LITERATURE MINING WITH LARGE LANGUAGE MODELS TO ASSIST THE DEVELOPMENT OF SUSTAINABLE BUILDING MATERIALS

Yifei Duan [1], **Yixi Tian** [1], **Soumya Ghosh** [2,3], **Richard Goodwin** [3], **Vineeth Venugopal** [1], **Jeremy Gregory** [1], **Jie Chen** [2,3], **Elsa Olivetti** [1]
[1] Massachusetts Institute of Technology, [2] MIT-IBM Watson AI Lab, [3] IBM Research
{duanyf99, yixitian, vineethv, jgregory, elsao}@mit.edu
{ghoshso, chenjie, rgoodwin}@us.ibm.com

ABSTRACT

Concrete industry, as one of the significant sources of carbon emissions, drives the urgency for its decarbonization that requires a shift to alternative materials. However, the absence of systematic knowledge summary remains a challenge for further development of sustainable building materials. This work offers a cost-efficient strategy for information extraction tasks in complex terminology settings using small (2.8B) large language models (LLMs) with well-designed instruction-completion schemes and fine-tuning strategies, introducing a dataset cataloging civil engineering applications of alternative materials. The Multiple Choice instruction scheme significantly improves model accuracies in entity inference from non-Noun-Phrase sources, with supervised fine-tuning benefiting from straightforward tokenized representations of choices. We also demonstrate the utility of the dataset by extracting valuable insights into promising applications of alternative materials from knowledge graph representations.

1 INTRODUCTION

Concrete production is a major contributor to industrial greenhouse gas (GHG) emissions, constituting 8-9% of the global CO_2 emissions (Ellis et al., 2020). As construction projects drive economic growth, the urgency to decarbonize building material consumption intensifies, necessitating a sustainable shift to alternative or secondary materials crucial for combating climate change as the majority of emissions are process- instead of energy-related. (Belaïd, 2022; Miller et al., 2021; Monteiro et al., 2017). Previous studies have extensively investigated using processed natural mineral materials (e.g., metakaolin), recycled demolition and construction waste, industrial residues (e.g., silica fume, coal ashes, metallurgical slags), and agricultural and municipal solid waste incineration (MSWI) residues to substitute the constituents of concrete according to their characteristics, including Portland cement, fine aggregate, and coarse aggregate (Juenger et al., 2019; Snellings et al., 2023; Kurniati et al., 2023).

The lack of a systematic summary impedes the advancement of commercially viable climate-friendly concrete production and the broader utilization of sustainable construction materials. For a comprehensive exploration of alternative material possibilities in construction for the decarbonization goals, an exhaustive literature review is indispensable to enrich the knowledge base of the applications and offer optimal recommendations for substitution mixture. Achieving this task is time- and labour-consuming both manually and through traditional Natural Language Processing (NLP) methods that rely on massive labeling. Therefore, we propose to leverage large language models (LLM) for the literature mining task to efficiently address this complexity.

NLP methods have been employed in chemistry and materials research for literature mining across various domains, such as drug discovery (Öztürk et al., 2021; Liu et al., 2021), solid-state synthesis (He et al., 2020), material discovery (Dunn et al., 2022; Xie et al., 2023; Munjal et al., 2023; Walker et al., 2023) etc. Despite the breadth of topics covered, previous studies have predominantly focused on named-entity recognition (NER) and relation extraction (RE) for molecules character-

ized by precise chemical compositions and well-defined properties, often with synthetic data. The clarity of compounds is seen as a prior condition for such literature-mining works (Krallinger et al., 2017; Gupta et al., 2022). However, within the concrete realm and particularly in the quest for sustainable strategies, the materials of interest encompass secondary and natural mineral materials, which exhibit highly intricate and variable compositions. Consequently, the materials and their applications in the studies are typically delineated through explanatory descriptions rather than noun phrases (NP), lacking universal standards in terminology usage, unlike most other scientific topics. Such complexity inherent in the corpus renders direct NER and RE extremely challenging (Nasar et al., 2021; Li et al., 2022), posing significant obstacles to effective academic communication and comprehension.

