Literature-Based Prediction of High-Performance Electrocatalysts

Lei Zhang[1], Markus Stricker[1]

[1]Interdisciplinary Centre for Advanced Materials Simulation, Ruhr-University Bochum, Universitaetsstraße 150, Bochum, Germany

lei.zhang-w2i@rub.de

Abstract

The discovery and optimization of high-performance materials for electrocatalysis is fundamental to the advancement of energy conversion technologies [1]. However, the vastness of the chemical design space, driven by the nearly infinite combinations of elements and processing conditions, poses a major challenge, often referred to as the ``combinatorial explosion'' [2]. Traditional approaches, which are heavily based on simulations and experimental screening, are constrained by time, cost, and the scarcity of reliable data. An underutilized yet powerful resource is the latent knowledge in the scientific literature [3] which includes, e.g., correlations between composition and material properties.

In this work, we present a literature-based framework that leverages natural language processing (NLP) techniques to predict promising electrocatalyst compositions. Using Word2Vec [4] to model semantic relationships between material compositions and performance descriptors extracted from scientific abstracts [5], we identify candidates for key electrochemical reactions [6], including the oxygen reduction reaction (ORR), hydrogen evolution reaction (HER), and oxygen evolution reaction (OER). To enhance prediction quality and data efficiency, we employ an iterative corpus refinement strategy [7] that prioritizes the most diverse and informative documents, allowing composition-property correlations to emerge more clearly in embedding space.

In regions of sparse experimental or simulation data, we combine these text-derived embeddings with Pareto front analysis to isolate high-performance candidates based solely on their linguistic similarity to target properties such as `conductivity' or `dielectric'. The resulting candidate predictions are experimentally validated and show excellent agreement with the *best* measured electrocatalytically active compositions. Our approach highlights the untapped potential of the scientific literature as a data source for predictive models in materials discovery and offers a scalable, data-efficient method for navigating large, unexplored compositional spaces.

References

[1] Z. W. Seh et al., Science, 355(6321) (2017) eaad4998
[2] S. H. Baeck et al., Journal of Combinatorial Chemistry, 4(6) (2002) 563–568.
[3] V. Tshitoyan et al., Nature, 571(7763) (2019) 95–98
[4] T. Mikolov et al., Proceedings of the International Conference on Learning Representations (ICLR), Workshop Track, (2013)
[5] L. Zhang and M. Stricker, SoftwareX, 26 (2024) 101654.
[6] L. Zhang and M. Stricker, arXiv preprint arXiv:2502.20860 (2025)
[7] L. Zhang and M. Stricker, ECML PKDD, (2025)

Figures

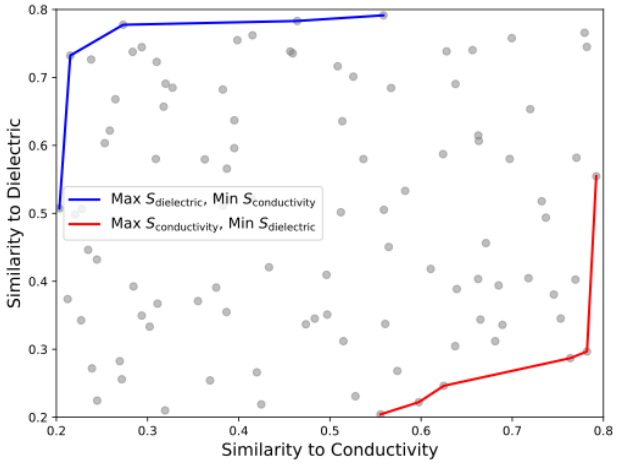

Figure 1. Illustration of Pareto-front trade-offs in conductivity–dielectric similarity space. Each point corresponds to a synthetic composition embedded by its similarity to conductivity ($S_{conductivity}$) and dielectric ($S_{dielectric}$). The blue Pareto front collects non-dominated compositions that maximise $S_{dielectric}$ while minimising $S_{conductivity}$, whereas the red Pareto front collects those that maximise $S_{conductivity}$ while minimising $S_{dielectric}$).