Iterative Corpus Refinement for Materials Property Prediction Based on Scientific Texts

Lei Zhang 🖂, Markus Stricker

Interdisciplinary Centre for Advanced Materials Simulation, Ruhr-University Bochum, Universitätsstraße 150, 44780 Bochum, Germany {lei.zhang-w2i, markus.stricker}@rub.de

Abstract. The discovery and optimization of materials for specific applications is hampered by the practically infinite number of possible elemental combinations and associated properties, also known as the 'combinatorial explosion'. By nature of the problem, data are scarce and all possible data sources should be used. In addition to simulations and experimental results, the latent knowledge in scientific texts is not yet used to its full potential. We present an iterative framework that refines a given scientific corpus by strategic selection of the most diverse documents, training Word2Vec models, and monitoring the convergence of composition-property correlations in embedding space. Our approach is applied to predict high-performing materials for oxygen reduction (ORR), hydrogen evolution (HER), and oxygen evolution (OER) reactions for a large number of possible candidate compositions. Our method successfully predicts the highest performing compositions among a large pool of candidates, validated by experimental measurements of the electrocatalytic performance in the lab. This work demonstrates and validates the potential of iterative corpus refinement to accelerate materials discovery and optimization, offering a scalable and efficient tool for screening large compositional spaces where reliable data are scarce or non-existent.

Keywords: text mining · materials informatics · corpus refinement

1 Introduction

The discovery of new materials has traditionally relied on experimental intuition and trial-and-error methods, where researchers manually combined and tested materials, guided by experience and theoretical knowledge [6,7,26]. These methods have led to significant breakthroughs, such as platinum-based electrocatalysts for fuel cells [5] and advanced alloys for aerospace applications [4]. However, they are time-consuming, resource-intensive, and difficult to scale, particularly as material systems containing more than one or two principle elements become more complex. *Modern* material systems, such as high-entropy alloys [15] or oxides [27], multi-principal element compounds [9], involve a large number of possible and tunable compositions, which represents a possibility to use them as 'discovery platforms' [2]. However, the sheer number of possible compositions makes

experimental searches impractical or even impossible. Furthermore, global challenges such as the switch to renewable energy and sustainability require faster and more efficient ways to discover new and optimize existing materials [11,17].

In recent years, computational methods have become powerful tools for materials discovery. Simulation techniques such as density functional theory (DFT) can predict material properties, reducing the need for extensive experimental trials at the cost of the energy spent in high performance computing centers [10,18,19]. Machine learning has added another dimension by finding patterns in structured datasets and predicting properties for unexplored compositions [13,16]. Although these methods have been successful in narrowing the search space, they have limitations. DFT simulations are computationally expensive, in particular when small ≤ 1 atom-% compositional changes require many calculation to achieve a statistically correct property prediction for one composition point due to the different random distributions of elements and their impact on material properties. Supervised machine learning models to substitute expensive DFT simulations require high-quality datasets that are available at scale and variety to match the parameter space of technologically interesting materials.

The scientific literature offers an alternative resource for material discovery. Research articles and patents contain hidden knowledge encoded in text about composition-property relationships from experimental results and theoretical approaches [25]. Natural language processing (NLP) methods like Word2Vec [14] and Doc2Vec [12] can extract this knowledge by turning textual data into vector representations. These vectors capture relationships and correlations between words, which allows to link 'material dimensions', such as composition, to properties, e.g. electrocatalytic performance, and to develop models for the prediction of new high-performing candidate compositions. Unlike simulations or structured datasets, text mining can leverage unstructured data, providing access to this latent knowledge.

Despite its potential, using the scientific literature effectively presents challenges. Not all text sources are equally relevant, and including too many irrelevant or redundant sources can reduce the predictive quality of the models. The increasing volume of scientific literature, including artificial intelligence (AI)generated content, makes this issue even more pressing. Many AI-generated texts are repetitive or inaccurate, introducing noise into the dataset, and if repetetively used to retrain the models eventually leading to their *collapse* [20,21]. Effective methods are therefore needed to filter out 'low-quality content' to ensure that models are trained on the most meaningful information.

