

Machine learning-assisted, literature-based screening of high-entropy alloys for solid state hydrogen storage

Jana Radaković¹, Katarina Batalović¹, Bojana Mamula Paskaš¹, Mirjana Medić¹, Bojana Kuzmanović¹, Nikola Novaković¹

¹ Department of nuclear and plasma physics, „VINČA" Institute of Nuclear Sciences - National Institute of the Republic of Serbia, University of Belgrade, Belgrade, Serbia

doi: 10.5937/MESCIS26039R

In recent years, the high-entropy alloys (HEAs) have emerged as a particularly promising class of candidate materials for hydrogen storage. However, their vast combinatorial composition space is far beyond what conventional trial-and-error synthesis can explore, making exhaustive experimental screening or first-principles simulations impractical. This overload of potential motivates a machine learning-based approach to materials screening to help identify thermodynamically favorable hydride-forming compositions. However, the missingness rate of reported physicochemical properties varies greatly by material class across different data sets. This leaves a high proportion of unreported labels, which inhibits predictions of target properties. Accordingly, machine learning-based predictions now often rely on text-derived chemical element or materials representations in format or word embeddings [1]. These embeddings offer feature representations that capture the chemical context found in published texts [2].

In this work, we present an end-to-end screening framework that combines data from multiple first-principles databases (Materials Project, AFLOW, and JARVIS-DFT) to assemble a training set of a great variety of materials. Each material is described by its inherent structural parameters and DFT-computed formation energy. In addition, we constructed composition-based feature vectors using the matminer library to capture materials descriptors, including valence-electron concentration, atomic radius mismatch, electronegativity variance, and d-band proxies relevant to hydrogen binding. The dataset is used to screen a systematically constructed virtual library of HEA compositions based on established systems, identifying candidates that satisfy thermodynamic windows for reversible hydrogen absorption near ambient conditions. Word embeddings trained on a materials science corpus are pooled by several different criteria into composition-level features and concatenated with classical phase-stability descriptors [3, 4], together with binary missingness indicators on the descriptors themselves. In contrast to predictive positioning of materials using combined descriptors and text-based representations, the similarity-based materials analysis operated in the feature space rather than the label space. The trained predictor scored phase-stable equiatomic candidates enumerated from a data-driven pool of chemical elements and annotated with predicted properties.

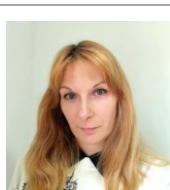
In addition to sparse property selection, the limitations of the current work stem from the fact that text-derived embeddings supporting predictive behavior inherit publication bias. Namely, chemistries investigated in greater detail have rich vector representations, while those neglected or underrepresented receive sparse vectors or very low-quality

representations. We further discuss how explicit label-side and feature-side missingness affects predictive calibration and discovery, and how it generalizes to other materials informatics tasks where literature corpora and labeled databases coexist with structurally non-random gaps.

The results demonstrate the potential of high-throughput computational screening combined with interpretable machine learning to accelerate the identification of novel HEA-based hydrogen storage materials, providing prioritized targets for experimental validation.

References

- [1] V. Tshitoyan, J. Dagdelen, L. Weston, A. Dunn, Z. Rong, O. Kononova, K.A. Persson, G. Ceder, A. Jain, *Nature* 571 (2019) 95–98. <https://doi.org/10.1038/s41586-019-1335-8>
- [2] J. Radaković, K. Batalović, N. Novaković, *Journal of Cheminformatics* 18 (2026), 20 <https://doi.org/10.1186/s13321-025-01149-3>
- [3] M. Witman, G. Ek, S. Ling, J. Chames, S. Agarwal, J. Wong, M.D. Allendorf, M. Sahlberg, V. Stavila, *Chem. Materials* 33 (2021) 4067–4076 <https://doi.org/10.1021/acs.chemmater.1c00647>
- [4] X. Liu, J. Zhang, Z. Pei, *Progress in Materials Science* 131 (2023) 101018. <https://doi.org/10.1016/j.pmatsci.2022.101018>



Dr. Jana Radakovic is a Principle Research Fellow at the Institute of Nuclear Sciences "Vinca", University of Belgrade, specializing in computational materials science and machine learning applications. Her research applies materials informatics methods to the design of advanced materials for renewable energy technologies, with a particular focus on hydrogen storage. Current work focuses on combining literature-derived representations with classical descriptors to support data-driven screening of materials for hydrogen storage applications.

Presenting author: Jana Radakovic **e-mail:** janar@vin.bg.ac.rs