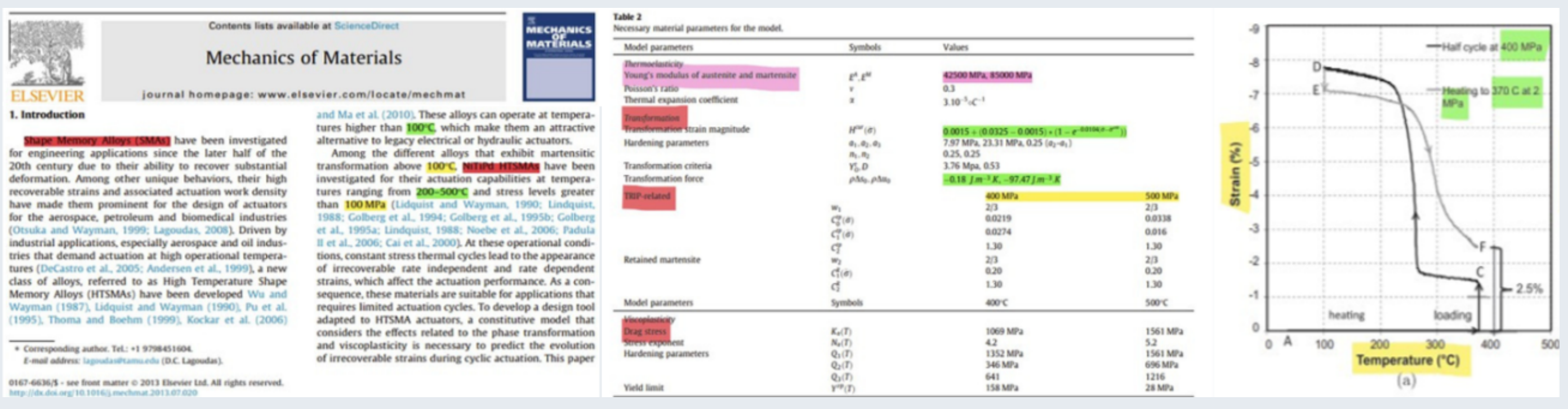
Development and Analysis of Multicomponent Alloy Database Using Large Language Models

Sai Mani Kumar Devathi[1] Aravindan Kamatchi Sundaram[1] B.Pabitramohan Prusty[1] Rohit Batra[1]

