Workflows for Artificial Intelligence

Jörg Behler[1,2], Gábor Csányi[3], ***Lucas Foppa**[4], Kisung Kang[4], Marcel F. Langer[5], Johannes T. Margraf[6], ***Akhil S. Nair**[4], Thomas A. R. Purcell[7], Patrick Rinke[8,9,10,11], Matthias Scheffler[4], Alexandre Tkatchenko[12,13], Milica Todorović[14], Oliver T. Unke[15], and Yi Yao[16]

[1]Lehrstuhl für Theoretische Chemie II, Ruhr-Universität Bochum, D-44780 Bochum, Germany
[2]Research Center Chemical Sciences and Sustainability, Research Alliance Ruhr, D-44780 Bochum, Germany
[3]Department of Engineering, University of Cambridge, Cambridge CB2 1PZ, U.K.
[4]The NOMAD Laboratory at the Fritz Haber Institute of the Max Planck Society, 14195 Berlin, Germany
[5]Laboratory of Computational Science and Modeling, Institut des Matériaux, École Polytechnique Fédérale de Lausanne, 1015 Lausanne, Switzerland
[6]Bavarian Center for Battery Technology (BayBatt), University of Bayreuth, Weiherstraße 26, 95448, Bayreuth, Germany
[7]Department of Chemistry and Biochemistry, University of Arizona, Tucson, AZ 85721, USA
[8]Department of Applied Physics, Aalto University, P.O. Box 11000, FI-00076 Aalto, Finland
[9]Physics Department, TUM School of Natural Sciences, Technical University of Munich, Garching, Germany
[10]Atomistic Modelling Center, Munich Data Science Institute, Technical University of Munich, Garching, Germany
[11]Munich Center for Machine Learning (MCML)
[12]Department of Physics and Materials Science, University of Luxembourg, L-1511 Luxembourg, Luxembourg
[13]Institute for Advanced Studies, University of Luxembourg, L-1511 Luxembourg, Luxembourg
[14]Department of Mechanical and Materials Engineering, University of Turku, 20014 Turku, Finland
[15]Google DeepMind, Berlin, Germany
[16]Molecular Simulations from First Principles e.V., D-14195 Berlin, Germany

Summary

The efficiency and reliability of artificial-intelligence (AI)-driven physics, chemistry, biophysics, materials science and engineering depends on the acquisition of sufficient, high-quality data. Due to its all-electron, full potential treatment, and its scalability to larger systems without precision limitations, FHI-aims provides accurate *ab initio* data from a wide range of computer simulations, such as electronic-structure calculations and molecular dynamics. To leverage the capabilities of AI models, workflows that seamlessly integrate AI tools with FHI-aims are essential. These workflows automate the acquisition of data and their use by AI. Thus, they facilitate the iterative data exchange between AI models and simulations, allowing FHI-aims to be used as a powerful AI-guided calculation engine. Also, interpretable AI models aid in analyzing the generated data. Furthermore, AI complements *ab initio* studies as it enables to perform simulations at larger time and length scales. In turn, also the AI models must incorporate the physics required for an accurate representation of the *ab initio* data. This contribution highlights workflows developed to integrate FHI-aims with AI and future challenges.

Current Status of the Implementation

FHI-aims [1] provides various approaches to the provision and communication of data, which can then be used in workflows for the development of AI models such as machine-learning interatomic potentials (MLIPs). The interface with the Atomic Simulation Environment (ASE) [2] library provides a framework for the generation or loading of input files, calculation of properties, and storage of calculation outcomes. Interfaces with high-throughput-calculation tools such as atomate2 [3] and Aiida [4] also allow rapid acquisition of data for training AI models (see contribution 8.1). These interfaces are particularly crucial in workflows where the AI model is retrained (updated) iteratively with more data. We refer to these workflows as sequential-active-learning (SAL) workflows. These SAL workflows rely on data acquisition strategies informed by the AI model, which ensure that the new data to be collected is relevant. For instance, new data might be acquired when the prediction of the AI model is unreliable [5].

In the following, we highlight examples of workflows involving FHI-aims and AI.

