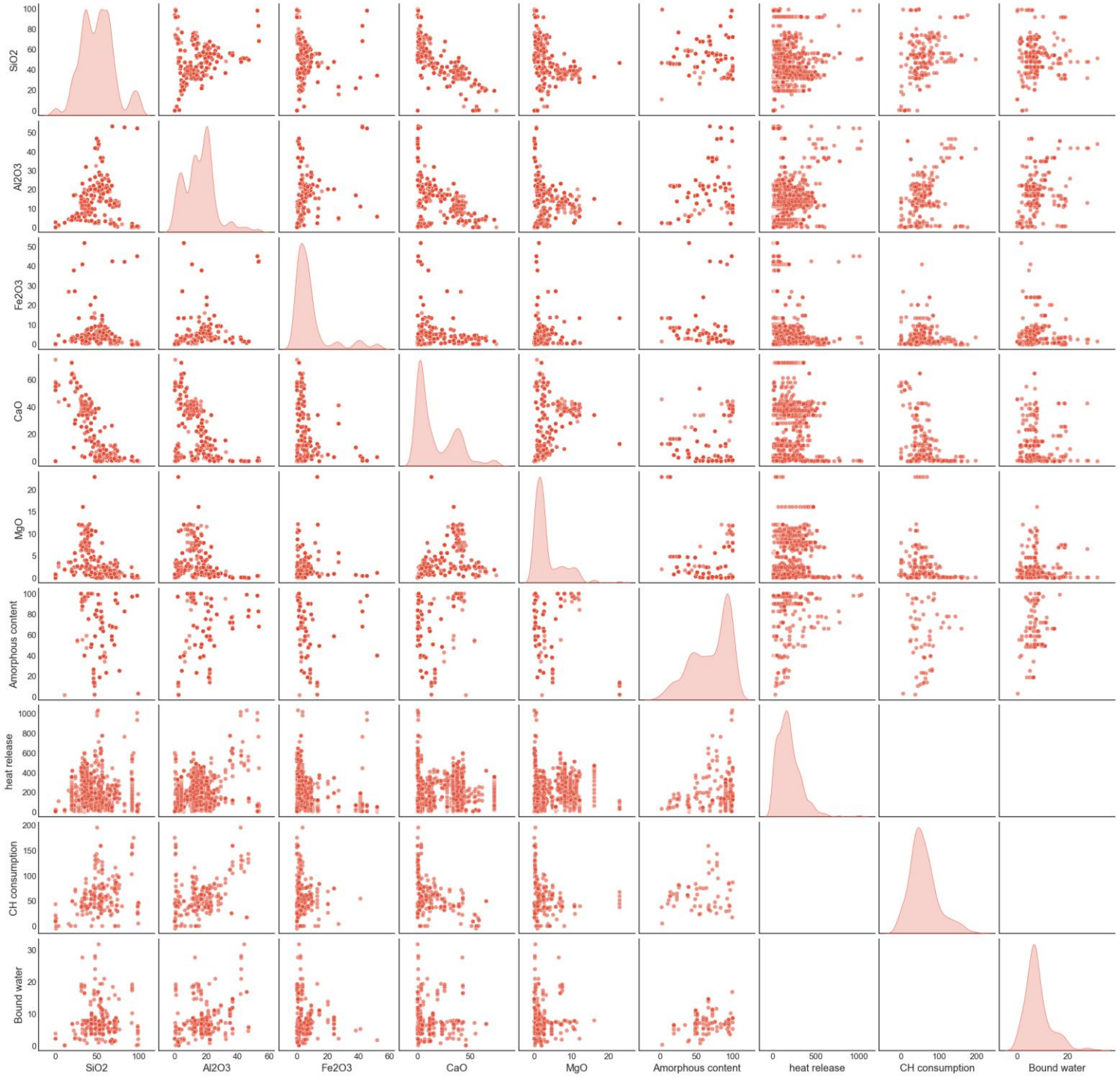


Isothermal calorimetry to measure the heat release

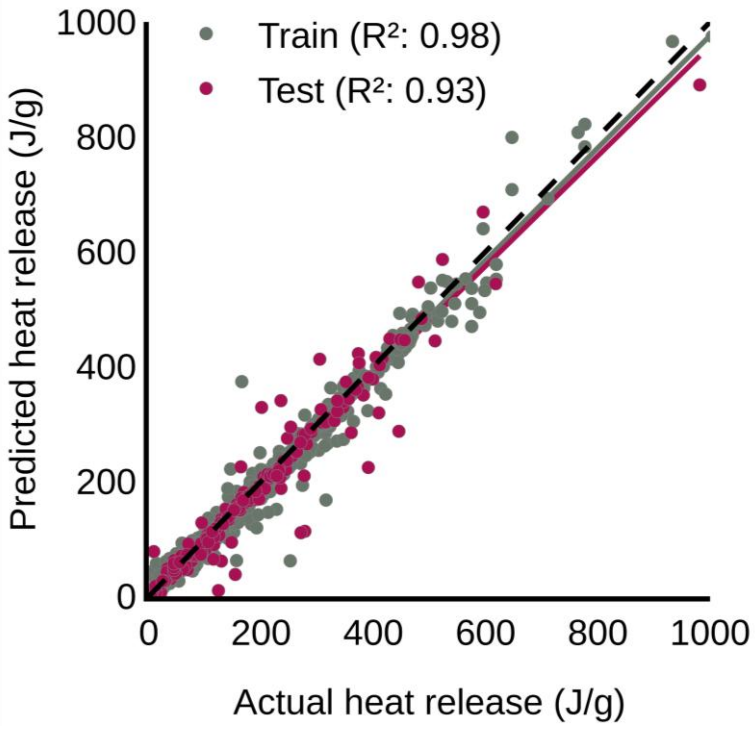


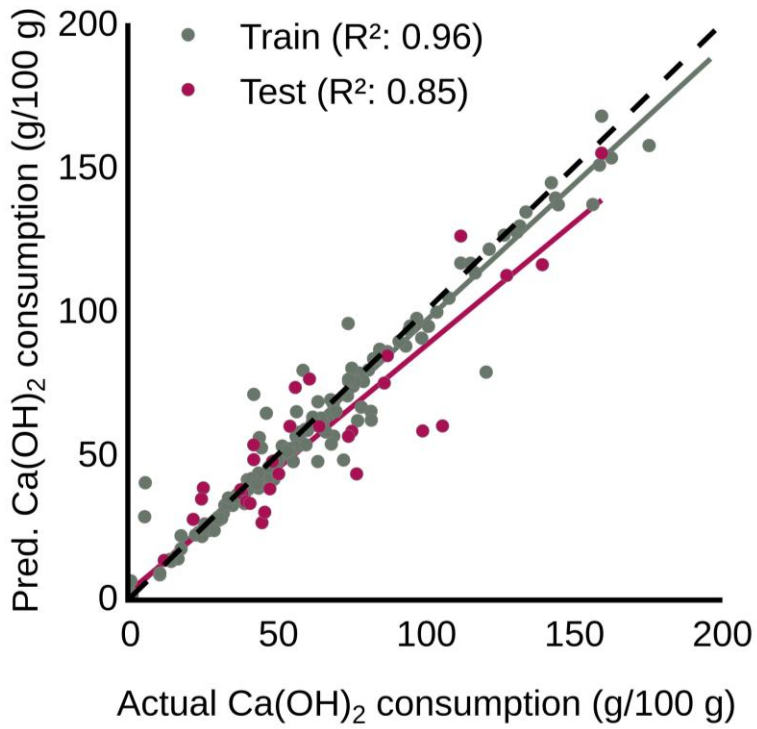
Dataset for reactivity prediction

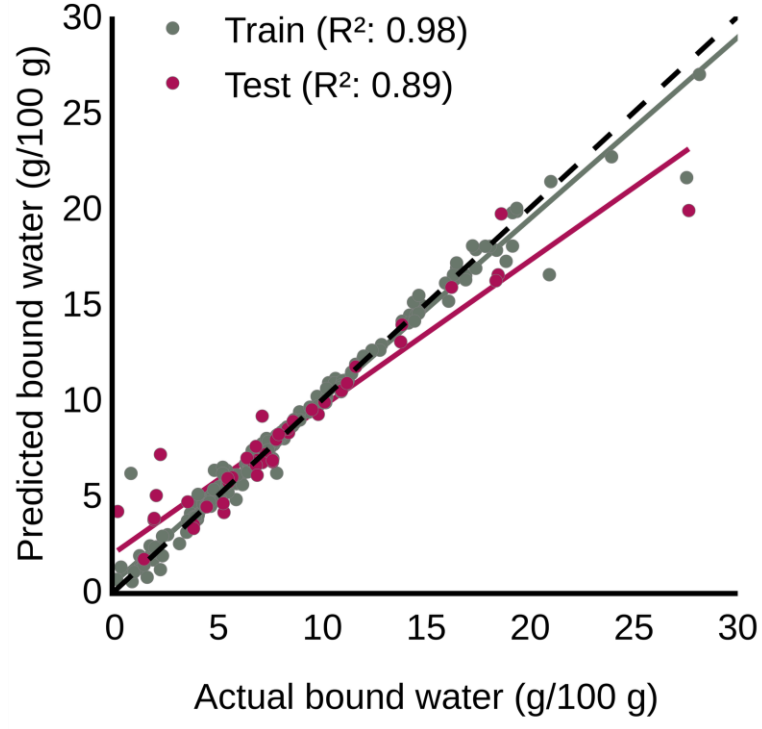
- ~1,800 data points are manually extracted from 30 publications for chemical reactivity prediction of SCMs
- Input variables: chemical compositions, specific gravity, mean particle size (D50), and amorphous/crystalline content
- Output variables: heat release, CH consumption, and bound water



The model accurately predicts heat release, $Ca(OH)_2$ consumption, and bound water content in concrete hydration reactions, showing high R^2 values across test (0.85–0.93) sets