This work presents a comprehensive methodology demonstrating how meticulous instruction schemes and fine-tuning strategies can achieve state-of-the-art performances in complex scientific information extraction. The extracted dataset cataloging the civil engineering applications of alternative materials, using normalized terms for summarization, is a first of its kind. The comprehensive instruction-completion schemes were developed to handle complex text, including non-NP sources. With small autoregressive language models (2.8B) achieving superior entity-level accuracy compared to pre-trained GPT-3.5 (175B) in scientific tasks, our work offers a cost-efficient strategy for extraction tasks in complex settings, unlike previous approaches that often lack entity-level accuracy report or rely on expensive large models (e.g., GPT-3/3.5) for higher accuracy. The dataset is a foundational resource for further studies into alternative materials applications in sustainable building materials. We also showcase the utility of the dataset through graph representation and descriptive and predictive analyses of material-application links.

2 METHODS

2.1 DATASET

An initial collection of 51,295 papers on concrete studies was retrieved and subsequently refined to 6,995 papers related explicitly to alternative materials. To fine-tune LLMs, 102 papers were manually annotated, extracting information regarding three key entities: (1) the alternative raw materials under investigation; (2) their respective applications; and (3) the laboratory products subject to engineering testing. Notably, a single paper may encompass multiple alternative raw materials and products, with each material potentially serving various applications. These 102 annotated papers were divided into a training corpus of 82 papers and a testing corpus comprising 20 papers, with examples then derived and augmented (see permutation in 2.2) to form the training and testing sets.

2.2 LLM LITERATURE MINING

Instruction-based Entity Inference Schemes:

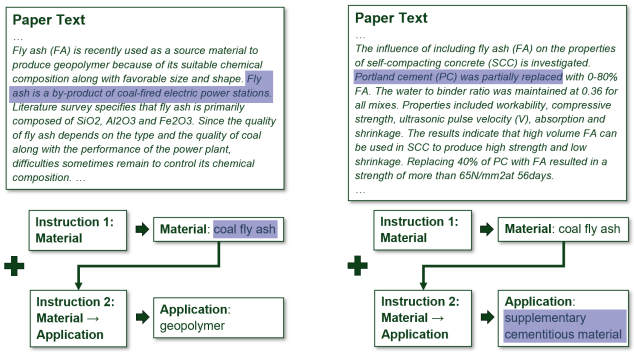

Figure 1: Information extraction examples, necessitating logical inference rather than conventional named-entity recognition

As illustrated in Figure 1, the tasks of this study require complex logical inference from non-NP source text to extract the desired information, making conventional named-entity recognition (NER)

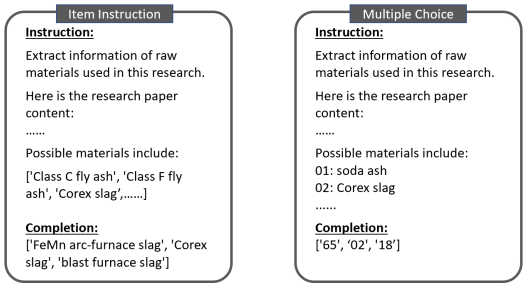

Figure 2: Comparison of two instruction-completion schemes.

unsuitable. Separate instructions were devised for extracting the three entities (material, application, product), with application extraction dependent on material extraction.

To accommodate the logical inference in extraction, the tasks were structured as multiple-choice problems, with provided choices in the instructions to guide model inference. A total of 75 material options, 13 application options, and 15 product options have been predefined by domain experts, covering potential studies comprehensively and an "unknown" option for each category to address potential model hallucination. Additionally, two distinct instruction-completion schemes were developed, with an example outlined in Figure 2.

- **Item Instruction:** the problem setting was multiple-choice, but possible choices were provided as a list of items without notations by any symbols.
- **Multiple Choice:** the choices were provided with double-digit notations.

Permutation of choice orders was performed for different examples to address the potential issue of LLM sensitivity to option ordering (Pezeshkpour & Hruschka, 2023), with permutated examples from a single paper being entirely in either training or testing set to avoid data leakage.

