Integrating Materials and Manufacturing Innovation
https://doi.org/10.1007/s40192-024-00344-8

THEMATIC SECTION: HARNESSING THE POWER OF MATERIALS DATA

High-Throughput Extraction of Phase–Property Relationships from Literature Using Natural Language Processing and Large Language Models

Luca Montanelli[1] · Vineeth Venugopal[1] · Elsa A. Olivetti[1] · Marat I. Latypov[2]

Received: 29 November 2023 / Accepted: 25 January 2024

Abstract
Consolidating published research on aluminum alloys into insights about microstructure–property relationships can simplify and reduce the costs involved in alloy design. One critical design consideration for many heat-treatable alloys deriving superior properties from precipitation are phases as key microstructure constituents because they can have a decisive impact on the engineering properties of alloys. Here, we present a computational framework for high-throughput extraction of phases and their impact on properties from scientific papers. Our framework includes transformer-based and large language models to identify sentences with phase-property information in papers, recognize phase and property entities, and extract phase-property relationships and their "sentiment." We demonstrate the application of our framework on aluminum alloys, for which we build a database of 7,675 phase–property relationships extracted from a corpus of almost 5000 full-text papers. We comment on the extracted relationships based on common metallurgical knowledge.

Keywords Natural language processing · Large language models · Aluminum alloys · Phase–property relationship

Introduction

Global demand for metals is expected to increase two to sixfold by 2100 [1–3]. This is especially true for aluminum as there is a growing demand for high-performance, lightweight, recyclable structural alloys across industries [4, 5]. In the context of recycling, shifts in end-uses lead to uncertainties in the future scrap stream compositions, which is further exacerbated by accumulation of detrimental elements as alloys are recycled [6]. For example, increasing recycling content often leads to the emergence of phases (e.g., iron-containing intermetallics) detrimental to mechanical and other properties [7, 8].

Understanding microstructure–property relationships is the foundation for any alloy design effort (including recyclability considerations). Microstructure constituents of special interest in aluminum alloys are *phases*—spatial regions of uniform crystal structure and chemistry. Many beneficial properties are achieved based on the formation of desirable phases, for example, in the form of fine precipitates [9]. Conversely, many performance characteristics sharply deteriorate in the presence of phases with undesirable size or morphology. Given the importance of phases as key microstructure constituents, a large body of work has been dedicated to experimental observation of phase formation in response to metallurgical processing. Systematically organizing the knowledge published in the literature over decades of research could greatly benefit the current alloy design endeavors.

In recent years, natural language processing (NLP) has emerged as a powerful tool for analysis of large sets of scientific texts. It has been applied to the design and discovery of battery materials [10], complex oxides [11], zeolites [12, 13],

✉ Luca Montanelli
montanel@mit.edu

Vineeth Venugopal
vineethv@mit.edu

Elsa A. Olivetti
elsao@mit.edu

Marat I. Latypov
latmarat@arizona.edu

[1] Department of Material Science and Engineering, Massachusetts Institute of Technology, Cambridge 02139, MA, USA

[2] Department of Material Science and Engineering, University of Arizona, Tucson 85721, AZ, USA

Published online: 19 March 2024

nanoparticles [14], and more [15, 16]. However, development and application of NLP to the design of structural alloys are still in early stages. Sample research includes the text-mining of millions of papers to efficiently design high-entropy alloys [17] as well as predicting the pitting potential for corrosion-resistant alloy design using embeddings of literature excerpts [18]. Relevant to aluminum, Liu et al. have created a labeled dataset of material entities from the literature focused on the Al-Si alloy system [19]. Their use of active learning to supplement their manual labeling of entities, however, highlights the need for an automated high-throughput extraction method applicable to different regions of the alloy space. On the other hand, Pfeiffer et al. considered the entire range of aluminum alloy series and extracted 14,884 aluminum alloy compositions, along with 1,278 properties from 5,172 research papers [20]. While covering wide independent ranges and distributions of engineering properties, their database does not contain links between compositions and properties.

To address this gap, we develop an NLP framework to automatically extract, from the literature, phases, and their "sentiment" in terms of positive or negative impact on properties. We leverage large language models (LLMs) to perform a wide variety of NLP tasks (including named entity recognition (NER) and relationship extraction (RE)), without the need for extensive manually labeled datasets [21]. By performing automated collection of relevant sentences, NER, and relationship inference using transformer-based models and LLMs, we create a database of existing phase–property relationships. We demonstrate the uses of this database for gaining insights that we confirm against established metallurgical knowledge. We focus on aluminum alloys, but the framework presented here is flexible and can be applied to other metallic systems. We develop the framework in Sect. 2 and show how it can derive key insights from the aluminum system in Sect. 3. The framework's uses and implications for researchers will be discussed in Sect. 4.

NLP Framework for High-Throughput Extraction of Phase–Property Relationships

In this section, we present an NLP framework for extraction of phase–property relationships in alloys from the literature applied to aluminum alloys. We (i) collect a corpus of relevant papers, (ii) extract sentences from full-body papers, (iii) perform NER and extract phase–property relationships, (iv) aggregate or disambiguate the extracted entities (Fig. 1).

Paper Corpus Collection

We first build a corpus of papers related to aluminum alloys culled from our in-house database of more than 5.7 million full texts of papers published in academic journals [22]. Our search for relevant papers included two strategies: (i) rule-based regular expression (regex) matching of words in titles and abstracts of the in-house database and (ii) querying the Scopus database [23]. In the first search strategy, we used the following five rules, which checked the presence of:

- The words alumin(i)um and alloy in the title,
- Alloy denominations in the title, (ex: "Al6061"),
- Alloy series in the title (ex: "7xxx", "6xxx"),
- Alloy names using chemical elements (ex: "Al-Si", "Al-Mg-Sc-Zr") along with the word alloy in the title,
- Alloy numbers consisting of 3 or 4 consecutive numbers (ex: "5182", "A382") with a mention of alumin(i)um in the title or abstract.