SAL workflows using MLIPs such as Gaussian Approximation Potential (GAP) [6] and, more recently, MACE [7] have been designed based on training data generated by FHI-aims. Examples of this include workflows for crystal structure prediction [8], battery materials [9] or surface catalysis [10, 11]. In these applications, it proved essential to have large flexibility with respect to simulation types (including global optimization, transition state searches, molecular dynamics and enhanced sampling methods) and training set selection (e.g., via uncertainty estimation or farthest point heuristics). These requirements have led to the development of the `wfl` package, a Python toolkit for interatomic potential creation and atomistic simulation workflows that emphasizes modularity and parallelisation over sets of atomic configurations [12, 13].

We note that it is important that the chosen MLIP model is able to describe all the relevant physics, since increasing the amount of data alone does not guarantee a model representing the system correctly. Examples are long-range electrostatic interactions beyond the local cutoff radii often employed in the construction of MLIPs for condensed systems, dispersion interactions, and non-local charge transfer. The latter is crucial in many types of chemical reactions, e.g., if the charge of a molecule is altered by (de)protonation or an atomic oxidation state changes due to electron transfer. For these cases, often non-local approaches like fourth-generation MLIPs may be needed, which take the global structure of the system into account for describing electrostatics [14]. An alternative solution for small systems is to employ explicit global machine-learning force fields like GDML/BIGMDL [15, 16] or, more generally, a graph neural network architecture combined with physical models for long-range interactions, such as GEMS [17] or SO3LR [18]. Another package that has been developed linking FHI-aims and MLIPs is the GKX package [19] and the FHI-vibes framework (see contribution 6.1). By using GKX, MLIPs trained on high-fidelity data generated with FHI-aims can be used to perform GPU-accelerated MD simulations for systems with thousands of atoms over timescales of nanoseconds. Such simulations can be used, for instance, to obtain converged thermal transport coefficients [20].

Despite the growing number of applications of MLIPs, concerns about their reliability arise when they are utilized to predict properties associated with configurations or chemical species that are significantly different from those in the training set [21]. Kang et al. developed `ALMOMD` (Active-Learning Machine-Operated Molecular Dynamics [22, 23]), a Python workflow package interfacing FHI-aims with the MLIP codes `NequIP` [24] and `so3krates` [25]. `ALMOMD` is designed to effectively train MLIP through a SAL scheme with an automated framework that samples unfamiliar data, e.g., rare events, based on the uncertainty estimates of MLIP predictions (Figure 1).

While MLIPs hold great promise, many material problems are high-dimensional in nature and involve costly evaluation of an objective function. This can benefit from SAL workflows that dramatically reduce sampling, such as those involving Bayesian optimization. This algorithm builds probabilistic N-dimensional surrogate models for materials energy or property landscapes, then refines them with smart sampling. The strategic acquisition strategy of blending data exploitation with design space ex-