Another challenge is scalability. NLP models can process large datasets, but training them on massive corpora is expensive and it is unclear if more data actually results in a better model. Static, corpus-based models also struggle to adapt to specific tasks. For instance, a model trained on a broad corpus may not effectively capture the relationships required to predict material properties for a specific system accurately. The term 'material system' here refers to a fixed set of elements which can be mixed in an arbitrary proportion. To address these challenges, there is a need for methods that allow to tailor a training corpus w.r.t. specific prediction tasks.

In this work, we propose an iterative framework to address these challenges. First we start with a broad collection of abstracts from scientific papers, furtheron referred to as 'documents', and use Doc2Vec embeddings [3] to create a map of abstracts. This map allows us to then derive an ordered list based on greedy selection which represents the most diverse documents to avoid information duplication and information fuzziness. From this list, we use batches of 50 documents to train Word2Vec [8,14] models. In their respective embedding space, we measure the change of the centroid of the embeddings w.r.t. our specific prediction task and stop adding new document batches once a convergence criterion is met. The resulting Word2Vec model then constitutes our material system-optimized model which we use to predict reaction-specific highperformance candidate compositions for electrocatalysis. To demonstrate the effectiveness of our approach, we predict highest performing compositions from three different material systems for three different electrocatalytic reactions: oxygen reduction (ORR), hydrogen evolution (HER), and oxygen evolution (OER) and validate them against experimental measurements. While electrocatalysts are used as a case study, our framework is general and can be applied to a wide range of materials discovery tasks.

2 Experiments

All code and data needed to reproduce our workflow are available. The code can be found in [29], references to all datasets are provided when they are introduced further below.

2.1 Corpus Collection and Preprocessing

Fig. 1 shows a schematic overview of our methodology. First, we collect a relevant corpus of documents using the PaperCollector module in MatNexus [28]. As source for the documents we use the application programming interfaces (APIs) of Scopus and and limit ourselves to use open access-only publications up to including the year 2023 which results in 6506 papers. The retrieved abstracts and metadata are stored in a structured comma-separated values (CSV) file. Text preprocessing is performed using the TextProcessor module, which removes licensing statements, filters common English stopwords, and retains domain-specific terms, such as chemical element symbols. Tokenization is applied to prepare the text for embedding generation.

2.2 Embedding Generation and Greedy Selection

The initial document embeddings are generated using a Doc2Vec model trained on the full corpus containing 6506 documents. The embeddings capture the semantic content of each document, mapping the tokens into a 200-dimensional



Fig. 1. Schematic overview for iterative corpus refinement framework from literature collection, document selection, Word2Vec model training to final predictions based on Pareto front analysis.

vector space, which offers a balanced trade-off between computational efficiency and representational capacity for capturing domain-specific linguistic nuances. A greedy selection algorithm is then applied to iteratively refine the subset of documents for analysis.

Initialization: We initialize the greedy selection algorithm with the central document in embedding space. This ensures that the starting point is representative of the corpus's overall thematic distribution, i.e. an "average abstract" w.r.t. the Doc2Vec embedding space.

Iterative Expansion: Subsequent documents are selected by identifying the farthest (in cosine distance) from all previously selected documents. It is a classic greedy farthest point sampling except the initial point is selected always the central document in embedding space instead of a random starting point. To improve computational efficiency and focus on principal directions of variance, the 200-dimensional Doc2Vec embeddings are reduced to two dimensions using principal component analysis (PCA) before distance calculations. The selection process continues until a predetermined batch size of 50 additional documents is reached, which serves as a heuristic balance between introducing sufficient diversity in each iteration and maintaining computational manageability. This process results in an ordered list of the most different documents in batches of 50 and constitutes the basis for subsequent analysis and optimization.

A selected subset is then used to train a Word2Vec [14,8] model via the Vec -Generator module of MATNEXUS [28]. The skip-gram architecture is used with a vector size of 200, a window size of 5, and hierarchical softmax. The vector size of 200 provides sufficient capacity to capture nuanced relationships between terms while maintaining training efficiency. A window size of 5 is used to reflect a

moderate contextual range, allowing the model to learn meaningful co-occurrence patterns without overextending semantic connections. Hierarchical softmax is chosen for its effectiveness in handling large vocabularies with improved training speed over full softmax.