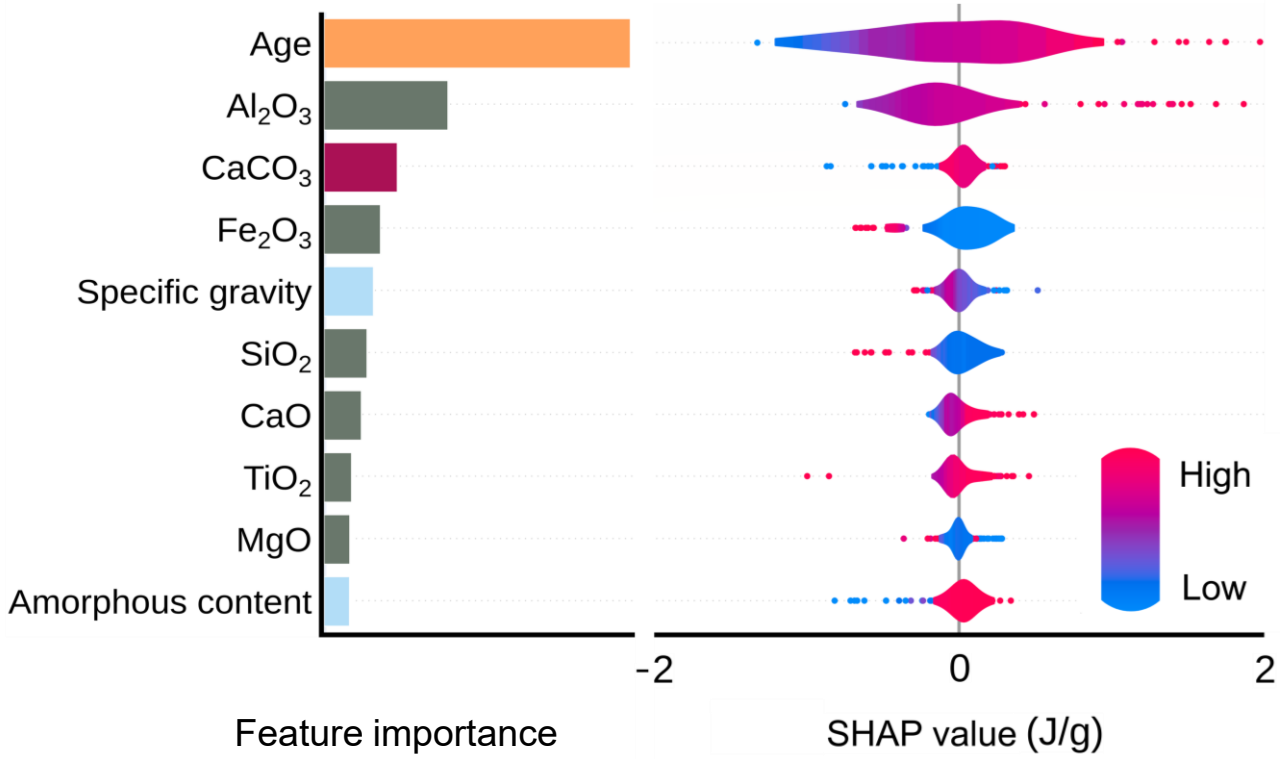






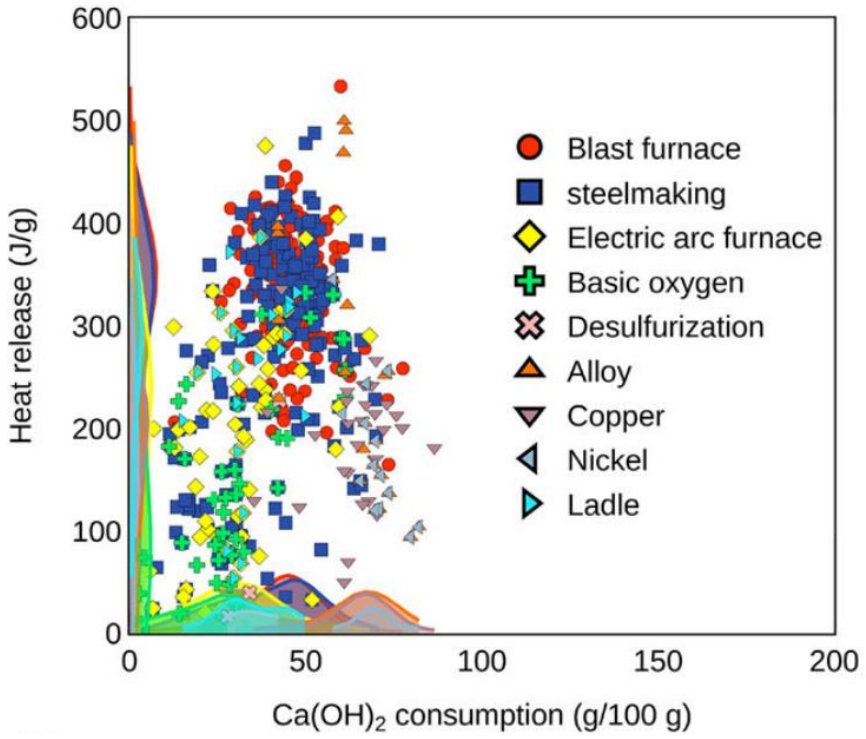
Explainable AI (XAI) was implemented to demystify the "black-box" nature of reactivity predictions, enabling users to understand the key drivers

- Age is the most influential factor driving heat release, with older ages strongly increasing predicted values.
- Al_2O_3 and CaO also contribute significantly, with higher contents of both oxides associated with increased heat release
- Other oxides (e.g. Fe_2O_3) and specific gravity show moderate to minor effects, with generally lower SHAP values