Models & Fine-tuning:

Two small, open-source LLMs, pythia-2.8B (Luo et al., 2023) and dolly-3B (Conver et al., 2023), were fine-tuned for the information extraction-inference tasks. In contrast to large models such as GPT-3.5 (175B), these small models offered reduced computational expenses and decreased time and memory usage. Furthermore, pythia and dolly were pertinent contrasts, sharing identical tokenizers, overall model architectures, and sizes (pythia-2.8B and dolly-3B differed only in naming convention). The primary distinction was that dolly models undergo additional fine-tuning with common sense instruction-following data after pre-training. In our work, supervised fine-tuning was performed using the instruction-completion data with different schemes.

3 RESULTS

Table 1: Test set performance of fine-tuned models and the GPT-3.5 few-shot baseline.

Model	Instruction Scheme	F1 Score	Precision	Recall
pythia-2.8B	Item Instruction	77.0	78.2	75.7
pythia-2.8B	Multiple Choice	**79.0**	**81.2**	**77.0**
pythia-2.8B	Without Options	30.5	33.3	28.1
dolly-3B	Multiple Choice	69.9	71.0	68.9
dolly-3B	Item Instruction	60.4	61.3	59.5
dolly-3B	Without Options	20.3	20.8	19.8
gpt-3.5 @4-shot	Item Instruction	57.2	62.8	52.6
gpt-3.5 @4-shot	Multiple Choice	51.9	46.8	58.1

Table 1 presents a comparison of test set accuracy performances among the 2.8B models fine-tuned with various instruction schemes and the GPT-3.5 few-shot baseline. The most notable performances are observed in pythia fine-tuned with the Multiple Choice scheme, achieving a test F1 score of 79.0%, precision of 81.2%, and recall of 77.0%. This model-scheme combination outperforms

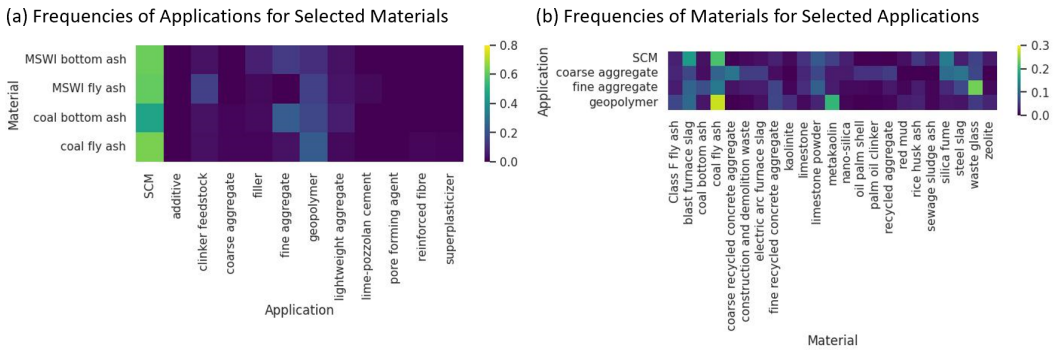

Figure 3: Normalized weights for selected material-application edges (values along each row add up to 1.0). Cell values show (a) the frequencies of linked applications for selected materials; (b) the frequencies of linked materials for selected applications

all others across all accuracy metrics. Furthermore, the best accuracies for both pythia and dolly following supervised fine-tuning surpass those of the pre-trained GPT-3.5 with few-shot learning, underscoring the potential of small, free models to attain state-of-the-art performances in scientific tasks requiring domain expertise.

The comparison of different instruction schemes reveals that including choices within the instructions enhances model performance for the complex information extraction tasks requiring logical inference. Specifically, the Multiple Choice scheme, utilizing double-digit notations, notably boosts accuracies. These results underscore the impact of tokenization simplicity on information extraction effectiveness. In supervised fine-tuning, straightforward tokenized representations associated with the options reduce uncertainties in answer generation, while for pre-trained large models, the Item Instruction scheme may be preferable for its clarity, given the limitations of the few-shot setting. Despite that the instruction tuning of dolly typically enhances its performance in common sense instruction-following tasks, pythia consistently outperforms dolly post-fine-tuning, highlighting the importance of considering knowledge domains and NLP task types in model selection.