Material Similarity Calculation: We create the representation of 'a material' w.r.t. its composition by a linearly weighted superposition of the embeddings for pure elments, e.g. 'Pt' for platinum, 'Pd' for palladium, etc. The linear weights model the composition. That is, the representation of a composition of 20% element A and 80% element B means to linearly superpose the word embeddings R_i of elements A and B $R_{A_{20}B_{20}} = 0.2R_A + 0.8R_B$. We then calculate the similarity scores of each composition to the embedding vectors of the two properties 'dielectric' and 'conductivity' using cosine similarity which we denote $S_{\text{dielectric}}$ and $S_{\text{conductivity}}$. Given a composition space and concentration resolution for a given material system containing several elements, we can use the two-dimensional similarity scores to calculate a centroid of the based on Ndifferent compositions in the material system as follows:

centroid =
$$\frac{1}{N} \sum_{i=1}^{N} \begin{bmatrix} S_{\text{dielectric}}(i) \\ S_{\text{conductivity}}(i) \end{bmatrix}$$
. (1)

Convergence Criterion: The previous step is then iterated with an increasing corpus size in batches of 50 documents. With each increase, we monitor the change of the centroid coordinates using Euclidean distance. If the distance between the previous and current centroid falls below a heuristically determined threshold of 0.03, we stop adding more documents and define the word embedding model *converged*. Our rationale is that when an additional batch of 50 documents does not significantly change the word embeddings, adding more documents is unnecessary or even detrimental to predictive performance.

2.3 Pareto Optimization for Candidate Selection

With the converged Word2Vec model for a given material system, Pareto optimization is applied to identify optimal trade-offs between material properties for three electrochemical reactions. For the Pareto front optimization we follow [30] which was shown to be very effective in identifying high-performing regions.

- For HER and ORR, the objective was to maximize similarity to conductivity while minimizing similarity to dielectric.
- For oxygen evolution OER, the objective was to maximize similarity to *di*electric while minimizing similarity to *conductivity*.

Compositions on the reaction-specific Pareto fronts in similarity space defined by S_i constitute our predictions for high-performing materials.



Fig. 2. Centroid distance as a function of iterations for different material systems.

3 Results

3.1 Centroid Convergence Over Iterations

Figure 2 shows the changes in centroid distance during each iteration for the six material systems we tested: AgPdPt, AgPdRu, AgPdPtRu (for ORR) [1], AgAuPdPtRh, AgAuPdPtRu (for HER) [22], and NiPdPtRu (for OER) [23]. The centroid distance measures the *shift* in the embedding space between iterations, and serves as a quantitative indicator of convergence. The first iteration constitutes the initialization and, therefore, does not have a previous iteration to compare against. Additionally, not all systems have values for every iteration because the embedding space is only defined when a trained Word2Vec model provides representations for all required elements as well as the target properties dielectric and conductivity. In other words, for some compositions, the first few batches of documents do not contain all tokens required to compute the word embeddings as well as their similarity scores. If the selected corpus for a given iteration lacks such similarities, the similarity space representation can not be established and, consequently, the centroid cannot be calculated. The number of iterations required to reach the convergence threshold of 0.03 for each material system is as follows:

- AgPdPt: Converged after 8 iterations, indicating steady refinement and early stabilization.
- AgPdRu: Required 18 iterations to stabilize, reflecting a gradual and extended refinement process.
- AgPdPtRu: Achieved convergence in 16 iterations, suggesting moderate complexity in refining its embedding space.

- AgAuPdPtRh: Stabilized after 15 iterations, indicating consistent refinement with gradual improvement.
- AgAuPdPtRu: Required 24 iterations to reach the threshold, showing a more extended effort to refine the embedding space.
- NiPdPtRu: Converged after 14 iterations, balancing between moderate refinement complexity and stabilization.

These results for different material systems exhibit a large variability in convergence behavior. AgPdPt has relatively fast convergence, while AgAuPdPtRu requires the largest number of iterations, reflecting the challenges of establishing a complex embedding space. Again, once the centroid distance falls below a user-defined, heuristically obtained threshold of 0.03, we consider the embedding space converged.