[1]Indian Institute of Technology Madras

Introduction

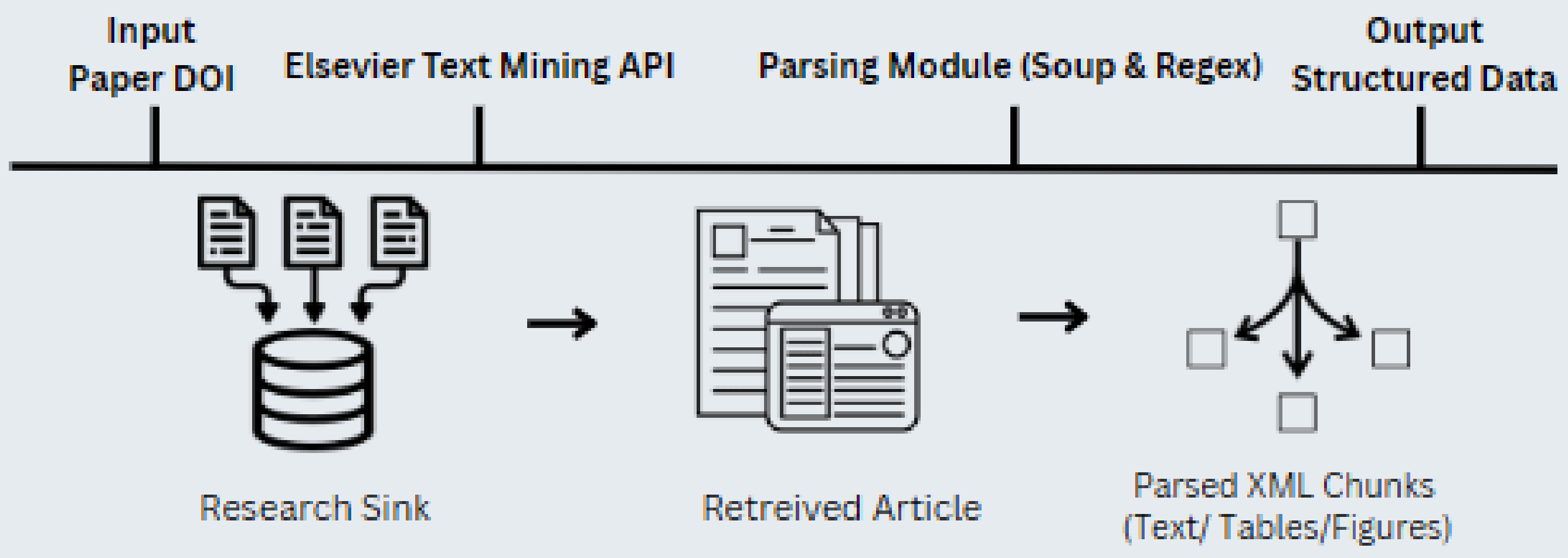
- ML-driven materials discovery is hampered by small, costly, manually curated datasets. Prior NLP/LLM approaches extract data from papers, but often miss critical alloy details or narrow scopes.
- We present an LLM-based extraction pipeline for both text and tables, tuned with prompt engineering and Retrieval Augmented Generation.
- Achieves F1 score **0.83** in textual fields (composition, processing, characterization, properties) and **0.96** in tabular data.
- We developed comprehensive methods to assess the LLMs' performance, testing the pipeline against existing alloy datasets.
- Applied to more than **10,000 papers**, producing the largest, most accurate public High Entropy Alloys dataset, readily extensible to polymers, MOFs, and ceramics.



Extractable Data from Research Articles

Methodology

1. Data Ingestion & Parsing



2. Text Extraction

- **Prompt Engineering:** Prompts consist of context, few–shot examples, and instructions.

Context

1. An alloy is a solid mixture of multiple elements. Its chemical composition conveys the amount of each element present in the alloy. For example, an alloy with chemical composition AlCoCr contains elements Al, Co and Cr in equal amount. Similarly, an alloy with chemical composition $Al_2Co_3Ni_5$ contains 2 parts Al, 2 parts Co and 5 parts Ni.

2. Here are some of the common structures adopted by alloys

...

n. Abbreviations for subsequent thermo-mechanical processes are: CR=Cold Rolled; FC=Furnace Cooled; FOR=Forged; HIP=Hot Isostatic Pressing; HPT=High Pressure Torsion; HR=Hot Rolled; VHP=Vacuum Hot Pressed; WQ=Water Quenched.

Few shot examples

1. Article: We have experimented on the chemical and mechanical high-entropy alloys CoCrFeMnNi. We have also done experiments with alloys after substituting elements,Ti for Co Mo or V for Cr. The factors affecting stability of various phases is studied.
Answer: CoCrFeMnNi, TiCrFeMnNi, CoMoFeMnNi, CoVFeMnNi