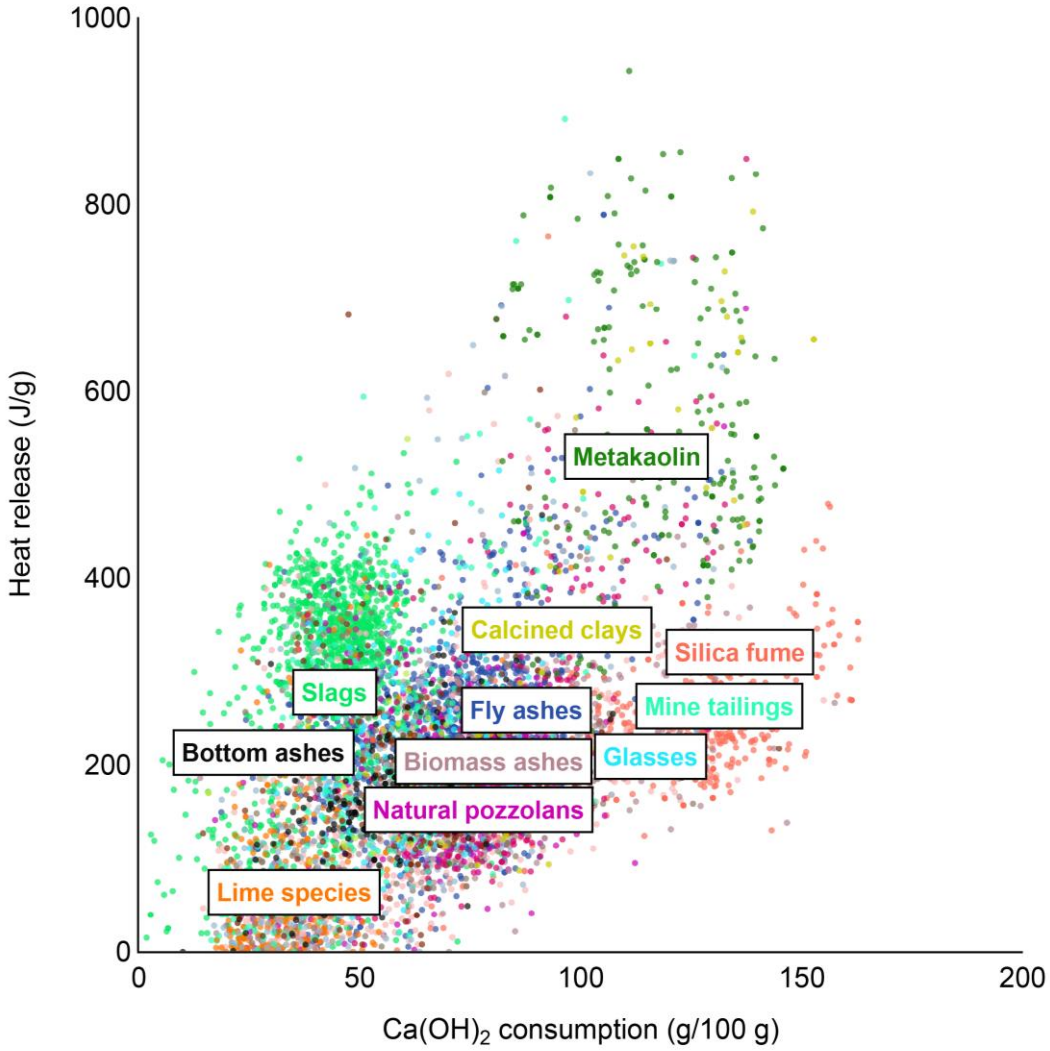


Shapley-based explanations for predicting the heat release

Reactivity landscape of materials



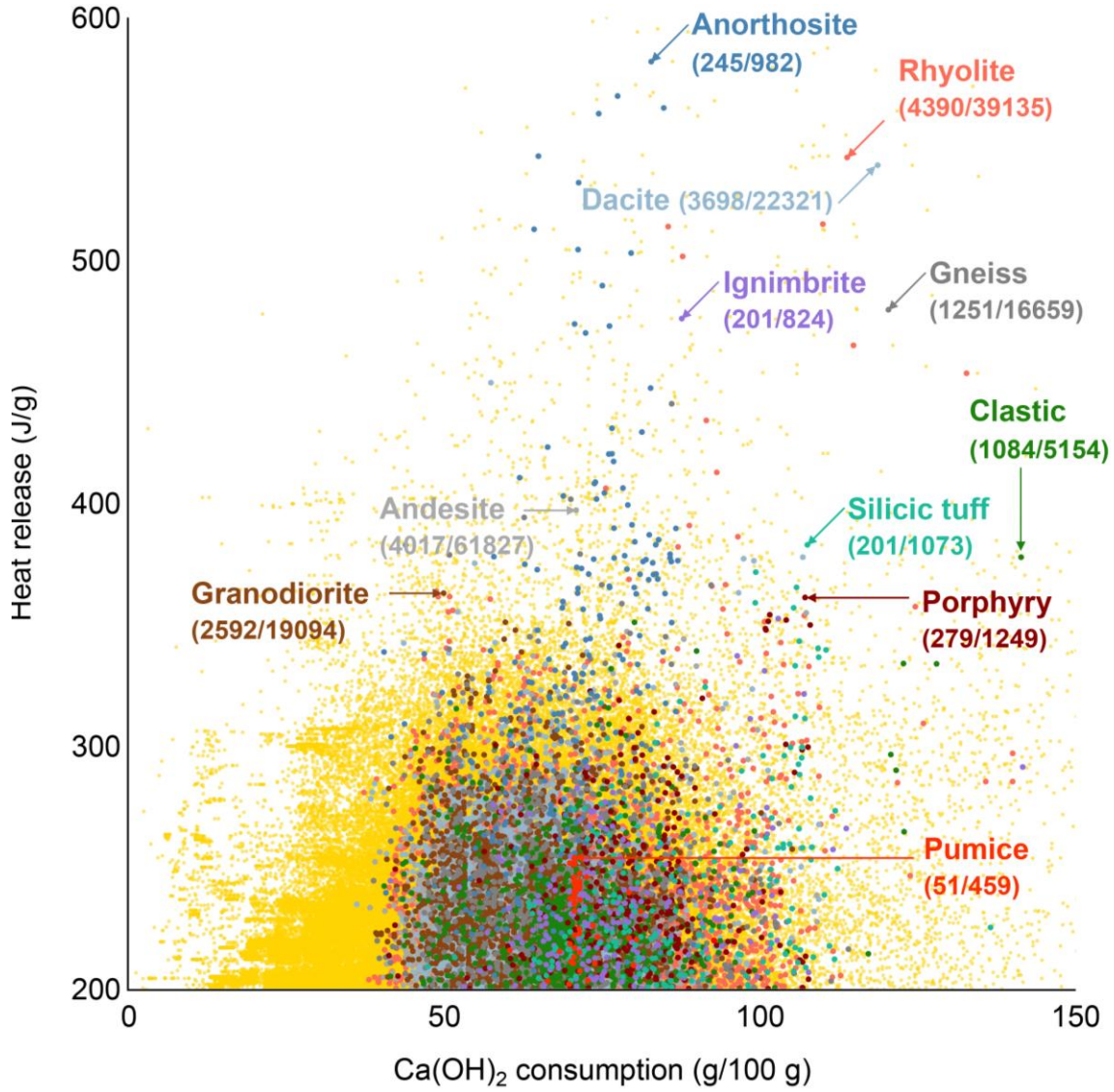
Slag reactivity varies significantly by subtype; *electric arc furnace* slags show a widespread in reactivity, highlighting intra-class variability