3.2 Pareto Analysis with Full Corpus vs. Selected Subset

After centroid convergence, we perform Pareto analysis following the method presented in [30] on each reaction type (ORR, HER, OER) using the final Word2Vec model for each material system. We compare the prediction of the candidate materials of the converged representations with predictions based on a model trained on all documents. This reference full-corpus model is the same for all three materials systems. Figures 3, 4, and 5 summarize the prediction metrics, highlighting the models' ability to predict high-performing materials in very different material systems and reaction types.

For ORR (Fig. 3), the selected-corpus model identified high-performing materials with subsequently measured minimum current densities at 850 mV of -0.50, -0.67, and $-0.37 \,\mathrm{mA/cm^2}$ for the AgPdPt, AgPdRu, and AgPdPtRu systems, respectively. These predictions align closely with best-performing experimentally measured electrocatalytic response, which recorded -0.58, -0.67, and $-0.37 \,\mathrm{mA/cm^2}$, respectively. The full-corpus model, while also successful in identifying high-performing materials, predicted slightly less optimal values of -0.44, -0.67, and $-0.37 \,\mathrm{mA/cm^2}$. This difference demonstrates the effectiveness of the selected-corpus model in refining the search space and focusing on the most relevant documents for a given composition space.

Fig. 4 shows the performance for the AgAuPdPtRh and AgAuPdPtRu systems for HER. The selected-corpus model achieved minimum current densities at -300 mV of -1.13 and -1.44 mA/cm², closely matching the best measured electrocatalytic response of -1.13 and -1.49 mA/cm². The full-corpus model predicts values of -1.11 and -1.41 mA/cm², which, while accurate, do not represent the actual highest-performing compositions. These results further validate the ability of the selected-corpus model to identify optimal candidates effectively.

For HER, Figure 4 presents the performance for the AgAuPdPtRh and AgAuPdPtRu systems. The selected-corpus model predicts minimum current densities at -300 mV of $-1.11 \text{ and } -1.27 \text{ mA/cm}^2$. For the AgAuPdPtRu system, the value is slightly less optimal than the full-corpus model predictions



Fig. 3. Comparison of minimum current density values at 850 mV for ORR systems (AgPdPt, AgPdRu, AgPdPtRu) derived from the original performance data, the full-corpus model, and the selected-corpus model.



Fig. 4. Comparison of minimum current density values at -300 mV for HER systems (AgAuPdPtRh, AgAuPdPtRu) derived from the original performance data, the full-corpus model, and the selected-corpus model.

of $-1.41 \,\mathrm{mA/cm^2}$. Despite this, the selected-corpus model captures essential information from the full corpus, demonstrating its ability to approximate high-performing materials with a significantly smaller number of documents.



Fig. 5. Comparison of maximum current density values at 1700 mV for the OER system (NiPdPtRu) derived from the original performance data, the full-corpus model, and the selected-corpus model.

For OER, as shown in Figure 5, the maximum current density at 1700 mV for the NiPdPtRu system is identical across all models, including the selected-corpus model, the full-corpus model, and the best measured data performance: 6.90 mA/cm^2 . This, again, confirms the consistency of our approach in capturing the best-performing materials using much fewer documents for creation of the representations.

All results collectively demonstrate that both the full-corpus and selectedcorpus models successfully predict high-performing materials across all reaction types. Notably, the selected-corpus models achieve this with a much smaller number of documents. By refining the training corpus to only the most diverse documents measured by a combination of Doc2Vec embeddings, greedy selection and convergence threshold based on the centroid change, the iterative selection process improves the predictive power of the model without compromising performance across very different scenarios.

Table 1. Statistical details for electrocatalysts under specified potentials. Abbreviations: Entries (Ori), the original number of possible compositions before text-mining selection; Entries (Full), number of compositions on the Pareto front for the Word2Vec model based on the full corpus; Entries (Selection), number of compositions on the Pareto front based on the converged model with iterative greedy selection; Selected Documents, the total number of documents used in the final iteration to train the selected-corpus model.