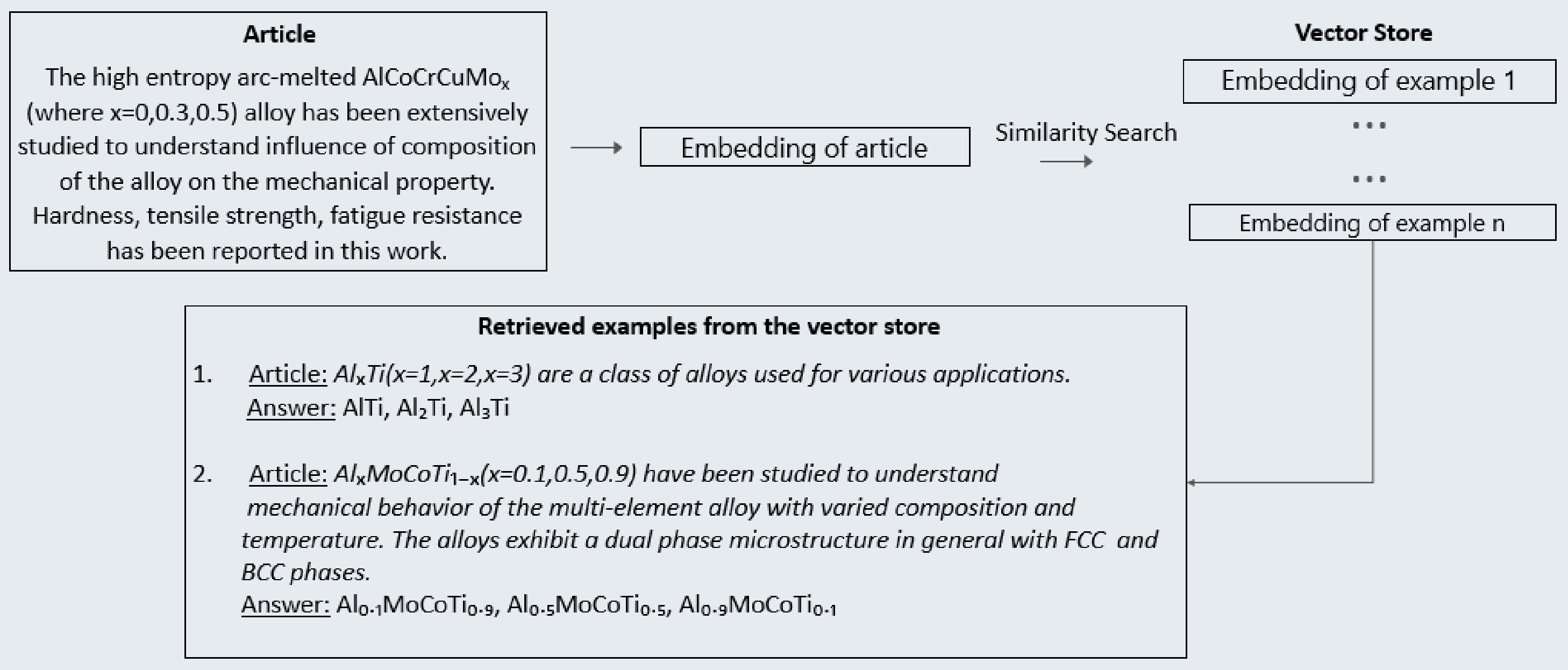
2. Article: AlCrFeCoNiCuTi alloy contains BCC1 phase, BCC2 phase and a FCC phase. AlCrFeCoNiCuV alloy contains two different phases with BCC, FCC structure, respectively.
Answer: AlCrFeCoNiCuTi, AlCrFeCoNiCuV

...

n. Article: $Al_xTi(x=1,x=2,x=3)$ are a class of alloys used for various applications.
Answer: AlTi, Al_2Ti, Al_3Ti

Few-shot prompt examples for text extraction

- **Retrieval-Augmented Generation:** RAG retrieves relevant domain-specific examples to augment the LLM prompt.



Article

The high entropy arc-melted $AlCoCrCuMo_x$ (where x=0,0.3,0.5) alloy has been extensively studied to understand influence of composition of the alloy on the mechanical property. Hardness, tensile strength, fatigue resistance has been reported in this work.

Retrieved examples from the vector store

1. Article: *$Al_xTi(x=1,x=2,x=3)$ are a class of alloys used for various applications.*
Answer: AlTi, Al_2Ti, Al_3Ti

2. Article: *$Al_xMoCoTi_{1-x}(x=0.1,0.5,0.9)$ have been studied to understand mechanical behavior of the multi-element alloy with varied composition and temperature. The alloys exhibit a dual phase microstructure in general with FCC and BCC phases.*
Answer: $Al_{0.1}MoCoTi_{0.9}$, $Al_{0.5}MoCoTi_{0.5}$, $Al_{0.9}MoCoTi_{0.1}$

RAG workflow: embedding → similarity search → retrieval

3. Table Extraction

Instructions

1. Identify properties from the context section.
2. <Formatting instructions>.

Context

1. Ultimate Tensile Strength (UTS):

Definition: Maximum stress a material withstands before breaking during a tensile test.
Units: Pascals (Pa), Megapascals (MPa), or Pounds per Square Inch (psi).
Variations: Referred to as tensile strength or breaking strength.