A paper satisfying any of those rules was considered an aluminum text. We found a total of 19,356 articles in our database of full texts. To complement this search, we also queried the Scopus database for papers on the subject of aluminum alloys. We queried papers that contained strings "alumin*um" or "Al-" and "alloy" in the titles but excluded those having "-Al" to remove papers with aluminum as an alloying element. The Scopus queries provided a further 1,164 articles that that were not already present in the list of relevant papers identified with the regex search. Having the combined list of articles on aluminum alloys, we downloaded full texts of these articles from our in-house database to obtain a final corpus of 20,520 full texts in the JSON format.

Sentence Dataset Collection

From the paper corpus, we then extract the sentences that contain information on phases and properties to build a sentence dataset. We choose the sentence as the main unit of text because papers in metallurgic literature often discuss multiple phases and properties in a single paper. Focusing on a smaller unit of text reduces the possibilities of ambiguous relationships. On the other hand considering larger units of text (e.g., paragraph) may challenge extraction of unambiguously coupled phase–property pairs and the sentiment of their relationships. Furthermore, we hypothesize that the description of phases and their impact on properties is captured at the

Fig. 1 Summary of our NLP framework

sentence level in the metallurgical literature at a sufficient level for insights to be gathered. Finally, focus on a small unit of text enables use of a wide spectrum of NLP tools and LLMs, including those with limited context windows.

The prototypical sentence that we targeted to include in the sentence dataset reads as *"[Phase A] leads to an increase in [property B]"*. Such sentence extraction can be approached as a classification problem, i.e., whether or not a given sentence contains the phase-property information, or whether or not it resembles our prototypical sentence. Here, we chose BERT-type transformer models coupled with a classification head to perform this task. For best performance, we fine-tuned and evaluated four BERT models: the uncased versions of the original BERT [24], SciBERT trained from BERT using 1.14M papers [25], MatBERT trained from BERT using 2 M papers [26], and MatSciBERT trained from SciBERT using 150k papers [27]. For fine-tuning and evaluation of the BERT models, we prepared 2000 hand labeled examples of sentences containing the phase-property information. Our fine-tuning also included hyperparameter optimization using cross-validation. The cross-validation scores were used as the optimization metric and their variance served to evaluate the spread of performance due to different initialization of parameters. The number of epochs, the batch size, and the learning rate were the hyperparameters tuned on a training and validation set while a test set was set aside for final comparison of the models.

Table 1 shows F1 scores for each model on the test set after hyperparameter tuning for BERT models. MatBERT and MatSciBERT exhibit comparable highest scores on the testing set (hidden from models during training and hyperparameter tuning). While MatBERT and MatSciBERT performed similarly, we selected MatBERT as the model of choice thanks to its higher test-set performance. The final MatBERT model was trained on the combined training and validation sets and performed with an F1 score of 78.87%, a precision of 74.67%, and a recall of 83.58% on the test set. The higher recall indicates that few undesired sentences enter our dataset at the cost of missing some desired sentences. However, this helps the downstream tasks by minimizing extraction and clean-up of undesired sentences.

Table 1 F1 scores of the 4 BERT models tested for sentence extraction

Model type	Test F1 (with 95% confidence)
BERT	69.06% (± 00.77)
SciBERT	73.94% (± 00.46)
MatBERT	**77.20% (± 00.43)**
MatSciBERT	77.18% (± 00.31)

The 95% confidence interval was measured using repeated evaluations to capture the variations in initializations

We used the selected model, MatBERT, on every sentence in our paper corpus to filter out all relevant sentences. Our resulting sentence dataset numbered a total of 10,213 unique sentences from 7,763 papers that were deemed to contain a phase, a property, and a relationship showing how the phase affects the property in question. We observed that most of the sentences with phase-property information were from the Introduction, Results, and/or Discussion sections of papers, which highlights in the importance of working with full texts rather than abstracts of papers.

Named Entity Recognition and Relationship Extraction

From the sentence dataset, we extract the entities (phase and property) along with the relationship between the two. In the prototypical example sentence: *"[Phase A] leads to an increase in [property B]*," we are looking to (i) recognize the phase entity "phase A" and property entity "property B," and (ii) extract their qualitative relationship ("increase," i.e., positive). To accelerate the implementation of the NER and RE tasks without the need for excessive annotated data, we leveraged a few-shot learning approach with LLMs. LLMs also allow completing both NER and RE tasks simultaneously.

Here we used the Cohere xlarge model with 52B parameters along with few-shot learning for simultaneous completion of the NER and RE tasks [28]. Cohere is commercial LLM platform and their xlarge model performance is comparable to the much larger models GPT-3, Jumbo, or BLOOM following the HELM protocol [29]. For few-shot learning, we provide a short list of examples containing the sentence, phase, property, and relationship as an instruction to the model of what information to extract (Fig. 2). The Cohere generation endpoint was prompted with the same set of instructions for each of the 10,213 sentences in the sentence dataset.

To facilitate relationship extraction, we have defined two possible relationships: positive and negative. A positive relationship is when a phase leads to an increase in a property. A negative relationship is the opposite. We use qualitative "positive" and "negative" relationships in the literal sense without engineering implications. For example, in the sentence: *"Phase A leads to an increase in corrosion*," the relationship is defined as "positive" even though an increase in corrosion is negative in the broader context of engineering alloy design. We limit the task to literal RE to minimize the ambiguity for LLMs. Indeed, it may require advanced reasoning abilities together with domain knowledge from LLMs to understand the subtle differences in such statements as "increase in corrosion resistance" versus "reduction in corrosion susceptibility." At the same time, presented with literal relationship attached to a particular property from an LLM, a human user can readily determine that a literally