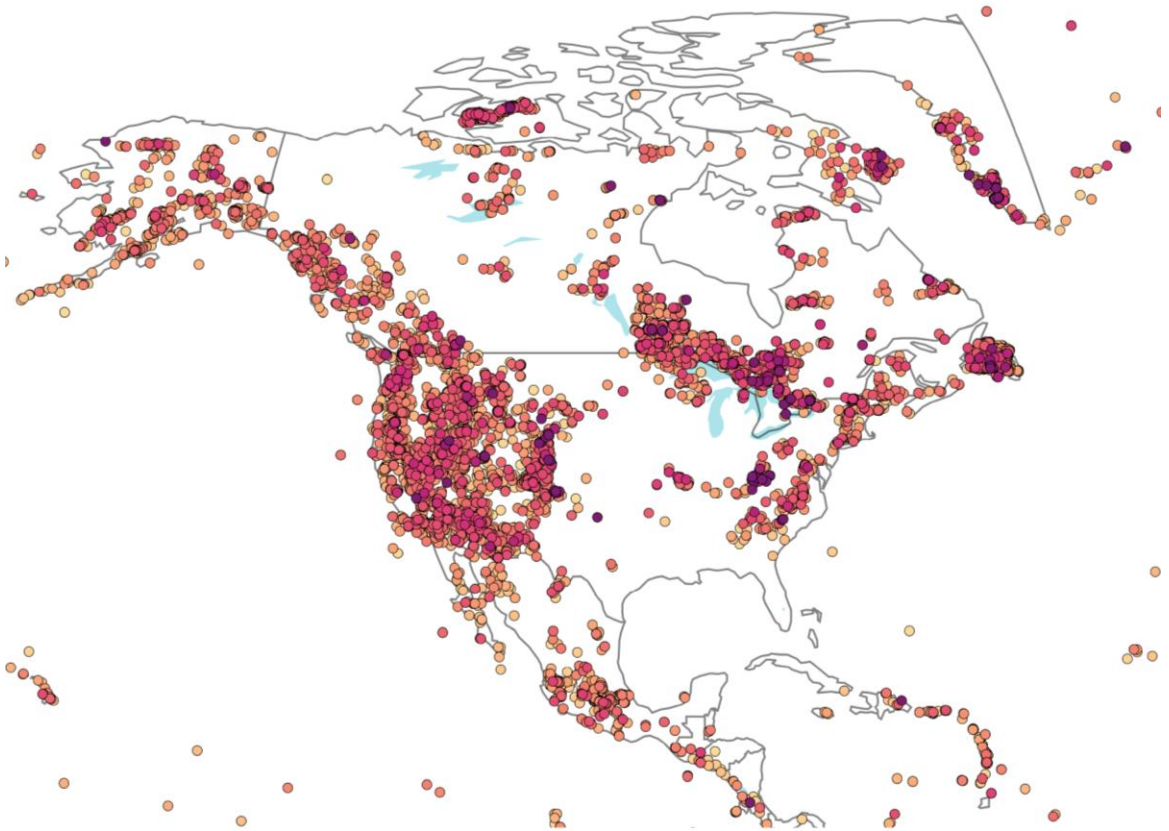
AI-predicted reactivity metrics reveal distinct clusters across SCM classes, with *metakaolin*, *silica fume*, and *calcined clays* exhibiting the highest heat release and $Ca(OH)_2$ consumption