...

n. Hardness:

Definition: Material's resistance to deformation or penetration by an indenter.
Units: Specific test scales like Vickers (HV), Rockwell (HRC).
Variations: Includes microhardness and nano-hard

Few shot examples

Table Caption: "Table 4. Shear Modulus(GPa) measured at various temperatures"

Temperature (K)	FeNiCoCr	FeNi	Ni
77	-	68	84
203	-	66	80
293	84	62	76

Alloy	Processing condition	Testing condition	Property	Value	Unit
FeNiCoCr	-	293	Shear Modulus	84	GPa
FeNi	-	77	Shear Modulus	68	GPa
FeNi	-	203	Shear Modulus	66	GPa
FeNi	-	293	Shear Modulus	62	GPa
Ni	-	77	Shear Modulus	84	GPa
Ni	-	203	Shear Modulus	80	GPa
Ni	-	293	Shear Modulus	76	GPa

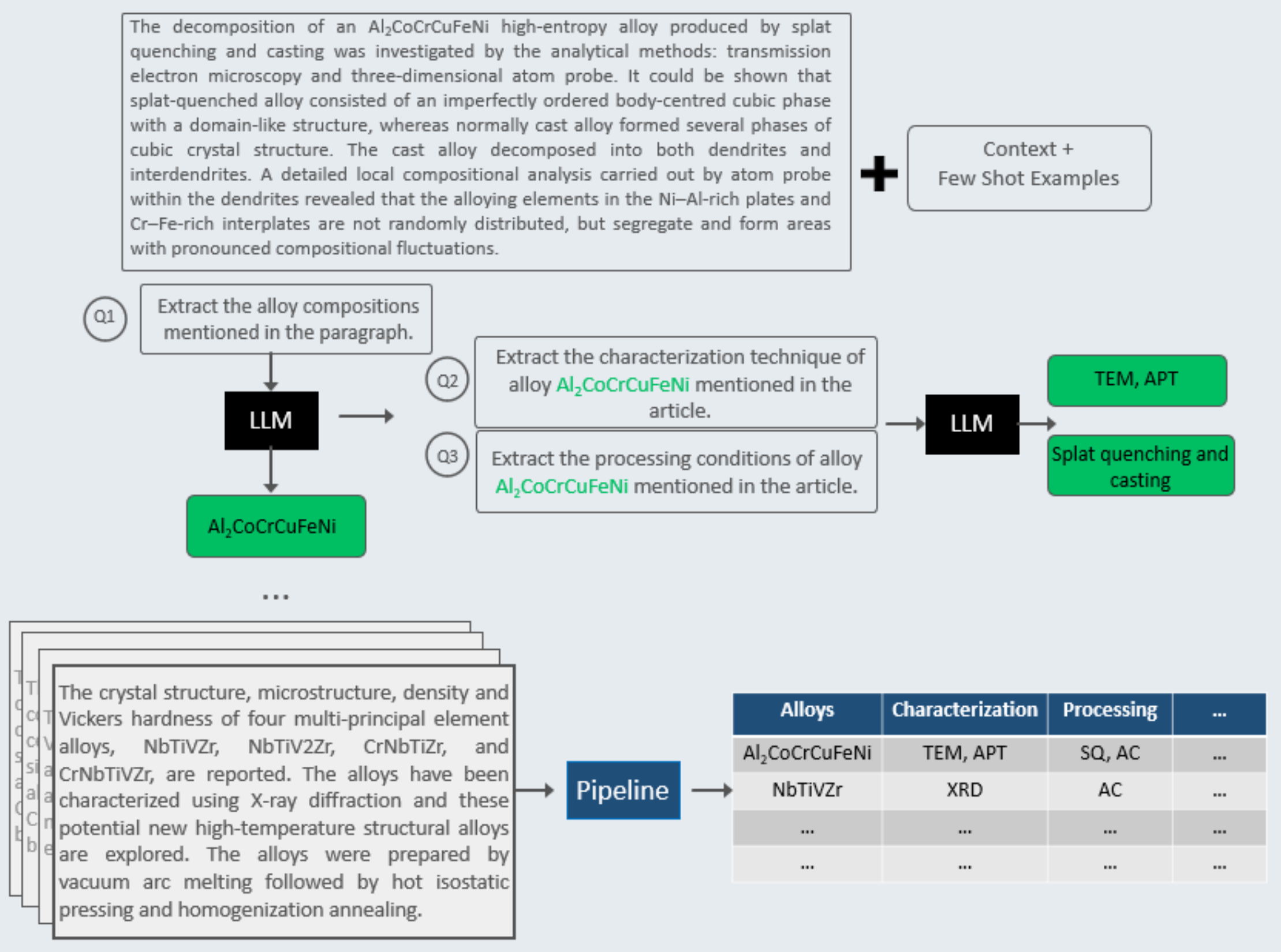
Table to extract from

Table Caption: "Table 4. Hardness of Al0.5CoCrCuFeNi alloys in four different states."

State	Hardness (HV)
As-cast	208 ± 2
As-forged	350 ± 3
As-homogenized-FC	265 ± 4
As-homogenized-WQ	154 ± 4

Table caption & cells → few-shot schema examples → CSV

Data Extraction Workflow


Context +
Few Shot Examples
Q1
Extract the alloy compositions mentioned in the paragraph.
LLM
$Al_2CoCrCuFeNi$
Q2
Extract the characterization technique of alloy $Al_2CoCrCuFeNi$ mentioned in the article.
Q3
Extract the processing conditions of alloy $Al_2CoCrCuFeNi$ mentioned in the article.
LLM
TEM, APT
Splat quenching and casting
...
Pipeline

The decomposition of an $Al_2CoCrCuFeNi$ high-entropy alloy produced by splat quenching and casting was investigated by the analytical methods: transmission electron microscopy and three-dimensional atom probe. It could be shown that splat-quenched alloy consisted of an imperfectly ordered body-centred cubic phase with a domain-like structure, whereas