...
Sentence: It can be found that the precipitation strengthening of the Al3(Sc, Zr) particles is the most remarkable.
Phase: Al3(Sc,Zr)
Property: precipitation strengthening
Relationship: positive
--
Sentence: The results show that the addition of Sc and Y2O3 nano-particles could significantly improve the mechanical property of the Al-Si alloy.
Phase: Sc && Y2O3
Property: mechanical property
Relationship: positive
--
Sentence: This characteristic of θ' phase contributes to better mechanical performance of 2A14 aluminum alloy after the NIA process.
Phase: *θ' phase*
Property: *mechanical performance*
Relationship: *positive*

Fig. 2 Example prompt used for entity and relationship extraction. The non-italicized parts represent the prompt given to the model and the blue italicized parts are an example of output to be parsed

positive relationship with an undesirable property means a negative impact of "phase A" on the overall engineering performance of the alloy. Relationships between phases and properties discussed in literature go well beyond the binary labels we have chosen. While focusing on extracting binary relationships from individual sentences, our approach allows us to go beyond overall binary relationships by analyzing the frequency of positive/negative samples for a given phase-property pair. Therefore, the nuance of the relationship for each phase-property pair arises from the relative frequency of positive versus negative extracted samples.

The performance of the LLM was evaluated using a manually labeled set of 50 sentences and ROGUE-L score, which quantifies similarity between two strings based on the length of their longest common subsequence [30]. We found that the ROUGE-L score of the LLM for phase recognition was 84.08%, while that for the property recognition was 75.86%. Those suggest that the model performs very well at the NER task. These results were obtained using the final prompt where a total of 23 example sentences were provided to the model. Figure 2 shows three examples, while a full set of the sentences can be found in the GitHub repository (see link in Code Availability section). Furthermore, the LLM showed an F1 score of 89.47% on the relationship extraction task. Additionally, we looked at the same metrics for a zero-shot version of the same LLM, and while the F1 score remained high at 88.23%, the recognition scores for phases and properties were much lower at 28.47% and 46.00%, respectively.

Some sentences in our dataset mentioned multiple phases and/or properties. For example, for a sentence containing *"phase A, phase B, and phase C increase property X*," we aim to extract three individual relationships between the three phases and the property. To resolve such cases, we used a convention of separating the entities with the symbols " &&" for parsing individual relationships from sentence with multiple entities. This convention was shown to the LLM with some examples in the prompt (see Fig. 2 second example). This enabled us to better parse the output after generation.

Entity Aggregation

Our approach with this work consisted in extracting all the insights from literature as opposed to querying the sentence dataset with a pre-defined list of phase and properties. An entity aggregation step is therefore needed in our framework to consolidate the different extracted entities into a smaller list of phases and properties. At the end of the NER and RE tasks, our database contained 5,671 unique phases and 1,769 unique properties. These represent unrealistically high numbers even for all aluminum alloy series considered here. Upon inspection, we found two sources of excessive numbers of extracted entities: (i) imperfect entity extraction (false positives) and (ii) numerous alias representations of the same phases or properties. False-positive entities typically included metallurgical terms that represent neither phases nor properties yet were extracted as those by the model. After inspection of the most frequent false positives, we complied lists of "stop-words" to remove entities irrelevant to phase–property relationships. The second problem—non-unique representations—emerges due to a variety of ways that authors refer to the same phases or properties. For example, "al2cu phase," "th (al2cu)," "cual2," "θ-phase" are examples of distinct entities that all represent the same physical phase, "$\theta - Al_2Cu$". To a domain expert, these entities evidently represent the same phase, yet they are much

harder for a machine to recognize as aliases. We thus aimed to consolidate aliases in order to not miss datapoints when capturing the relationships between phase-property pairs. The next two sections describe our strategies of removing false positives and merging aliases for phases and properties given the specific characteristics of their text descriptions. Those strategies, while developed on the aluminum system, are applicable to other systems with minimal modification.

Phases

Aggregation and cleaning of phase entities was performed using a series of filters to remove, alter towards a unified format, and merge entities. The first step of phase name aggregation removed stop-words and modified the entities towards a unified format of phase and property representations. We distinguished two types of stop-words. The first are hard stop-words whose presence anywhere in the entity string indicated a false positive and thus led to deletion of the sentence from the database. Examples of those words are "alloy," "fibers," or "nanotube." If a sample contained the entity "Al-Mg alloy," the phase entity was removed entirely from the database along with its associated sentence, property, and relationship. The second type are soft stop-words that were simply removed from the entity while the rest was preserved. Examples of soft stop words include "precipitate," "intermetallic," or "particle:" the entity "Mg2Si precipitate" would change to "Mg2Si". Stop words of both types were identified by manual inspection as words whose presence indicated that the entity was not a phase. Other modifications included the proper handling of Greek letters (changing "eta" to "η" or "th" to "θ", where "eta" and "th" are the ways the paper database handled Greek letters in some of the full texts), the change to the same types of apostrophes (i.e., removing Unicode characters U+2032 and U+2033), and conversion of chemical elements' full names to symbols.

The second step included merging entities with similar chemical formulae. A list of formulae was made manually to obtain the most common and unified version of the same phase names. For example, such phases entities as "mg2al3" and "al3mg2" were merged together. Apostrophe symbols contained in entities (denoting metastable phases) were kept to distinguish stable phases from their metastable counterparts.

The third step was a more targeted series of similar steps. First of all, strings containing only element names were removed to focus on phase–property relationships rather than the effects of individual alloying elements. In this context, treating entities containing "Si" deserved special attention. In addition to "Si" as an element, "Si" also commonly refers to silicon as a phase. To isolate "Si" as a phase, we used the presence of the words "eutectic," "phase," or "primary" in the same sentence, which we found to be uniquely co-occurring with the silicon phase (but not Si element). Finally, we created a dictionary to map codified phase names into our selected unified notation. For example, changing "T1" and "Al_2CuLi" into "T1-Al_2CuLi" or "$Al_3Sc_x Zr_{1-x}$" and "$Al_3Sc_{1-x}Zr_x$" into "Al_3(Sc,Zr)."