Material	Potential	Entries	Entries	Entries	Selected
Systems	(mV)	(Ori)	(Full)	(Selection $)$	Documents
AgPdPt	850	341	11	63	400
AgPdRu	850	342	15	10	900
AgPdPtRu	850	341	27	4	800
AgAuPdPtRh	-300	327	16	29	750
AgAuPdPtRu	-300	335	23	3	1200
NiPdPtRu	1700	4026	374	168	700

Table 2. Minimum and maximum current densities (mA/cm^2) for electrocatalysts under specified potentials. Abbreviations: Min/Max (Ori, Full, Selection), the minimum and maximum current densities for each of the three scenarios (original data, full corpus model, selected-corpus models).

Material	Min	Min	Min	Max	Max	Max
Systems	(Ori)	(Full)	(Selection $)$	(Ori)	(Full)	(Selection)
AgPdPt	-0.58	-0.44	-0.50	-0.06	-0.16	-0.06
AgPdRu	-0.67	-0.67	-0.67	-0.07	-0.36	-0.36
AgPdPtRu	-0.37	-0.37	-0.37	-0.06	-0.06	-0.34
AgAuPdPtRh	-1.13	-1.11	-1.11	-0.69	-0.73	-0.72
AgAuPdPtRu	-1.49	-1.41	-1.27	-0.79	-1.14	-1.06
NiPdPtRu	0.24	0.60	0.53	6.90	6.90	6.90

3.3 Current Density Measurements and Statistical Overview

Table 1 and Table 2 present the statistical details and measured current densities (in mA/cm^2) for each system at its corresponding potential.

ORR Systems (AgPdPt, AgPdRu, AgPdPtRu at 850 mV):

- In AgPdPt, the selection-based model retains 63 entries compared to 11 in the full-corpus model. The minimum current density shifts from -0.44 mA/cm² in the full model to -0.50 mA/cm² in the selection model, while the maximum rises to -0.06 mA/cm².
- In AgPdRu, the minimum and maximum current densities remain the same across selection-based model and full-corpus model, at -0.67 mA/cm^2 and -0.36 mA/cm^2 , respectively, with 10 entries retained in the selection model compared to 15 in the full model.
- In AgPdPtRu, the selection-based model retains 4 entries compared to 27 in the full model. The minimum current density remains -0.37 mA/cm^2 , while the maximum is lower at -0.34 mA/cm^2 compared to -0.06 mA/cm^2 in the full model.

11

HER Systems (AgAuPdPtRh, AgAuPdPtRu at -300 mV):

- In AgAuPdPtRh, the selection-based model retains 29 entries compared to 16 in the full model. The minimum current density is consistent at -1.11 mA/cm^2 , while the maximum slightly increases from -0.72 mA/cm^2 in the selection model to -0.73 mA/cm^2 in the full model.
- In AgAuPdPtRu, the selection-based model retains 3 entries compared to 23 in the full model. The minimum current density slightly increases from -1.41 mA/cm^2 in the full model to -1.27 mA/cm^2 in the selection model, while the maximum decreases from -1.14 mA/cm^2 to -1.06 mA/cm^2 .

OER System (NiPdPtRu at 1700 mV):

- NiPdPtRu starts with a large set of 4026 candidate compositions. The selection-based model retains 168 entries compared to 374 in the full-corpus model. The minimum current density decreases from 0.60 mA/cm^2 in the full model to 0.53 mA/cm^2 in the selection model, while the maximum remains consistent at 6.90 mA/cm².

All these results demonstrate that the iterative selection-based approach effectively narrows the dataset while preserving or even improving the ability to predict high-performing electrocatalysts from a large list of possible candidate compositions. The evidence in Table 2 supports the conclusion that a focused corpus can adequately capture and even surpass predicting the key characteristics of the full-corpus dataset.

4 Discussion

Our iterative selection approach consistently results in models that match or exceed the performance of the model built using the full corpus. For ORR and HER systems, the method identifies more negative (better) current densities, indicating higher catalytic performance. For OER, the selected-corpus model suggests the same high-performing compositions as the full corpus model. These results confirm that reducing the training set—when done strategically—does not compromise, and even improves the capability to predict high-performance compositions.

4.1 Advantages of Iterative Corpus Selection

A feature of our approach is the *stability criterion* that checks whether adding more information significantly changes the model's representation of materials. Once the difference between subsequent centroids in embedding space stays below a user-defined, heuristically found threshold, we stop adding more information. Note, all models use the same threshold. This ensures:

1. **Reduced Noise:** Unnecessary or very similar information is left out, preventing additional fuzziness of key semantic signals related to electrocatalytic properties.