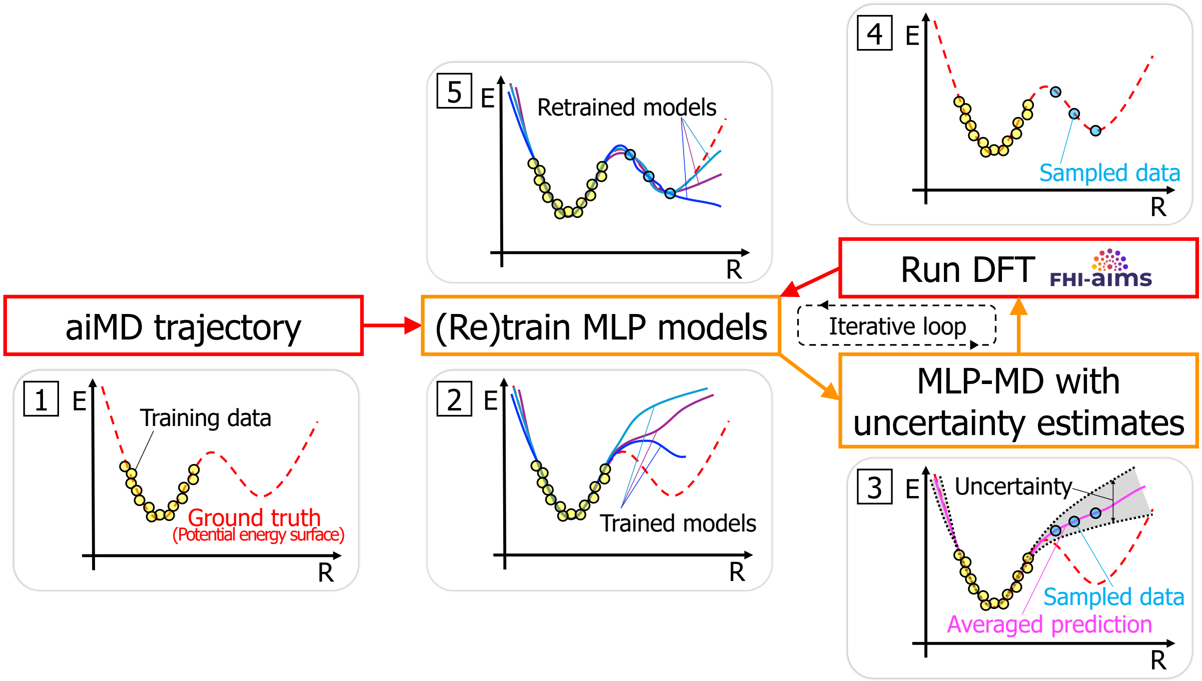

Figure 1: The overall iterative workflow of the `ALMOMD`. White boxes display indexed sequential steps for exploring the configurational space using MLIP-MD and sampling training data via uncertainty estimates.

ploration ensures fast identification of optimal solutions. Such a probabilistic algorithm is encoded into the Bayesian Optimization Structure Search (BOSS) Python tool for materials optimization [26, 27] and made interoperable with FHI-aims and ASE.

Finally, high-quality materials data are sparse, demanding data-efficient AI approaches. Moreover, interpretability, i.e., the ability to inspect the model and gain insights into its reasoning, is critical, highlighting the importance of using descriptor-based AI methods [28]. Nair et al. have developed a SAL workflow integrating the sure-independence screening and specifying operator (SISSO) approach [29, 30] with FHI-aims. SISSO is a symbolic-regression method that utilizes compressed sensing to identify analytical expressions correlated with materials' properties or functions. It is a data-efficient method and offers better interpretability compared to widely used ML approaches in materials science such as neural networks. SISSO identifies analytical expressions that contain key physicochemical parameters, from many offered ones. The developed workflow utilizes an interface of FHI-aims with the high-throughput utility Taskblaster [31] for executing multiple tasks, such as geometry optimization, band-structure calculation, etc., for a large number of materials. Such workflows achieve efficient data acquisition and they mitigate the issue of redundant data [32].

Usability and Tutorials

(around 450 words)

This section illustrates the applications of the workflows with tutorials, that researchers can adapt to their specific projects.

The `ALMOMD` framework is demonstrated in the context of atomistic simulations of strongly anharmonic materials. Incomplete MLIP training often happens due to the absence or insufficiency of data within training regions. For example, MLIPs may be unable to predict rare dynamical events, like defect creations, that are not included in training data due to their infrequency. Consequently, it leads to critical deviations in predictions for transport properties. The ALMOMD framework can actively learn these unfamiliar data missed during MLIP training and correct the potential erroneous predictions during molecular dynamics (MD) simulations. ALMOMD consists of two important steps: exploration and data-sampling. The efficient exploration of configurational space is achieved by explorative MD employing MLIPs (MLIP-MD). Uncertainty estimates serve as a warning signal indicating when MLIP-MD goes beyond its trained area, and thus, it can identify unfamiliar data that need to be sampled and retrained for MLIPs in subsequent steps. `ALMOMD` provides the user with an automated workflow environment and online tutorials [23].