The fine-tuned model extracts information from an extensive corpus of unannotated papers, which is subsequently utilized to construct a knowledge graph. Figure 3 illustrates edge weights normalized by material and application, highlighting promising applications and the frequently studied materials. Across various ash residues, supplementary cementitious materials (SCMs) predominate as the most studied application, with frequencies exceeding 60%, except for coal bottom ash (46.3%). Geopolymer emerges prominently for coal fly ash, clinker feedstock is notable for MSWI fly ash, and fine aggregate is significant for coal and MSWI bottom ash. Alongside industry-adopted SCMs coal fly ash, silica fume, and blast furnace slag (Young et al., 2019; DeRousseau et al., 2019), whose supplies are expected to decline (Juenger et al., 2019), commonly studied materials for SCMs also include promising alternatives like limestone powder, rice husk ash, and waste glass. Metakaolin and coal fly ash emerge as promising raw materials in the geopolymer realm. The studies on waste glass and recycled concrete aggregate for fine and coarse aggregate purposes, respectively, surpass those on industrial residues. Most of these links point to less developed applications in industrial practice, providing valuable insights into priority areas for deploying alternative materials. The knowledge graph representation facilitates graph analysis, offering deeper insights for future studies. For instance, link prediction can be employed to guide further research towards previously overlooked potential applications of alternative materials. Appendix B illustrates one approach to achieve this, utilizing node similarity analysis in conjunction with existing material-application edge weights.

4 Conclusions

This work demonstrates a novel methodology showcasing how precise instruction-completion schemes and fine-tuning strategies yield cutting-edge performance using small LLMs in scientific information extraction. It introduces a pioneering dataset cataloging civil engineering applications of alternative materials, serving as a foundational resource for further studies to facilitate the sustainable transition of the concrete industry. It offers a cost-efficient strategy for extraction tasks in complex settings. The Multiple Choice scheme, with more straightforward notations and resulting tokenized representations, substantially enhances model accuracies for complex entity inference from non-NP sources. The knowledge graphs were enabled to construct from unannotated papers in-

formation extraction, revealing valuable insights into the most promising applications of alternative materials, providing directions for industrial practice and further researches.

REFERENCES

Fateh Belaïd. How does concrete and cement industry transformation contribute to mitigating climate change challenges? *Resources, Conservation amp; Recycling Advances*, 15:200084, November 2022. ISSN 2667-3789. doi: 10.1016/j.rcradv.2022.200084. URL http://dx.doi.org/10.1016/j.rcradv.2022.200084.

Mike Conver, Matt Hayes, and Ankit Mathur. Free dolly: Introducing the world's first truly open instruction-tuned llm. 2023. URL https://www.databricks.com/blog/2023/04/12/dolly-first-open-commercially-viable-instruction-tuned-llm.

M.A. DeRousseau, E. Laftchiev, Joseph R. Kasprzyk, Rajagopalan Balaji, and Wil V. Srubar III. A comparison of machine learning methods for predicting the compressive strength of field-placed concrete. *Construction and Building Materials*, 228:116661, 2019.

Alexander Dunn, John Dagdelen, Nicholas Walker, Sanghoon Lee, Andrew S Rosen, Gerbrand Ceder, Kristin Persson, and Anubhav Jain. Structured information extraction from complex scientific text with fine-tuned large language models. *arXiv preprint arXiv:2212.05238*, 2022. URL https://doi.org/10.48550/arXiv.2212.05238.

Leah D. Ellis, Andres F. Badel, Miki L. Chiang, and Yet-Ming Chiang. Toward electrochemical synthesis of cement—an electrolyzer-based process for decarbonating caco3 while producing useful gas streams. *Proceedings of the National Academy of Sciences*, 117:12584–12591, 2020.

