Data Generation, Heuristics, and Machine Learning for Materials Discovery and Simulation

Janine George[1,†]

[1]BAM Berlin, Germany. 2 University of Jena, Germany

† corresponding author's email:Janine.george@bam.de

Machine learning (ML) offers powerful new strategies for accelerating the discovery and design of functional materials. In our work, we develop ML models and software frameworks for large-scale screening and advanced materials simulations, starting from robust high-throughput quantum-chemical workflows, such as those implemented in atomate2.[1,2] These automated workflows enable the creation of large, high-quality materials databases that form the foundation for data science and machine learning. In addition to experimentally known crystal structures, increasingly generative models are used to extend materials databases, which also need to be evaluated.[3]

To build predictive, scientifically grounded ML models, we use chemical bonding concepts, incorporating quantum-chemical bonding strengths and related descriptors as physically meaningful features to predict vibrational properties and heat transport.[4,5] Beyond property prediction, we address the challenge of determining which hypothetical materials are synthesizable. To this end, we introduced co-training into a positive–unlabeled (PU) learning framework, enabling ML-based classification even in the absence of true negative data—an essential step for screening synthesizable compounds.[6,7]

To advance atomistic simulations of complex materials, we further developed automated training pipelines for ML interatomic potentials that support both general-purpose and system-specific potential development, as implemented in our software autoplex.[8] This automated approach has already facilitated detailed investigations of challenging systems, including the computational exploration of amorphous arsenic.[9]

Together, these developments provide a toolbox spanning workflow automation, automated ML potential training, and ML models for materials properties and synthesis, enabling scalable, data-driven discovery and understanding of advanced materials.

References:

[1] M. K. Horton, P. Huck, R. X. Yang, J. M. Munro, S. Dwaraknath, A. M. Ganose, R. S. Kingsbury, M. Wen, J. X. Shen, T. S. Mathis, A. D. Kaplan, K. Berket, J. Riebesell, J. George, A. S. Rosen, E. W. C. Spotte-Smith, M. J. McDermott, O. A. Cohen, A. Dunn, M. C. Kuner, et al., *Nat. Mater.* **2025**, *24*, 1522–1532.

[2] A. M. Ganose, H. Sahasrabuddhe, M. Asta, K. Beck, T. Biswas, A. Bonkowski, J. Bustamante, X. Chen, Y. Chiang, D. C. Chrzan, others, *Digit. Discov.* **2025**, *4*, 1944–1973.

[3] P. Hagemann, S. Müller, J. George, P. Benner, **2025**, DOI 10.48550/arXiv.2512.09514.

[4] A. A. Naik, C. Ertural, N. Dhamrait, P. Benner, J. George, *Sci. Data* **2023**, *10*, 610.

[5] A. A. Naik, N. Dhamrait, K. Ueltzen, C. Ertural, P. Benner, G.-M. Rignanese, J. George, **2026**, DOI 10.48550/arXiv.2602.12109.

[6] S. Amariamir, J. George, P. Benner, *Digit. Discov.* **2025**, *4*, 1437–1448.

[7] J. Bustamante, A. Ghata, A. A. Naik, C. Ertural, K. Ueltzen, W. G. Zeier, J. George, *J. Mater. Chem. A* **2026**, DOI 10.1039/D5TA08709K.

[8] Y. Liu, J. D. Morrow, C. Ertural, N. L. Fragapane, J. L. A. Gardner, A. A. Naik, Y. Zhou, J. George, V. L. Deringer, *Nat. Commun.* **2025**, *16*, 7666.

[9] Y. Liu, Y. Zhou, R. Ademuwagun, L. Walterbos, J. George, S. R. Elliott, V. L. Deringer, *J. Am. Chem. Soc.* **2026**, DOI 10.1021/jacs.5c18688.