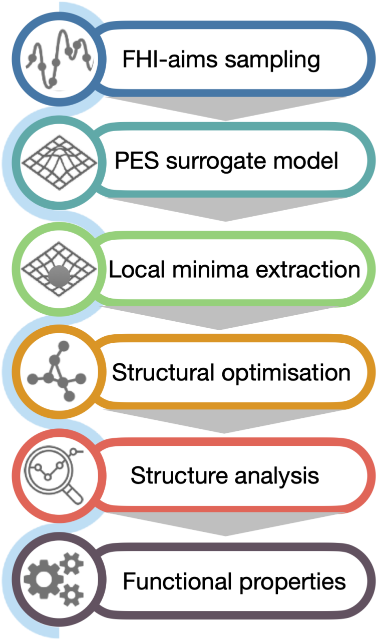

Figure 2: BOSS AI workflow: from surrogate models to optimal solutions.

BOSS was applied to study molecular conformers [33] and surface adsorbates [34, 35], thin film growth [36], solid-solid interfaces [37] and even combine multiple fidelity simulations. The computation workflow illustrated in Figure 2 relies on uncertainty-aware and interpretable surrogate model landscapes to extract optimal solution basins, from which structural optimization leads to final structures and associated functional properties. The BOSS website facilitates adaptations of this workflow to different use cases [27], with the code, manual and extensive tutorials available to the research community [38, 39].

The SISSO-based SAL workflow is applied to the efficient discovery of acid-stable oxides for water splitting reaction from a large space of candidate materials, i.e., with a reduced number of calculations compared to high-throughput screening (Figure 3). Ensembles of SISSO models are used to obtain not only mean predictions, but also estimates of the prediction uncertainties. This opened the opportunity to use SISSO as a surrogate model in the aforementioned Bayesian optimization approach. DFT calculations were carried out for these materials by leveraging an efficient implementation of hybrid functionals (see contribution 3.2). The SISSO-guided workflow enabled the identification of 13 ternary oxides as potential candidate materials for water splitting. A tutorial demonstrating the workflow can be found at ref [40]. Such a workflow reduces the risk of overlooking potentially interesting portions of the materials space that were disregarded in the initial training data and hence enables efficient materials discovery.

Future Plans and Challenges

(around 350 words)

Currently, the majority of MLIP-based methods which have been integrated into workflows to guide FHI-aims simulations, function as standalone packages. The next step would be to enable direct execution of MLIPs within the FHI-aims environment, creating a more streamlined process that integrates MLIPs into the simulation workflow without the need for separate executions. For SAL workflows, obtaining accurate and reliable uncertainty estimates is challenging, as overconfidence could lead to inefficient sampling of the materials or configuration spaces. Additionally, for systems like strongly correlated materials, workflows must be adapted to integrate advancements in beyond-DFT methods, such as GW or RPA.

Apart from the use as a calculation engine for data acquisition, AI models could also be used to accelerate FHI-aims calculations or improve their accuracy. For example, promising research directions include AI-based initial guesses for wavefunctions based on learning Kohn-Sham matrices [41, 42], the prediction of electron density in 3D space [43] or the development of novel density functionals [44, 45]. The training data for such approaches could itself be generated by FHI-aims, allowing to improve AI models