The fourth step was disambiguation, which, for our database, mostly focused on the "β" phase. In aluminum literature, "β" can refer to multiple distinct phases: Mg_2Si, Al_5FeSi, or Al_3Mg_2. We disambiguated β phase entities by looking at chemical names appearing in the same sentence or the paragraph containing the sentence. For example, if exactly one the chemical formulae (Mg_2Si, Al_5FeSi, or Al_3Mg_2) appeared in the sentence or the paragraph then this was used to determine the specific β phase. Therefore, "β-Mg_2Si" represents entities specifically disambiguated as Mg_2Si. In all other cases, "β" was left to denote an aggregate of the phases that were not disambiguated even after look-up of the additional information at the sentence and paragraph levels.

Properties

Of the 1,769 unique property entities, 61.3% datapoints represent property terms encountered only once in the corpus. This observation together with manual inspection of the extracted entities suggests that the vast majority of the 1,769 entities need to be grouped into a much smaller number of actual property categories. From the data perspective, this task becomes one of clustering to find which sets of entities refer to the same property and thus must be grouped. In the case of properties, more semantic information is typically available at the string level to warrant the use of more powerful NLP methods of aggregation that require less manual intervention than phases. Here, we approached property aggregation as topic modeling. We adopted the BERTopic model [31] for the task and evaluated a variety of configurations to obtain most complete and accurate alias property aggregation. We settled on using a seeded version of the model, which uses a list of topics and example words for each topic to guide the model in matching those as closely as possible. We used the top 20 properties in the raw list of properties as the 20 seed topics. The model then found topic by clustering the embeddings of the raw property strings and created topic labels. Labels were further refined by manual inspection. Furthermore, two additional changes were made to the raw topics obtained by the model. First, we split such extracted properties as "wear" and "corrosion" properties by the presence of the words "resistance," "resistant," or "protection" to create the new "corrosion resistance" and "wear resistance" properties. For the second change, we subclassified every aggregate/collective entity such as "strength"

or "tensile properties" into a more specific property, e.g., "UTS" if the entity contained the word "ultimate."

Insights from the Application on Aluminum Alloys

This work resulted in the extraction of 7,675 sentences from 4,891 full-text papers with recognition of 2,955 unique phases and 25 unique properties, which constituted our phase-property database.

Figure 3 shows the histograms of the top most frequent phases and properties in our database. From Fig. 3a, we see that the database is dominated by the Si phase that occurs in Al-Si alloys, notably cast alloys [9]. Following closely is the β-Mg_2Si phase commonly present in 6xxx alloys as the main strengthening phase [32, 33]. We also observe a large number of reinforcement phases/compounds amongst which are SiC, TiB_2, B_4C, Al_3Ti, TiC, and Al_2O_3 [34–36]. We found that these phases/compounds are often associated with wear resistance and thermal stability.

Figure 3b demonstrates a domination of the property "strength" in our database, as it constitutes a quarter of all property occurrences. The high frequency of the strength property is associated with two factors. First, many sentences contain the general term "strength" when referring to the more specific properties such as yield strength, ultimate tensile strength, specific strength, etc. Second, we find that the strength characteristics are much more extensively studied and reported in the literature on aluminum alloys compared to other properties.

We further analyze the phase–property relationships obtained with our NLP framework. Figure 4 shows a heatmap of phase and property pairs mentioned in the sentences for the top 10 phases and properties. First of all, Fig. 4 shows the frequency of co-occurrence of phases and properties. For example, β-Mg_2Si appears frequently with the hardening property, and its metastable β" variant also appears with hardness. This aligns with metallurgical intuition as these phases along with β' form during the aging process and typically result in increased hardness and hardening ability of aluminum alloys [37, 38]. The Si phase is frequently mentioned with mechanical properties as this is a topic of interest for Al-Si alloys that are often used for structural components in the automotive and aerospace industries [39, 40]. Similarly, its co-occurrence with brittleness is expected