Gupta, Tanishq, Mohd Zaki, N.M. A. Krishnan, and Mausam. Matscibert: A materials domain language model for text mining and information extraction. *npj Computational Materials*, 8:102, 2022.

Tanjin He, Wenhao Sun, Haoyan Huo, Olga Kononova, Ziqin Rong, Vahe Tshitoyan, Tiago Botari, and Gerbrand Ceder. Similarity of precursors in solid-state synthesis as text-mined from scientific literature. *Chemistry of Materials*, 32:7861–7873, 2020.

Maria C.G. Juenger, Ruben Snellings, and Susan A. Bernal. Supplementary cementitious materials: New sources, characterization, and performance insights. *Cement and Concrete Research*, 122: 257–273, August 2019. ISSN 0008-8846. doi: 10.1016/j.cemconres.2019.05.008. URL http://dx.doi.org/10.1016/j.cemconres.2019.05.008.

Martin Krallinger, Obdulia Rabal, Analia Lourenco, Julen Oyarzabal, and Alfonso Valencia. Information retrieval and text mining technologies for chemistry. *Chemical reviews*, 117:7673–7761, 2017.

Eka Oktavia Kurniati, Federico Pederson, and Hee-Jeong Kim. Application of steel slags, ferronickel slags, and copper mining waste as construction materials: A review. *Resources, Conservation and Recycling*, 198:107175, November 2023. ISSN 0921-3449. doi: 10.1016/j.resconrec.2023.107175. URL http://dx.doi.org/10.1016/j.resconrec.2023.107175.

Jing Li, Aixin Sun, Jianglei Han, and Chenliang Li. A survey on deep learning for named entity recognition. *IEEE Transactions on Knowledge and Data Engineering*, 34:50–70, 2022.

Zhichao Liu, Ruth A Roberts, Madhu Lal-Nag, Xi Chen, Ruili Huang, and Weida Tong. Ai-based language models powering drug discovery and development. *Drug Discovery Today*, 26:2593–2607, 2021.

Ziyang Luo, Can Xu, Pu Zhao, Qinfeng Sun, and Xiubo Geng. Wizardcoder: Empowering code large language models with evol-instruct. *arXiv preprint arXiv:2306.08568*, 2023. URL https://doi.org/10.48550/arXiv.2306.08568.

Sabbie A. Miller, Guillaume Habert, Rupert J. Myers, and John T. Harvey. Achieving net zero greenhouse gas emissions in the cement industry via value chain mitigation strategies. *One Earth*, 4(10):1398–1411, October 2021. ISSN 2590-3322. doi: 10.1016/j.oneear.2021.09.011. URL http://dx.doi.org/10.1016/j.oneear.2021.09.011.

Paulo J. M. Monteiro, Sabbie A. Miller, and Arpad Horvath. Towards sustainable concrete. *Nature Materials*, 16(7):698–699, June 2017. ISSN 1476-4660. doi: 10.1038/nmat4930. URL http://dx.doi.org/10.1038/nmat4930.

Mrigi Munjal, Thorben Prein, Vineeth Venugopal, Kevin Huang, and Elsa Olivetti. Extracting a database of challenges and mitigation strategies for sodium-ion battery development. In *AI for Accelerated Materials Design*, 2023. URL https://openreview.net/pdf?id=3GiwwOJ1be.

Zara Nasar, Syed Waqar Jaffry, and Muhammad Kamran Malik. Named entity recognition and relation extraction: State-of-the-art. *ACM Computing Surveys (CSUR)*, 54:1–39, 2021.

Pouya Pezeshkpour and Estevam Hruschka. Large language models sensitivity to the order of options in multiple-choice questions. *arXiv preprint arXiv:2308.11483*, 2023. URL https://doi.org/10.48550/arXiv.2308.11483.

Ruben Snellings, Prannoy Suraneni, and Jørgen Skibsted. Future and emerging supplementary cementitious materials. *Cement and Concrete Research*, 171:107199, September 2023. ISSN 0008-8846. doi: 10.1016/j.cemconres.2023.107199. URL http://dx.doi.org/10.1016/j.cemconres.2023.107199.

