Cold-Starting Active Learning Loops Using Multiple Data Modalities

Doaa Mohamed[1], Felix Thelen[2], Rico Zehl[2], Alfred Ludwig[2], Markus Stricker[1]

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

[2]Institute for Materials, Ruhr University Bochum, Bochum, Germany

Doaa.Mohamed@ruhr-uni-bochum.de

Discovering new materials in compositionally complex materials is time-consuming and costly due to the large number of measurements required to explore vast composition-property spaces. Active learning offers an acceleration strategy by reducing the number of required measurements while maintaining high surrogate model accuracy and low uncertainty. A key challenge, however, is the cold-start problem: selecting informative initial points in the absence of labeled data [1].

We systematically evaluate multiple cold-start initialization strategies based on different inexpensive ("cheap") data modalities, as well as their multimodal combination, to initialize the active learning loop for materials characterization. These modalities are numerical, visual, and textual information that provide complementary priors for cold-start initialization [2]. The numerical modality is derived from elemental compositions measured by EDX, which are available prior to resistance measurements. Each measurement area is represented as a vector in composition space (e.g., atomic fractions of Ag, Au, Pd, Pt, etc.). The visual modality uses photographs of the materials libraries. Image features extracted from the RGB color space capture spatial variations in surface appearance that correlate with film thickness, microstructure, and indirectly with electrical resistance. Clustering these features enables the selection of visually distinct regions as initialization measurement areas for active learning. The textual modality provides a literature-informed prior. Word embeddings trained on scientific abstracts map elemental compositions and material properties (e.g., "resistance") into a shared semantic space. Similarity between composition and property representations reflects reported composition–property associations, guiding initialization toward chemically relevant regions. These strategies provide diverse and representative starting points, enabling rapid model convergence compared to a baseline cold start using an evenly fixed-grid initialization baseline [3]. We compare two acquisition functions, Uncertainty Sampling (US) and Self-Adjusting Weighted Expected Improvement (SAWEI), that automatically balance exploration and exploitation during iterative sampling. Active learning is stopped dynamically based on the normalized mean predictive variance of the surrogate model.

We apply our framework to predict the electrical resistance as a function of composition using composition-spread thin-film materials libraries, evaluating eight experimental libraries with varying compositional complexity. Our approach significantly reduces the number of required measurements, achieving up to an 87% reduction for individual libraries and an average reduction of 85% using a multimodal cold-start strategy compared to the fixed, evenly spaced four-point probe initialization baseline. Overall, SAWEI consistently outperforms uncertainty sampling. This work demonstrates a practical multimodal cold-start active learning framework that accelerates autonomous experimental characterization toward autonomous materials discovery.

References

[1] M. Stricker, L. Banko, N. Sarazin, N. Siemer, J. Janssen, L. Zhang, J. Neugebauer, A. Ludwig, npj Computational Materials, **2025**, 12, 2.

[2] D. Mohamed, S. G. Vázquez, B. Mardani, V. Dudarev, A. Ludwig, M. Acosta, M. Stricker, arXiv, **2026**, arXiv:2601.09359.

[3] F. Thelen, L. Banko, R. Zehl, S. Baha, A. Ludwig, *Digital Discovery*, **2023**, *2*, 1612–161.