normally cast alloy formed several phases of cubic crystal structure. The cast alloy decomposed into both dendrites and interdendrites. A detailed local compositional analysis carried out by atom probe within the dendrites revealed that the alloying elements in the Ni–Al-rich plates and Cr–Fe-rich interplates are not randomly distributed, but segregate and form areas with pronounced compositional fluctuations.

The crystal structure, microstructure, density and Vickers hardness of four multi-principal element alloys, NbTiVZr, NbTiV2Zr, CrNbTiZr, and CrNbTiVZr, are reported. The alloys have been characterized using X-ray diffraction and these potential new high-temperature structural alloys are explored. The alloys were prepared by vacuum arc melting followed by hot isostatic pressing and homogenization annealing.

Alloys	Characterization	Processing	...
$Al_2CoCrCuFeNi$	TEM, APT	SQ, AC	...
NbTiVZr	XRD	AC	...
...	...	...	...
...	...	...	...

Results

Evaluation Metrics

$$\text{precision} = \frac{|\{\text{relevant entries}\} \cap \{\text{retrieved entries}\}|}{|\{\text{retrieved entries}\}|} \qquad \text{recall} = \frac{|\{\text{relevant entries}\} \cap \{\text{retrieved entries}\}|}{|\{\text{relevant entries}\}|} \qquad \text{F1-score} = \frac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}}$$

Query Set 1: Text Extraction

Evaluation on Review Articles:

Metric	Value
Precision	0.80
Recall	0.86
F1 Score	0.83

10K Papers Implementation:

- GPT-4o mini for standard processing
- Cost: $160 total ($0.015 per article)
- Processing time: 230 hours (90 sec/article)
- Successfully processed all 10,000+ papers

Query Set 2: Table Extraction

Evaluation on Review Articles:

