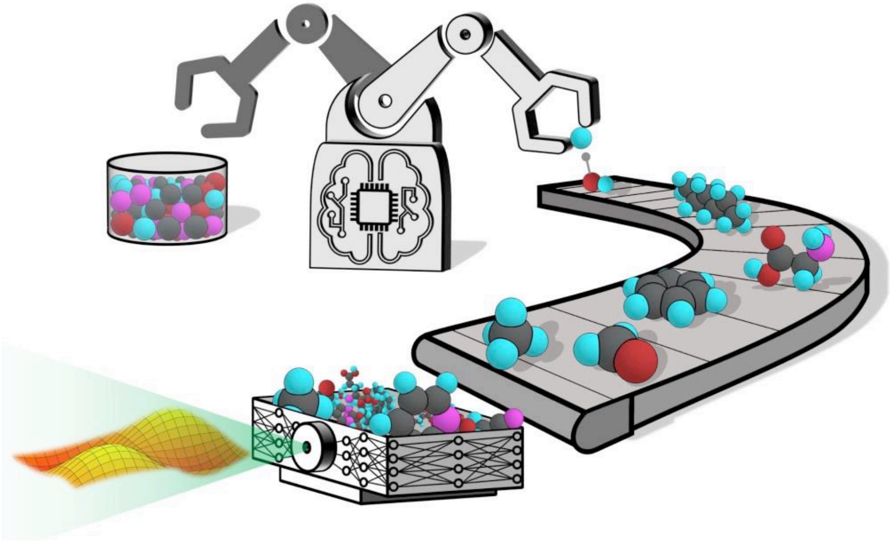
Data Generation for Machine Learning



Caption: Graphic image depicting the importance of data generation, usage and sharing for machine-learning based simulation of chemical processed and materials.

Scientific Achievement

This Review summarizes the methodologies, challenges, and opportunities that underpin critical aspects of machine-learning driven sampling and data collection.

Significance and Impact

The field of data-driven chemistry and materials science is undergoing an evolution, driven by innovations in machine learning models for predicting molecular properties and behavior. The key determinant defining reliability of these simulations is the quality of the training data.

Research Details

- Discussion of data acquisition, combination and integrity.
- Summarized publicly available databases and discuss best practices for data preparation and sharing.
- Outlined active learning principles for autonomous data generation.

Kulichenko, M.; Nebgen, B.; Lubbers, N.; Smith, J. S.; Barros, K.; Allen, A.; A. Habib, Shinkle, E.; Fedik, N.; Li, Y. W.; Messerly, R. A.; Tretiak, S. “Data Generation for Machine Learning Interatomic Potentials and Beyond.” *Chemical Reviews*, 124, 13681–13714 (2024)

Work was performed, in part, at the Center for Integrated Nanotechnologies.

U.S. DEPARTMENT OF ENERGY | Office of Science

CINT Los Alamos NATIONAL LABORATORY NVIDIA

https://science.osti.gov/