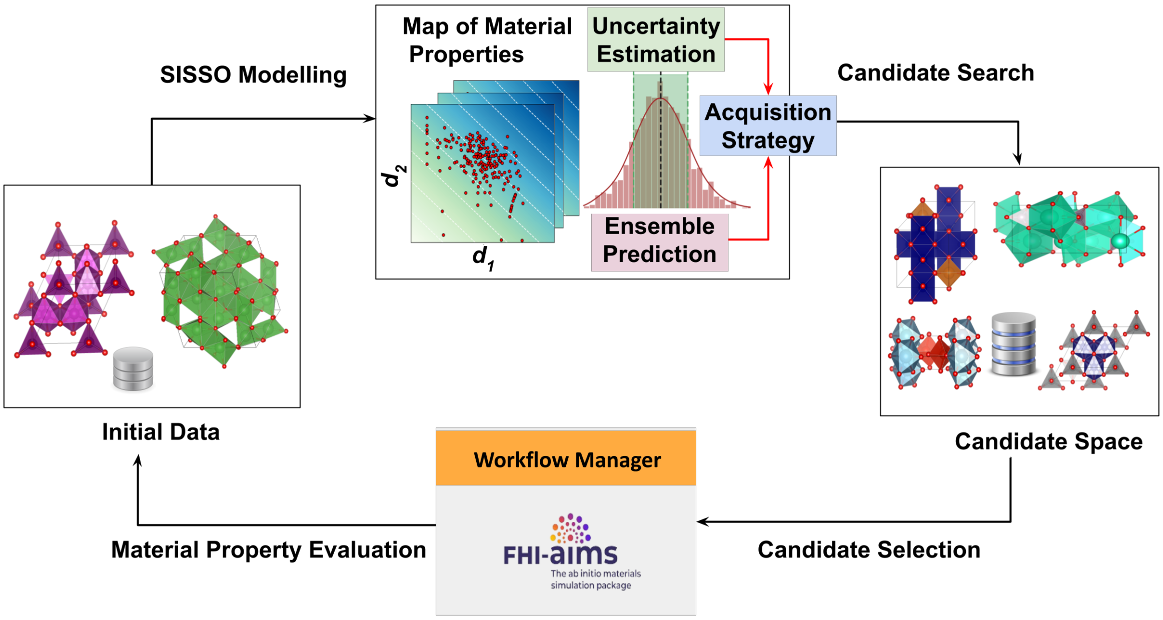

Figure 3: Schematic representation of the workflow integrating SISSO and FHI-aims. d_1 and d_2 represent the descriptors constituting a materials map where the initially available data is labeled with red circles.

in a feedback loop. Such methods and their potential applications are discussed in greater detail in the contribution 8.4.

Acknowledgements

We acknowledge the funding of the NOMAD Center of Excellence (European Union's Horizon 2020 research and innovation program, Grant Agreement No. 951786) and of the ERC Advanced Grant TEC1p (European Research Council, Grant Agreement No 740233).

References

[1] V. Blum, R. Gehrke, F. Hanke, P. Havu, V. Havu, X. Ren, K. Reuter, and M. Scheffler, "Ab initio molecular simulations with numeric atom-centered orbitals", Computer Physics Communications **180**, 2175–2196 (2009).

[2] A. H. Larsen, J. J. Mortensen, J. Blomqvist, I. E. Castelli, R. Christensen, M. Du lak, J. Friis, M. N. Groves, B. Hammer, C. Hargus, E. D. Hermes, P. C. Jennings, P. B. Jensen, J. Kermode, J. R. Kitchin, E. L. Kolsbjerg, J. Kubal, K. Kaasbjerg, S. Lysgaard, J. B. Maronsson, T. Maxson, T. Olsen, L. Pastewka, A. Peterson, C. Rostgaard, J. Schiøtz, O. Schütt, M. Strange, K. S. Thygesen, T. Vegge, L. Vilhelmsen, M. Walter, Z. Zeng, and K. W. Jacobsen, "The atomic simulation environment—a python library for working with atoms", Journal of Physics: Condensed Matter **29**, 273002 (2017).

[3] Andrei Sobolev and Uthpala Herath, *High-throughput workflows with FHI-aims and atomate2*, `https://workflows-with-atomate2-fhi-aims-club-tutorials-a0acc2efc2fba83.gitlab.io/`, [Online; accessed 26-Sep-2024], 2024.

[4] Andrei Sobolev, *Working with FHI-aims and AiiDA*, `https://fhi-aims-club.gitlab.io/tutorials/fhi-aims-with-aiida/`, [Online; accessed 26-Sep-2024], 2024.

[5] S. Bauer, P. Benner, T. Bereau, V. Blum, M. Boley, C. Carbogno, C. R. A. Catlow, G. Dehm, S. Eibl, R. Ernstorfer, et al., "Roadmap on data-centric materials science", Modelling and Simulation in Materials Science and Engineering **32**, 063301 (2024).