Nicholas Walker, Sanghoon Lee, John Dagdelen, Kevin Cruse, Samuel Gleason, Alexander Dunn, Gerbrand Ceder, A Paul Alivisatos, Kristin A Persson, and Anubhav Jain. Extracting structured seed-mediated gold nanorod growth procedures from scientific text with llms. *Digital Discovery*, 2:1762–1782, 2023.

Tong Xie, Yuwei Wan, Wei Huang, Yufei Zhou, Yixuan Liu, Qingyuan Linghu, Shaozhou Wang, Chunyu Kit, Clara Grazian, Wenjie Zhang, and Bram Hoex. Large language models as master key: Unlocking the secrets of materials science with gpt. *arXiv preprint arXiv:2304.02213*, 2023. URL https://doi.org/10.48550/arXiv.2304.02213.

Benjamin A. Young, Alex Hall, Laurent Pilon, Puneet Gupta, and Gaurav Sant. Can the compressive strength of concrete be estimated from knowledge of the mixture proportions?: New insights from statistical analysis and machine learning methods. *Cement and Concrete Research*, 115:379–388, 2019.

Hakime Öztürk, Arzucan Özgür, Philippe Schwaller, Teodoro Laino, and Elif Ozkirimli. Exploring chemical space using natural language processing methodologies for drug discovery. *Drug Discovery Today*, 25:689–705, 2021.

A KNOWLEDGE GRAPH REPRESENTATION

With the construction of the knowledge graph from extracted information, the summary of commonly studied material-application links shown in Figure 3 can also be analyzed in forms including subgraph representations (Figure A.1) and un-normalized edge weights (Figure A.2).