- 12 L. Zhang and M. Stricker
- 2. Computational Efficiency: Fewer documents result in faster training and analysis.
- 3. Focus: The final Word2Vec embeddings capture terminology tailored to the studied compositions, reactions, and the properties *dielectric* and *conductiv-ity*.

4.2 Practical Impact on Electrocatalyst Screening

By identifying fewer yet relevant documents for training, this approach can speed up the overall pipeline of electrocatalyst design in particular for scenarios where experimental data are scarce and simulation-based data prove computationally intractable. Fewer experimental tests are needed because the Pareto analysis, driven by these tailored embeddings, allows to narrow down a large pool of candidate compositions to a much small, experimentally accessible likely highperforming compositions.

4.3 Mitigating the Growing Volume of AI-Generated Text

The rise of AI-generated information, specifically texts, poses new challenges. Large language models sometimes produce repetitive, paraphrased, or even wrong content by hallucination [31]. On a large scale, if generative AI models which are retrained on recursively generated output (meaning their own output), the resulting models become defective – they start producing noise [20]. Our iterative selection loop, which operates in the embedding space and discards information that does not shift the centroid meaningfully, offers a possibility to filter noise or at least avoid duplicated information: our framework avoids documents which contribute little new information. As AI-generated content continues to grow, such filtering mechanisms may become essential to maintaining reliable text-based models.

4.4 Limitations and Potential Extensions

While the proposed framework demonstrates significant strengths in refining training corpora and often even improves predictive performance, it is not without limitations. A critical challenge lies in balancing coverage and specificity. If the initial subset of documents is too small or biased, it risks omitting important details, potentially reducing the general applicability of the model. Additionally, reliance on predefined target properties, such as *dielectric* and *conductivity* as in our case require domain knowledge to be defined. Such terms might be difficult to find for other prediction scenarios and broader applicability of our method. Future work will address these challenges by exploring the following extensions:

- Human in the loop: Involve domain experts to review and validate the selected documents, ensuring that important topics or underrepresented domains are not overlooked. For example, experts could identify emerging areas or validate key materials that align with research objectives.

13

- Hybrid Models: Integrate advanced NLP methods, such as transformerbased encoders (e.g., MatBERT [24]), to complement the iterative selection process. These models may provide additional layers of semantic evaluation or help identify nuanced relationships that Word2Vec with fixed representations might miss.
- Property Expansion: Extend the framework to incorporate a broader range of material properties or performance metrics. For instance, multiobjective optimization could include properties like catalytic efficiency, stability under reaction conditions, or environmental impact, enabling more targeted materials discovery and optimization pipelines.

5 Conclusion

Our contribution introduces an iterative framework for refining scientific corpora to create word2vec representations of with a minimal set of documents for specific prediction scenarios. By dynamically selecting a subset of 'most diverse' documents sampled in Doc2Vec embedding space with a greedy selection strategy and subsequently training Word2Vec models, our approach results in representations with minimal noise w.r.t. user-defined composition spaces and target properties as well as their relationships. Applied to material systems for ORR, HER, and OER, our framework reliably identifies high-performing compositions from a large candidate pool using fewer training documents, matching or exceeding the performance of models trained on full corpora. Our predictions are verified with experimental data. This verification demonstrates the ability of our framework to efficiently use text-based knowledge for composition-property correlations.

Acknowledgments. The authors gratefully acknowledge the financial support provided by the China Scholarship Council (CSC, CSC number: 202208360048) and funding by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) through CRC 1625, project number 506711657, subprojects INF, A05.

Disclosure of Interests. The authors have no competing interests to declare.