Metric	Value
Precision	0.99
Recall	0.96
F1 Score	0.96

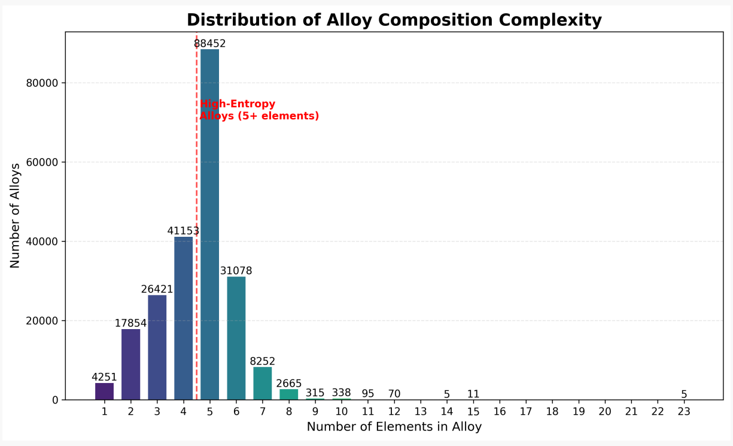
10K Papers Implementation:

- GPT-4o for complex table processing
- 5,294 articles with tables extracted
- Cost: $150.70 total
- Processing time: 49.3 hours

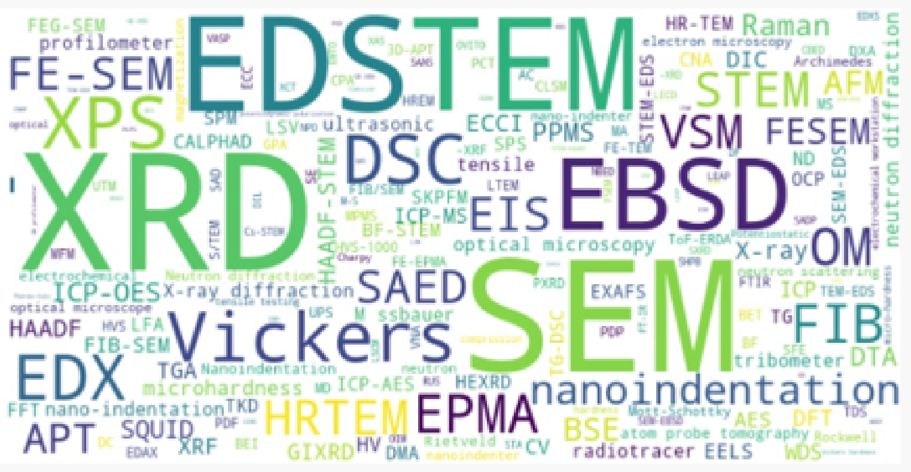
Database Statistics

- **10,829** processed journal articles
- **37,556** alloy systems extracted
- **32,846** (88%) directly usable alloy entries
- **15,998** unique alloy compositions
- **3** unique alloy compositions per article
- **2,202** alloy compositions discussed in multiple articles

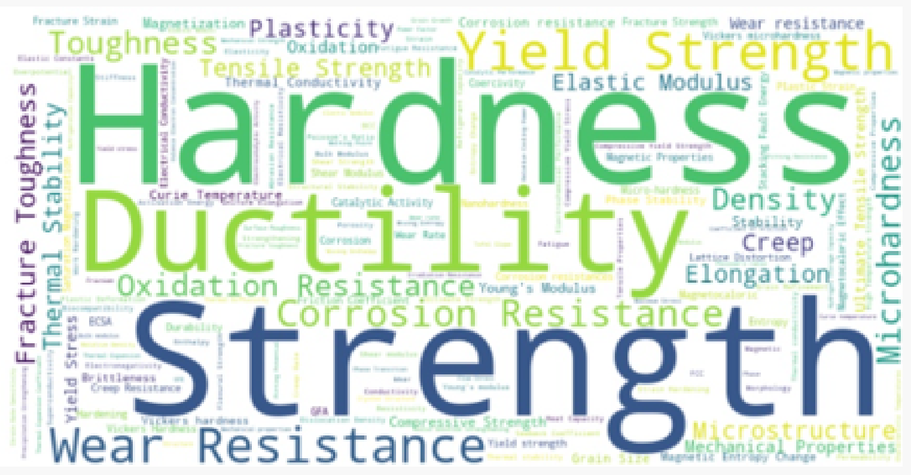
Alloy Composition Complexity Distribution