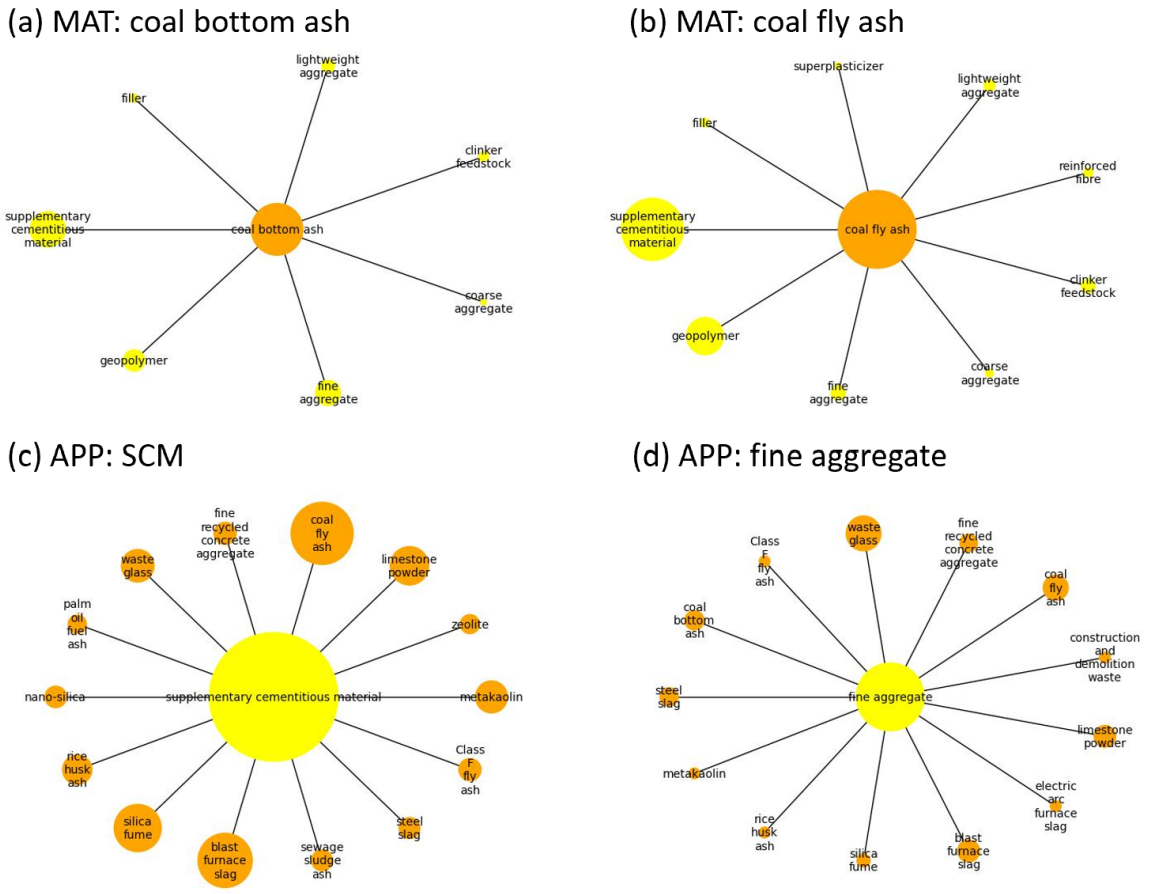

Figure A.1: Knowledge Graph Representations of Selected Subgraphs

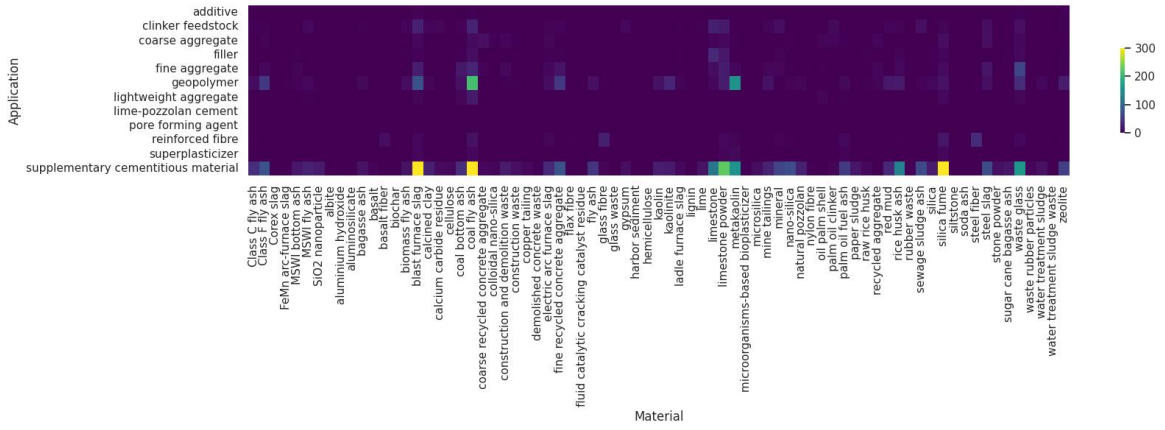

Figure A.2: Heatmap showing MAT-APP Edge Weights before any normalization

B LINK PREDICTION

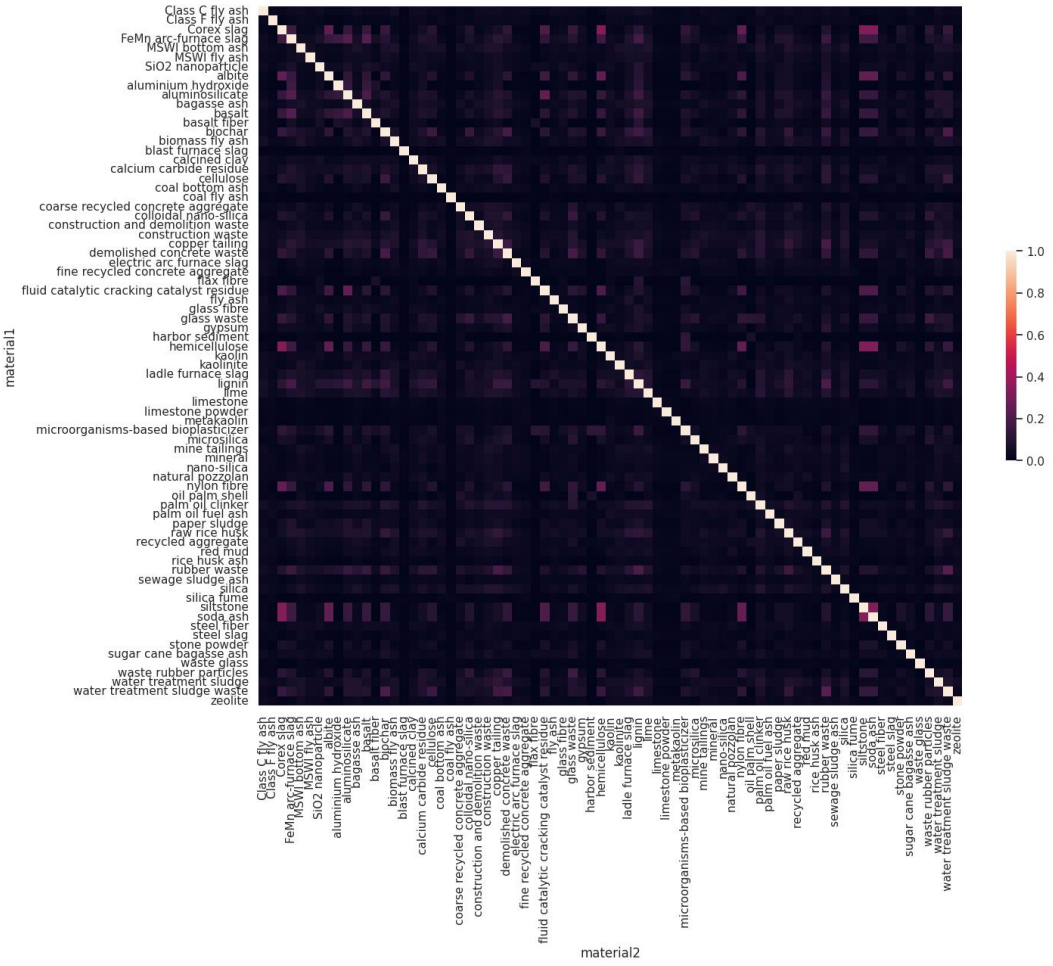

Figure B.1: Heatmap showing material similarity based on knowledge graph local structural similarity

Figure B.1 shows one way to quantify material similarity using Jaccard coefficients calculated through local graph structural similarity of the material nodes.