Discovery of natural cementitious precursors

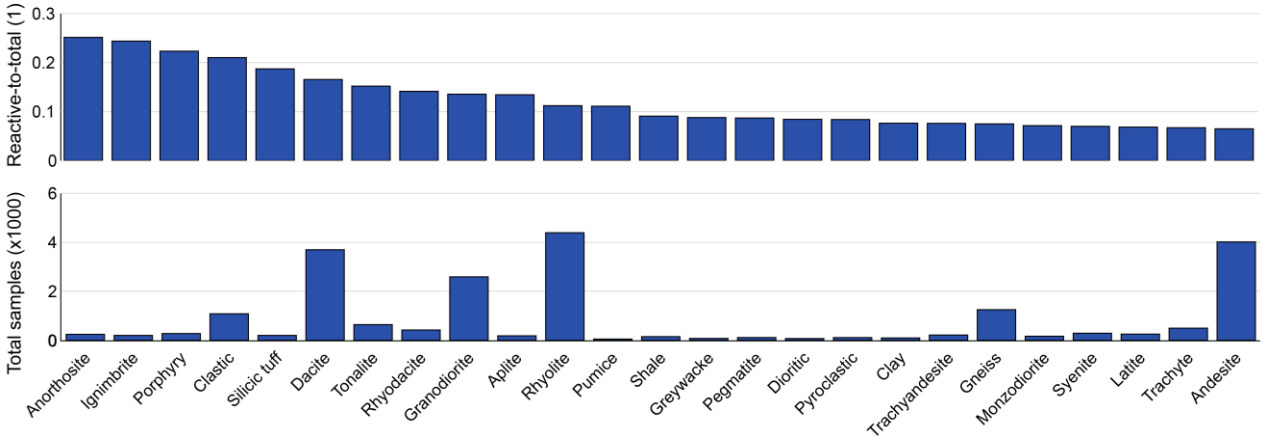
- 1,000,000 samples representing 16,000 distinct rock types were evaluated for reactivity
- 20 rock types were identified as potentially reactive, defined by having over 5% of their samples exhibiting reactivity
- In total, approximately 3% of all tested rock types can be classified as reactive



Geospatial and reactivity profiling of natural cementitious materials in the U.S.



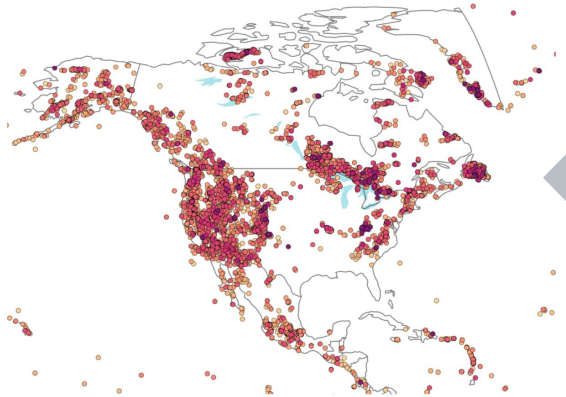
Over 6,000 geolocated natural samples across North America reveal widespread availability of potentially reactive volcanic and sedimentary rocks