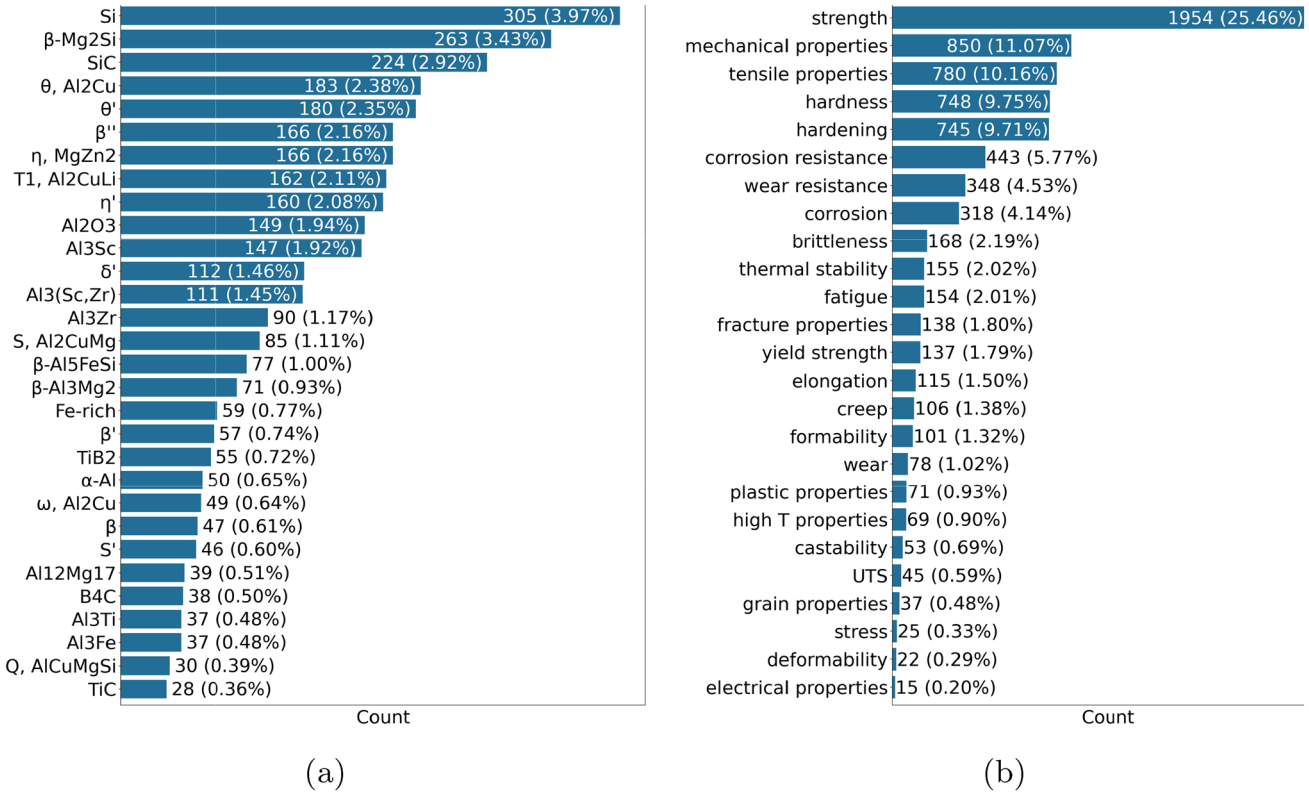

Fig. 3 **a** Frequency plot of the top 30 phases found in our database. Names with commas indicate that we aggregated the chemical and alias names. In the case of the β phase, we disambiguated names such that "β-Mg_2Si" only contains Mg_2Si and β is an aggregate of Mg_2Si, Al_5FeSi, and Al_3Mg_2 that we could not separate. The metastable versions of the β phase (β' and β'') have been checked and are only composed of metastable version of Mg_2Si. **b** Frequency plot of the 25 properties along with absolute numbers and percentage occurrence

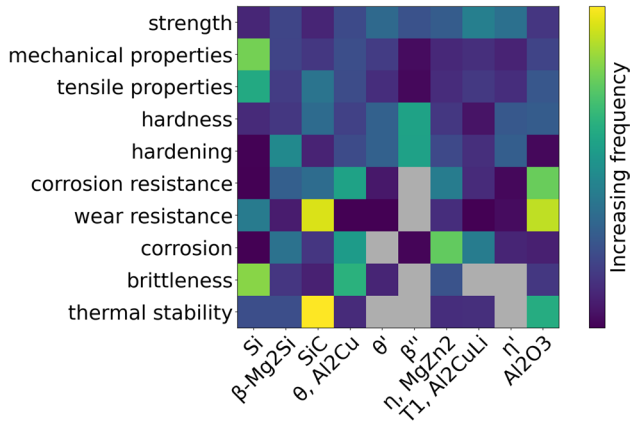

Fig. 4 Heat map representing frequencies of pairs for the top 10 phases and top 10 properties. Yellow indicates higher frequencies, while dark blue are lower frequencies. Gray squares show the absence of the pair in our database. Each row and column is normalized by the total mentions of the corresponding property or phase, respectively, in order to highlight trends. Without this step, the upper left corner would be mostly yellow giving us no added information

as coarse Si particles are mostly detrimental to the ductility of those alloys [41, 42]. Furthermore, we see that SiC and Al_2O_3 are mentioned with wear resistance and thermal stability more than the average as these compounds typically improve both properties [43–45].

We analyze the qualitative relationship for a few phase-property pairs. Figure 5 shows the positive (green) and negative (red) sentiment of the top phases in terms of their impact on a selected property indicated in the title. The dashed lines provide a point of comparison showing the average positive sentiment for all mentions of that property. We see that the metastable phases θ' and η' have more positive correlations with strength compared to the other non-metastable phases. The non-metastable phases appear in a richer variety of contexts including sentences on over-aging as well as different phase morphologies or sizes which can deteriorate strength, which adds to the corresponding "negative" sentiment fraction in 5a. Corrosion resistance (5b) shows two clear extremes. Al_2O_3 is predominantly regarded as beneficial for corrosion resistance mainly due to its protective properties when forming a layer on the surface of alloys [46, 47]. At the same time, the θ-Al_2Cu phase appears detrimental to corrosion resistance, which is due to the formation of galvanic cells between the noble Al_2Cu and the Al matrix [48, 49]. Similarly the η-$MgZn_2$ phase is also negatively correlated to corrosion resistance. In Fig. 4, we also see a strong correlation between corrosion and with the η-$MgZn_2$. Investigating the literature shows a similar sentiment as for the θ phase, the η phase is often reported as being the cause of stress corrosion cracking and intergranular corrosion [50, 51].

Discussion

The exploration of the literature on aluminum alloys shown in Sect. 3 highlights the retrieval of specialized knowledge from research papers. We were able to not only see which phase-property pairs were most mentioned, but also observe the positive or negative impact of phases. The insights from our database and its visualization are in concordance with metallurgical intuition and domain knowledge. While demonstrated on aluminum alloys, our framework can be used for other families of alloys where phases are important, including new emerging alloys with less established phase–property relationships. Multiprincipal element alloys is an example of potential application, especially given an explosion of the number of papers being published on the topic in recent years. Informing design of sustainable aluminum alloys with