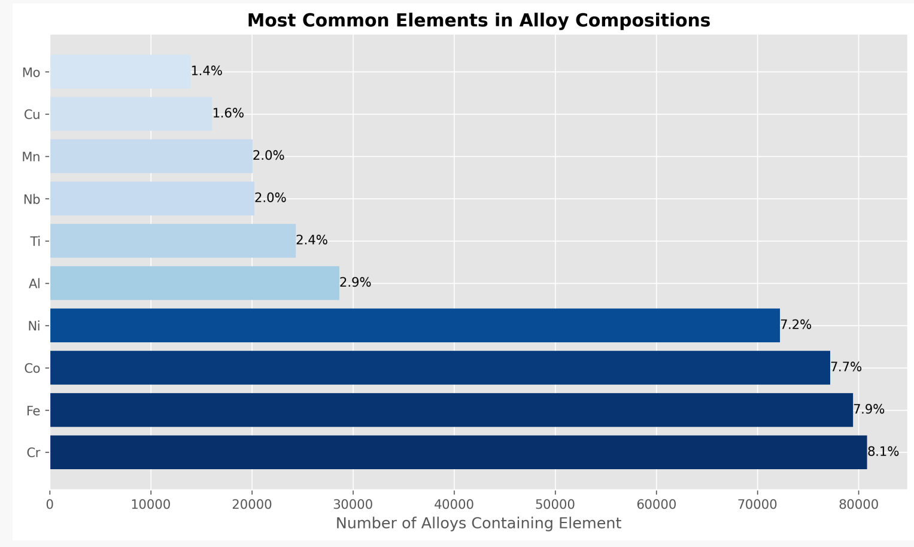
Word cloud of Processing Conditions



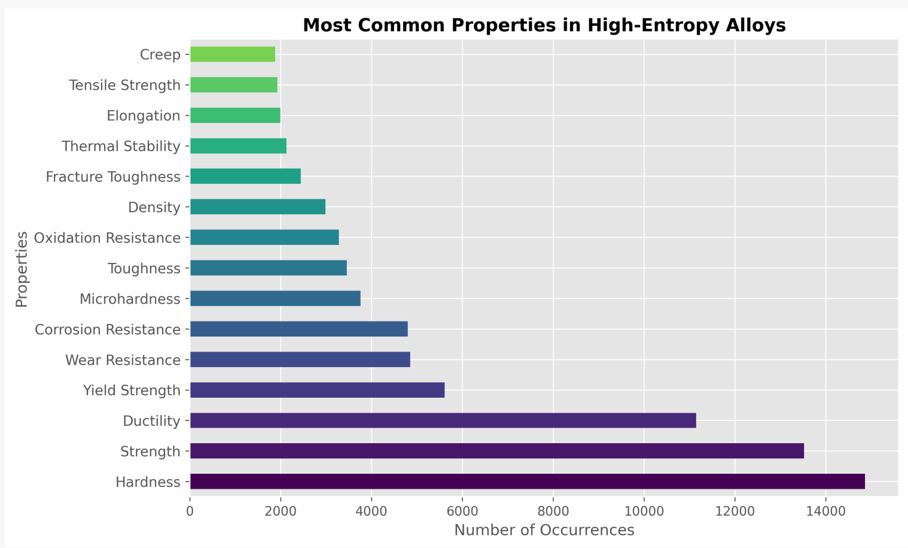
Word cloud of Properties Extracted



Most Common Elements



Most Common Properties



Sample Extracted Data

	Alloys	Characterization techniques	Processing conditions	Properties	Value	Units
1	NbMoTaW	X-Ray diffraction	Vacuum arc-melting	Yield Stress	1390	MPa
2	NbTiVZr	X-Ray diffraction	As-cast, Splat quenched	-	-	-
3	$Al_{0.5}CoCrCuFeNi$	Scanning electron microscopy, Electron dispersive spectroscopy	Arc-melting, water quenched and cold rolled	Hardness	208	HV

Acknowledgements

WSAI IITM

RBCDSAI IITM

da24m016@smail.iitm.ac.in