With the adjacency matrix containing edge weights of the existing material-application links, as well as the Jaccard matrix containing node similarity of materials, likelihood of potential new links can be calculated. Figure B.2 illustrates the results, with all non-zero entries indicating the likelihood of new link between material-application pairs that don't exhibit existing edges.

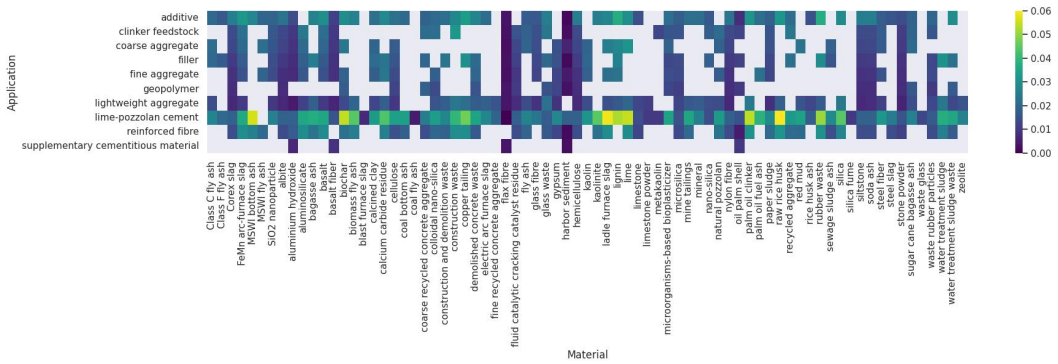

Figure B.2: Heatmap showing material-application link prediction results