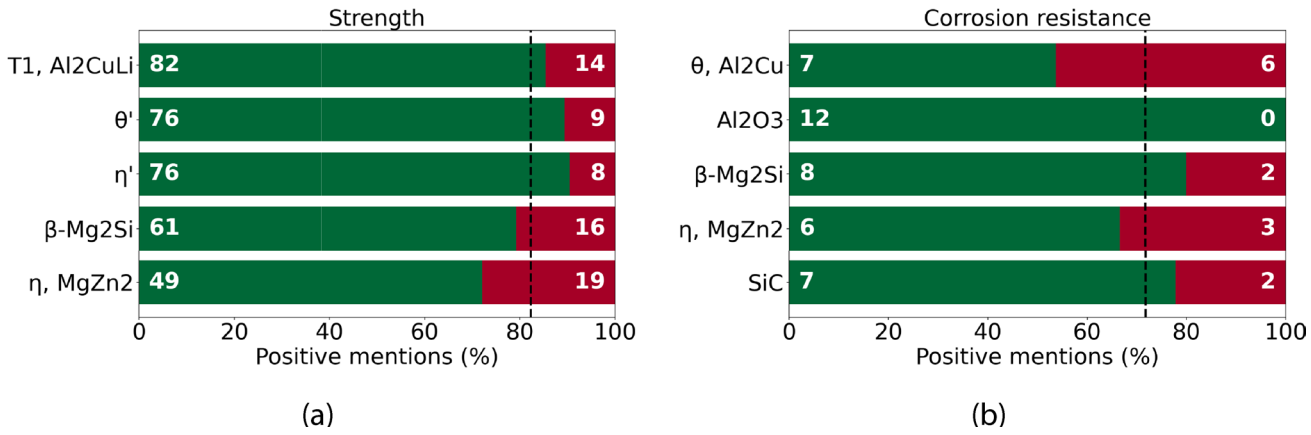

Fig. 5 Positive and negative mentions for the top phases associated with the properties **a** strength, and **b** corrosion resistance. The colored bars represent fractions of positive (green) and negative (red) mentions. The dashed line represents average positive mentions for that property and the numbers on either side of the plot represent the number of mentions

increased recycling rates, where multiple harmful phases can emerge is another potential high-impact application. Rapid inspection of large body of published works can therefore contribute to accelerated alloy designs in these highly active research directions.

In addition to helping researchers rapidly navigate metallurgical literature, our NLP framework can streamline data extraction for training quantitative machine learning models (e.g., for microstructure–property relationships). Extracted sentence datasets can also serve as a basis for constructing materials knowledge graphs [52], which in turn can form a foundation for interactive systems of fast and user-friendly retrieval of materials information. Our sentence dataset can be utilized as information-dense source of text data that can be used as domain-specific context for conversational LLMs (e.g., for retrieval-augmented generation [53]).

We finally note the key role of LLMs in building our framework without the need in excessive amounts of manually labeled data. Specifically, we observed a remarkable performance of LLMs in NER and RE tasks using only a handful of labeled examples (Sect. 2.3). The manual annotation of sentences for NER and RE tasks with more traditional NLP approaches would have been extremely time consuming. Furthermore, using few-shot learning, we could significantly improve the model performance without expensive fine-tuning. The sentence classification was addressed by fine-tuning BERT-type models, for which constituting a manually annotated dataset requires significantly less effort than NER and RE tasks.

Limitations and Future Opportunities

In this work, we developed a framework for high-throughput extraction of phases, properties, and their relationships from published literature on aluminum alloys. Ideally the framework should be fully automated, however, in the current state, some (semi-)manual intervention was still needed, most notably the aggregation of alias notations of the same phases/properties and their verification. For example, 15% of the extracted samples of the property "strength" have been aliased from otherwise worded terms referring to strength. Similarly, our database contains the property "corrosion," which aggregates not only the term "corrosion" itself but also other related terms that constituted 79% of the final aggregated "corrosion" samples. We expect that rapid progress in NLP and LLMs will eliminate the need in these additional steps and allow extraction of one-to-one relationships of unique phases and properties.

This study focused on *qualitative* relationships between phases and properties, i.e., whether any given phase has a positive or negative impact on a property. Next efforts in this direction can pursue *quantitative* relationships as well as additional extraction of alloy chemical composition to further aid computational alloy design.

Finally, Fig. 5 shows that relationships described in literature are about 70–80% positive. This indicates a clear bias towards reporting "positive results," e.g., phases and phenomena that are beneficial to alloys properties. This bias results in unbalanced extracted datasets regardless how good the NLP framework for extraction is. The computational alloy design leveraging state-of-the-art NLP could benefit from a more balanced reporting of both negative and positive research results from the community.

Conclusion

In summary, we present a novel methodology for extracting phase–property relationships from metallurgic literature using natural language processing and large language models. The study focuses on the aluminum system and leverages the power of NLP and LLMs to systematically organize knowledge from a vast corpus of research papers. The insights generated from the extracted database show its use as a valuable guide for alloy designers and researchers seeking to optimize alloy performance.

The results presented here show that this framework is useful for rapidly extracting insights from literature on alloys. The knowledge we have derived on the aluminum system would, traditionally, be held in textbooks that would have taken years to write by experts. As research on alloy properties continues to grow, these tools will become an indispensable to quickly screen literature and gain insights.

Acknowledgements The authors gratefully acknowledge the support from Novelis and NSF (grant CBET-2243914). We express our gratitude to Mrigi Munjal and Thorben Prein for providing a source of inspiration for the approach we used as well as important code snippets.

Funding 'Open Access funding provided by the MIT Libraries'.

Code Availability The code and database are openly available on GitHub at the following address: https://github.com/olivettigroup/phase-sentiment.

Declarations

Conflict of interest On behalf of all authors, the corresponding author states that there is no conflict of interest.