References

- Banko, L., Krysiak, O., Schumann, W., Ludwig, A.: Electrochemical activity of several compositions in the system ag-pd-pt-ru for the oxygen reduction reaction in 0.05 m koh solution ph 12.5 (Oct 2024). https://doi.org/10.5281/zenodo. 13992986
- Batchelor, T.A., Pedersen, J.K., Winther, S.H., Castelli, I.E., Jacobsen, K.W., Rossmeisl, J.: High-entropy alloys as a discovery platform for electrocatalysis. Joule 3(3), 834-845 (2019). https://doi.org/https://doi.org/10.1016/j. joule.2018.12.015
- Bilgin, M., Sentürk, I.F.: Sentiment analysis on twitter data with semi-supervised doc2vec. In: 2017 International Conference on Computer Science and Engineering (UBMK). pp. 661–666 (2017). https://doi.org/10.1109/UBMK.2017.8093492

- 14 L. Zhang and M. Stricker
- Ezugwu, E., Bonney, J., Yamane, Y.: An overview of the machinability of aeroengine alloys. Journal of Materials Processing Technology 134(2), 233 – 253 (2003). https://doi.org/10.1016/S0924-0136(02)01042-7
- Ferreira, P., La O', G., Shao-Horn, Y., Morgan, D., Makharia, R., Kocha, S., Gasteiger, H.: Instability of pt/c electrocatalysts in proton exchange membrane fuel cells: A mechanistic investigation. Journal of the Electrochemical Society 152(11), A2256-A2271 (2005). https://doi.org/10.1149/1.2050347
- Fujishima, A., Rao, T.N., Tryk, D.A.: Titanium dioxide photocatalysis. Journal of Photochemistry and Photobiology C: Photochemistry Reviews 1(1), 1 – 21 (2000). https://doi.org/10.1016/S1389-5567(00)00002-2
- Geim, A., Novoselov, K.: The rise of graphene. Nature Materials 6(3), 183 191 (2007). https://doi.org/10.1038/nmat1849
- Goldberg, Y., Levy, O.: word2vec explained: deriving mikolov et al.'s negativesampling word-embedding method (2014), https://arxiv.org/abs/1402.3722
- Huang, P.K., Yeh, J.W., Shun, T.T., Chen, S.K.: Multi-principal-element alloys with improved oxidation and wear resistance for thermal spray coating. Advanced Engineering Materials 6(1-2), 74 - 78 (2004). https://doi.org/10.1002/adem. 200300507
- Jain, A., Shin, Y., Persson, K.A.: Computational predictions of energy materials using density functional theory. NATURE REVIEWS MATERIALS 1(1) (JAN 2016). https://doi.org/10.1038/natrevmats.2015.4
- Larcher, D., Tarascon, J.M.: Towards greener and more sustainable batteries for electrical energy storage. Nature Chemistry 7(1), 19 - 29 (2015). https://doi. org/10.1038/nchem.2085
- 12. Le, Q., Mikolov, T.: Distributed representations of sentences and documents (2014)
- Liu, Y., Zhao, T., Ju, W., Shi, S.: Materials discovery and design using machine learning. Journal of Materiomics 3(3), 159 - 177 (2017). https://doi.org/10. 1016/j.jmat.2017.08.002
- Mikolov, T., Chen, K., Corrado, G., Dean, J., Sutskever, L., Zweig, G.: word2vec. URL https://code.google.com/p/word2vec 22, 795 (2013)
- Miracle, D., Senkov, O.: A critical review of high entropy alloys and related concepts. Acta Materialia 122, 448 - 511 (2017). https://doi.org/10.1016/j. actamat.2016.08.081
- Raccuglia, P., Elbert, K.C., Adler, P.D.F., Falk, C., Wenny, M.B., Mollo, A., Zeller, M., Friedler, S.A., Schrier, J., Norquist, A.J.: Machine-learning-assisted materials discovery using failed experiments. Nature 533(7601), 73 - 76 (2016). https:// doi.org/10.1038/nature17439,
- Ragauskas, A.J., Williams, C.K., Davison, B.H., Britovsek, G., Cairney, J., Eckert, C.A., Frederick Jr., W.J., Hallett, J.P., Leak, D.J., Liotta, C.L., Mielenz, J.R., Murphy, R., Templer, R., Tschaplinski, T.: The path forward for biofuels and biomaterials. Science **311**(5760), 484 - 489 (2006). https://doi.org/10.1126/ science.1114736,
- Saal, J.E., Kirklin, S., Aykol, M., Meredig, B., Wolverton, C.: Materials design and discovery with high-throughput density functional theory: The open quantum materials database (oqmd). JOM 65(11), 1501 – 1509 (2013). https://doi.org/ 10.1007/s11837-013-0755-4
- Schmidt, J., Cerqueira, T.F., Romero, A.H., Loew, A., Jäger, F., Wang, H.C., Botti, S., Marques, M.A.: Improving machine-learning models in materials science through large datasets. Materials Today Physics 48, 101560 (2024). https://doi. org/https://doi.org/10.1016/j.mtphys.2024.101560