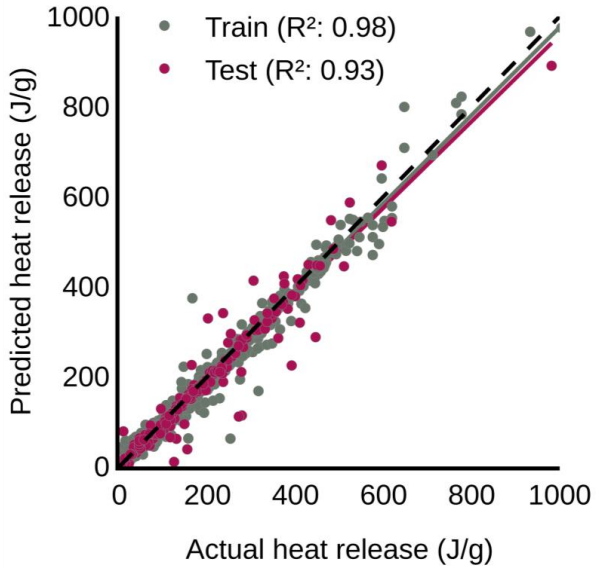


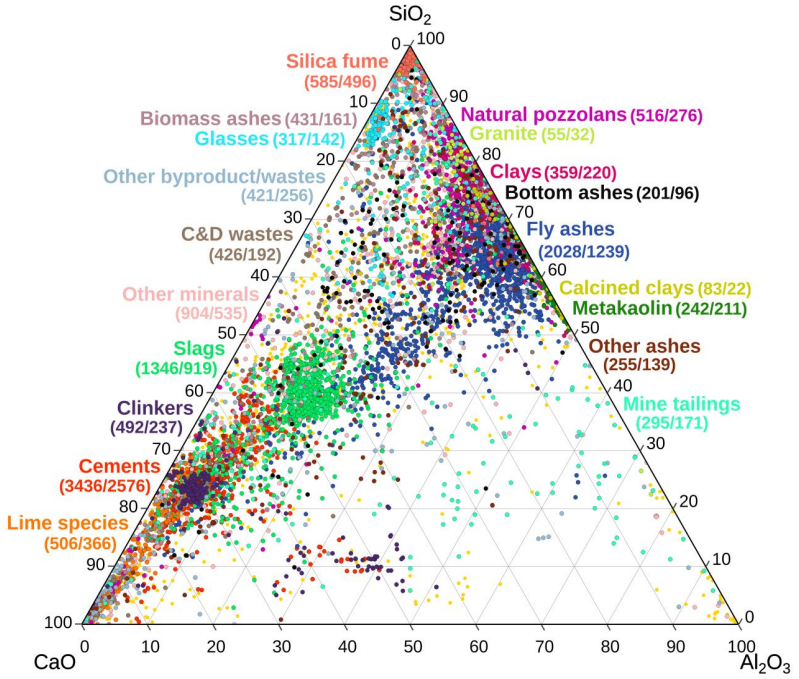
Andesite and rhyolite dominate in sample count, while anorthosite and ignimbrite show the highest average predicted reactivity

Key takeaways

- **14,434 unique materials** extracted with distinct chemical signatures from diverse sources.
- **LLM-powered classification** enabled accurate identification of material types and descriptors.
- **LLM-powered composition mining** enabled accurate extraction of chemical compositions.
- **Multi-output machine learning models** successfully predicted reactivity metrics.
- **Mapped and ranked natural precursors** across the U.S., unveiling high-potential candidates for low-carbon cementitious use.







Any questions?

mahjoubi@mit.edu

Quality assessment: Embedding-based retrieval

- Cumulative Gain: Quantifies the incremental value added by each correctly identified relevant table
- 993 out of the first 1,000 tables selected by an embedding model called all-MiniLM-L6-v2 is relevant for information extraction

Quality of retrieval: metrics

Embedding model	Cutoff	Precision	Recall	F1	CG*	NDCG‡
all-mpnet-base-v2	**10**	**1.00**	**0.01**	**0.02**	**10**	**1.00**
	100	0.97	0.10	0.19	97	0.98
	1000	0.94	1.00	0.97	940	0.94
all-MiniLM-L6-v2	**10**	**1.00**	**0.01**	**0.02**	**10**	**1.00**
	100	**0.99**	**0.10**	**0.18**	**99**	**0.99**
	1000	**0.99**	**1.00**	**1.00**	**993**	**0.99**
MatSciBERT	10	0.90	0.02	0.04	9	0.93
	100	0.74	0.15	0.26	74	0.78
	1000	0.48	1.00	0.65	484	0.52

*CG: cumulative gain; ‡NDCG: normalized discounted cumulative gain