References

1. Saevarsdottir G, Kvande H, Welch B (2019) Aluminum production in the times of climate change: the global challenge to reduce the carbon footprint and prevent carbon leakage. JOM 11:72. https://doi.org/10.1007/s11837-019-03918-6
2. Cullen JM, Allwood JM (2013) Mapping the global flow of aluminum: from liquid aluminum to end-use goods. Environ Sci Technol 47(7):3057–3064. https://doi.org/10.1021/es304256s
3. Watari T, Nansai K, Nakajima K (2021) Major metals demand, supply, and environmental impacts to 2100: a critical review. Resour Conserv Recycl 164:105107. https://doi.org/10.1016/j.resconrec.2020.105107
4. Raabe D, Ponge D, Uggowitzer PJ, Roscher M, Paolantonio M, Liu C et al (2022) Making sustainable aluminum by recycling scrap: the science of "dirty'' alloys. Prog Mater Sci 128:100947. https://doi.org/10.1016/j.pmatsci.2022.100947
5. Raabe D, Tasan C, Olivetti E (2019) Strategies for improving the sustainability of structural metals. Nature 11(575):64–74. https://doi.org/10.1038/s41586-019-1702-5
6. Gaustad G, Olivetti E, Kirchain R (2011) Toward sustainable material usage: evaluating the importance of market motivated agency in modeling material flows. Environ Sci Technol 45(9):4110–4117. https://doi.org/10.1021/es103508u
7. Yang H, Ji S, Fan Z (2015) Effect of heat treatment and Fe content on the microstructure and mechanical properties of die-cast Al–Si–Cu alloys. Mater Des 85:823–832. https://doi.org/10.1016/j.matdes.2015.07.074
8. Basak C, Hari Babu N (2017) Improved recyclability of cast al-alloys by engineering β-Al9Fe2Si2 phase. In: Light metals. Springer, pp 1139–1147
9. Wang J (2018) Physical metallurgy of aluminum alloys. In: Aluminum science and technology. ASM International. https://doi.org/10.31399/asm.hb.v02a.a0006503
10. Huang S, Cole J (2020) A database of battery materials auto-generated using ChemDataExtractor. Sci Data 08:7. https://doi.org/10.1038/s41597-020-00602-2
11. Young SR, Maksov A, Ziatdinov M, Cao Y, Burch M, Balachandran J et al (2018) Data mining for better material synthesis: the case of pulsed laser deposition of complex oxides. J Appl Phys 123(11):115303. https://doi.org/10.1063/1.5009942
12. Schwalbe-Koda D, Kwon S, Paris C, Bello-Jurado E, Jensen Z, Olivetti E et al (2021) A priori control of zeolite phase competition and intergrowth with high-throughput simulations. Science 374(6565):308–315. https://doi.org/10.1126/science.abh3350
13. Jensen Z, Kwon S, Schwalbe-Koda D, Paris C, Gómez-Bombarelli R, Román-Leshkov Y et al (2021) Discovering relationships between OSDAs and zeolites through data mining and generative neural networks. ACS Cent Sci 7(5):858–867. https://doi.org/10.1021/acscentsci.1c00024
14. Cruse K, Trewartha A, Lee S, Wang Z, Huo H, He T et al (2022) Text-mined dataset of gold nanoparticle synthesis procedures, morphologies, and size entities. Sci Data 05(9):234. https://doi.org/10.1038/s41597-022-01321-6
15. Tshitoyan V, Dagdelen J, Weston L, Dunn A, Rong Z, Kononova O et al (2019) Unsupervised word embeddings capture latent knowledge from materials science literature. Nature 571(7763):95–98. https://doi.org/10.1038/s41586-019-1335-8
16. Lee J, Lee M, Min K (2023) Natural language processing techniques for advancing materials discovery: a short review. Int J Precis Eng Manufact Green Technol 06:10. https://doi.org/10.1007/s40684-023-00523-6
17. Pei Z, Yin J, Liaw PK, Raabe D (2023) Toward the design of ultrahigh-entropy alloys via mining six million texts. Nat Commun. https://doi.org/10.1038/s41467-022-35766-5
18. Sasidhar KN, Siboni NH, Mianroodi JR, Rohwerder M, Neugebauer J, Raabe D (2023) Enhancing corrosion-resistant alloy design through natural language processing and deep learning. Sci Adv 9(32):eadg7992. https://doi.org/10.1126/sciadv.adg7992
19. Liu Y, Yao C, Niu C, Li W, Yin J, Shen T (2021) Text mining of hypereutectic Al–Si alloys literature based on active learning. Mater Today Commun 26:102032. https://doi.org/10.1016/j.mtcomm.2021.102032
20. Pfeiffer O, Liu H, Montanelli L, Latypov M, Sen F, Hegadekatte V et al (2022) Aluminum alloy compositions and properties extracted from a corpus of scientific manuscripts and US patents. Sci Data 03(9):128. https://doi.org/10.1038/s41597-022-01215-7
21. Dunn A, Dagdelen J, Walker N, Lee S, Rosen AS, Ceder G, et al (2022) Structured information extraction from complex scientific text with fine-tuned large language models. arXiv:2212.05238
22. Kim E, Huang K, Tomala A, Matthews S, Strubell E, Saunders A et al (2017) Machine-learned and codified synthesis parameters of oxide materials. Sci Data 4:sdata2017127. https://doi.org/10.1038/sdata.2017.127
23. Boyle F, Sherman D (2006) Scopus^{TM}: the product and its development. Ser Libr 49(3):147–153. https://doi.org/10.1300/J123v49n03_12
24. Devlin J, Chang MW, Lee K, Toutanova K (2018) Bert: pre-training of deep bidirectional transformers for language understanding. arXiv:1810.04805
25. Beltagy I, Lo K, Cohan A (2019) SciBERT: a pretrained language model for scientific text. arXiv:1903.10676
26. Trewartha A, Walker N, Huo H, Lee S, Cruse K, Dagdelen J et al (2022) Quantifying the advantage of domain-specific pre-training on named entity recognition tasks in materials science. Patterns 3(4):100488. https://doi.org/10.1016/j.patter.2022.100488
27. Gupta T, Zaki M, Krishnan NMA, Mausam M (2022) MatSciBERT: a materials domain language model for text mining and information extraction. NPJ Comput. Mater. 12:8. https://doi.org/10.1038/s41524-022-00784-w
28. Cohere LLM API. Accessed 30 Sept 2023. https://cohere.com/
29. Liang P, Bommasani R, Lee T, Tsipras D, Soylu D, Yasunaga M, et al (2022) Holistic evaluation of language models. arXiv:2211.09110
30. Lin CY (2004) Rouge: A package for automatic evaluation of summaries. In: Text summarization branches out, pp 74–81
31. Grootendorst M (2022) BERTopic: neural topic modeling with a class-based TF-IDF procedure. arXiv:2203.05794
32. Mrówka G (2010) Influence of chemical composition variation and heat treatment on microstructure and mechanical properties of 6xxx alloys. Arch Mater Sci Eng 12:46
33. Usta M, Glicksman M, Wright R (2004) The effect of heat treatment on Mg2Si coarsening in aluminum 6105 alloy. Metall Mater Trans A 02(35):435–438. https://doi.org/10.1007/s11661-004-0354-7
34. Jawalkar C, Verma AS, Suri N et al (2017) Fabrication of aluminium metal matrix composites with particulate reinforcement: a review. Mater Today Proc 4(2):2927–2936. https://doi.org/10.1016/j.matpr.2017.02.174
35. Arunkumar S, Sundaram MS, Vigneshwara S et al (2020) A review on aluminium matrix composite with various reinforcement particles and their behaviour. Mater Today Proc 33:484–490. https://doi.org/10.1016/j.matpr.2020.05.053
36. Wang X, Jha A, Brydson R (2004) In situ fabrication of Al_3Ti particle reinforced aluminium alloy metal-matrix composites.