- 20. Shumailov, I., Shumaylov, Z., Zhao, Y., Papernot, N., Anderson, R., Gal, Y.: Ai models collapse when trained on recursively generated data. Nature 631(8022), 755–759 (Jul 2024), https://doi.org/https://doi.org/10.1038/ s41586-024-07566-y
- Shumailov, I., Shumaylov, Z., Zhao, Y., Papernot, N., Anderson, R., Gal, Y.: Ai models collapse when trained on recursively generated data. Nature 631(8022), 755 - 759 (2024). https://doi.org/10.1038/s41586-024-07566-y,
- 22. Thelen, F., Zehl, R., Limani, N., Schuhmann, W., Ludwig, A.: High-throughput seccm and edx data for the hydrogen evolution reaction in ag-au-pd-pt-ru and ag-au-pd-pt-rh thin-film materials libraries (Mar 2025). https://doi.org/10.5281/zenodo.14959252
- Thelen, F., Zehl, R., Zerdoumi, R., Bürgel, J.L., Schuhmann, W., Ludwig, A.: Dataset - accelerating combinatorial electrocatalyst discovery with bayesian optimization: A case study in the quaternary system ni-pd-pt-ru for the oxygen evolution reaction (Feb 2025). https://doi.org/10.5281/zenodo.14891704
- Trewartha, A., Walker, N., Huo, H., Lee, S., Cruse, K., Dagdelen, J., Dunn, A., Persson, K.A., Ceder, G., Jain, A.: Quantifying the advantage of domainspecific pre-training on named entity recognition tasks in materials science. Patterns 3(4), 100488 (2022). https://doi.org/https://doi.org/10.1016/j. patter.2022.100488,
- Tshitoyan, V., Dagdelen, J., Weston, L., Dunn, A., Rong, Z., Kononova, O., Persson, K.A., Ceder, G., Jain, A.: Unsupervised word embeddings capture latent knowledge from materials science literature. Nature 571(7763), 95 98 (2019). https://doi.org/10.1038/s41586-019-1335-8
- Yeh, J.W., Chen, S.K., Lin, S.J., Gan, J.Y., Chin, T.S., Shun, T.T., Tsau, C.H., Chang, S.Y.: Nanostructured high-entropy alloys with multiple principal elements: Novel alloy design concepts and outcomes. Advanced Engineering Materials 6(5), 299 - 303 (2004). https://doi.org/10.1002/adem.200300567
- 27. Zerdoumi, R., Ludwig, A., Schuhmann, W.: High entropy intermetallic compounds: A discovery platform for structure-property correlations and materials design principles in electrocatalysis. Current Opinion in Electrochemistry 48, 101590 (2024). https://doi.org/https://doi.org/10.1016/j.coelec.2024.101590
- Zhang, L., Stricker, M.: Matnexus: A comprehensive text mining and analysis suite for materials discovery. SoftwareX 26, 101654 (2024). https://doi.org/10.1016/ j.softx.2024.101654
- 29. Zhang, L., Stricker, M.: Code for "iterative corpus refinement for materials property prediction based on scientific texts" (2025), https://github.com/lab-mids/word_embedding_paper_selection
- 30. Zhang, L., Stricker, M.: Electrocatalyst discovery through text mining and multiobjective optimization (2025), https://arxiv.org/abs/2502.20860
- 31. Zhang, Y., Li, Y., Cui, L., Cai, D., Liu, L., Fu, T., Huang, X., Zhao, E., Zhang, Y., Chen, Y., Wang, L., Luu, A.T., Bi, W., Shi, F., Shi, S.: Siren's song in the ai ocean: A survey on hallucination in large language models (2023), https://arxiv.org/abs/2309.01219