Mater Sci Eng, A 364(1–2):339–345. https://doi.org/10.1016/j.msea.2003.08.049
37. Menzemer C, Lam PC, Srivatsan TS, Wittel CF (1999) An investigation of fusion zone microstructures of welded aluminum alloy joints. Mater Lett 41(4):192–197. https://doi.org/10.1016/S0167-577X(99)00129-9
38. Myhr OR, Grong Ø, Fjær HG, Marioara CD (2004) Modelling of the microstructure and strength evolution in Al–Mg–Si alloys during multistage thermal processing. Acta Mater 52(17):4997–5008. https://doi.org/10.1016/j.actamat.2004.07.002
39. Robles Hernández FC, Sokolowski JH (2006) Comparison among chemical and electromagnetic stirring and vibration melt treatments for Al-Si hypereutectic alloys. J Alloy Compd 426(1):205–212. https://doi.org/10.1016/j.jallcom.2006.09.039
40. Dash SS, Chen D (2023) A review on processing–microstructure–property relationships of Al–Si alloys: recent advances in deformation behavior. Metals. https://doi.org/10.3390/met13030609
41. Kim JC, Nishida Y, Arima H, Ando T (2003) Microstructure of Al-Si-Mg alloy processed by rotary-die equal channel angular pressing. Mater Lett 57(11):1689–1695. https://doi.org/10.1016/S0167-577X(02)01053-4
42. Natori K, Utsunomiya H, Tanaka T (2017) Improvement in formability of semi-solid cast hypoeutectic Al-Si alloys by equal-channel angular pressing. J Mater Process Technol 240:240–248. https://doi.org/10.1016/j.jmatprotec.2016.09.022
43. Al-Qutub AM, Allam IM, Qureshi TW (2006) Effect of submicron Al2O3 concentration on dry wear properties of 6061 aluminum based composite. J Mater Process Technol 172(3):327–331. https://doi.org/10.1016/j.jmatprotec.2005.10.022
44. Mahdavi S, Akhlaghi F (2011) Effect of SiC content on the processing, compaction behavior, and properties of Al6061/SiC/Gr hybrid composites. J Mater Sci 03(46):1502–1511. https://doi.org/10.1007/s10853-010-4954-x
45. Yu H, Huang X, Lei F, Tan X, Han Y (2013) Preparation and electrochemical properties of Cr(III)-Ti-based coatings on 6063 Al alloy. Surf Coat Technol 03(218):137–141. https://doi.org/10.1016/j.surfcoat.2012.12.042
46. Chong Z, Yang X, Wang Y, Zhang DQ, Chen Y (2019) Synergistic effect between glutamic acid and rare earth cerium (III) as corrosion inhibitors on AA5052 aluminum alloy in neutral chloride medium. Ionics 03:25. https://doi.org/10.1007/s11581-018-2605-4
47. Li T, Li X, Dong C, Cheng Y (2010) Characterization of atmospheric corrosion of 2A12 aluminum alloy in tropical marine environment. J Mater Eng Perform 06(19):591–598. https://doi.org/10.1007/s11665-009-9506-7
48. Ghosh R, Venugopal A, Rao S, Narayanan P, Pant B, Cherian RM (2018) Effect of temper condition on the corrosion and fatigue performance of AA2219 aluminum alloy. J Mater Eng Perform 01(27):423–433. https://doi.org/10.1007/s11665-018-3125-0
49. Osório WR, Spinelli JE, Ferreira IL, Garcia A (2007) The roles of macrosegregation and of dendritic array spacings on the electrochemical behavior of an Al-4.5wt% Cu alloy. Electrochimica Acta. 52(9):3265–3273. https://doi.org/10.1016/j.electacta.2006.10.004
50. Ma J, Wen J, Li Q, Zhang Q (2013) Electrochemical polarization and corrosion behavior of Al–Zn–In based alloy in acidity and alkalinity solutions. Int J Hydrogen Energy 38(34):14896–14902. https://doi.org/10.1016/j.ijhydene.2013.09.046
51. Andreatta F, Terryn H, de Wit JHW (2004) Corrosion behaviour of different tempers of AA7075 aluminium alloy. Electrochimica Acta 49(17):2851–2862. https://doi.org/10.1016/j.electacta.2004.01.046
52. Venugopal V, Pai S, Olivetti E (2022) MatKG: the largest knowledge graph in materials science–entities, relations, and link prediction through graph representation learning. arXiv:2210.17340
53. Lewis P, Perez E, Piktus A, Petroni F, Karpukhin V, Goyal N et al (2020) Retrieval-augmented generation for knowledge-intensive NLP tasks. Adv Neural Inf Process Syst 33:9459–9474