[6] V. L. Deringer, A. P. Bartók, N. Bernstein, D. M. Wilkins, M. Ceriotti, and G. Csányi, "Gaussian process regression for materials and molecules", Chemical Reviews **121**, 10073–10141 (2021).

[7] I. Batatia, D. P. Kovacs, G. Simm, C. Ortner, and G. Csányi, "Mace: higher order equivariant message passing neural networks for fast and accurate force fields", Advances in Neural Information Processing Systems **35**, 11423–11436 (2022).

[8] S. Wengert, G. Csányi, K. Reuter, and J. T. Margraf, "Data-efficient machine learning for molecular crystal structure prediction", Chem. Sci. **12**, 4536–4546 (2021).

[9] C. G. Staacke, H. H. Heenen, C. Scheurer, G. Csányi, K. Reuter, and J. T. Margraf, "On the role of long-range electrostatics in machine-learned interatomic potentials for complex battery materials", ACS Applied Energy Materials **4**, 12562–12569 (2021).

[10] S. Stocker, H. Jung, G. Csányi, C. F. Goldsmith, K. Reuter, and J. T. Margraf, "Estimating free energy barriers for heterogeneous catalytic reactions with machine learning potentials and umbrella integration", Journal of Chemical Theory and Computation **19**, 6796–6804 (2023).

[11] H. Jung, L. Sauerland, S. Stocker, K. Reuter, and J. T. Margraf, "Machine-learning driven global optimization of surface adsorbate geometries", npj Computational Materials **9**, 114 (2023).

[12] E. Gelžinytė, S. Wengert, T. K. Stenczel, H. H. Heenen, K. Reuter, G. Csányi, and N. Bernstein, "wfl Python toolkit for creating machine learning interatomic potentials and related atomistic simulation workflows", The Journal of Chemical Physics **159**, 124801 (2023).

[13] *Libatoms*, `https://libatoms.github.io/workflow/index.html` (visited on 2023).

[14] T. W. Ko, J. A. Finkler, S. Goedecker, and J. Behler, "General-purpose machine learning potentials capturing nonlocal charge transfer", Acc. Chem. Res. **54**, 808–817 (2021).

[15] S. Chmiela, V. Vassilev-Galindo, O. T. Unke, A. Kabylda, H. E. Sauceda, A. Tkatchenko, and K.-R. Müller, "Accurate global machine learning force fields for molecules with hundreds of atoms", Science Advances **9**, eadf0873 (2023).

[16] H. E. Sauceda, L. E. Gálvez-González, S. Chmiela, L. O. Paz-Borbón, K.-R. Müller, and A. Tkatchenko, "Bigdml—towards accurate quantum machine learning force fields for materials", Nature communications **13**, 3733 (2022).

[17] O. T. Unke, M. Stöhr, S. Ganscha, T. Unterthiner, H. Maennel, S. Kashubin, D. Ahlin, M. Gastegger, L. Medrano Sandonas, J. T. Berryman, et al., "Biomolecular dynamics with machine-learned quantum-mechanical force fields trained on diverse chemical fragments", Science Advances **10**, eadn4397 (2024).

[18] J. T. Frank, O. T. Unke, K.-R. Müller, and S. Chmiela, "A euclidean transformer for fast and stable machine learned force fields", Nature Communications **15**, 6539 (2024).

[19] *Gkx*, `https://github.com/sirmarcel/gkx` (visited on 2024).

[20] M. F. Langer, F. Knoop, C. Carbogno, M. Scheffler, and M. Rupp, "Heat flux for semilocal machine-learning potentials", Phys. Rev. B **108**, L100302 (2023).

[21] L. Kahle and F. Zipoli, "Quality of uncertainty estimates from neural network potential ensembles", Physical Review E **105**, 015311 (2022).

[22] K. Kang, T. A. R. Purcell, C. Carbogno, and M. Scheffler, *Accelerating the training and improving the reliability of machine-learned interatomic potentials for strongly anharmonic materials through active learning*, 2024.

[23] Available at: `https://gitlab.com/FHI-aims-club/ALmoMD`.

[24] S. Batzner, A. Musaelian, L. Sun, M. Geiger, J. P. Mailoa, M. Kornbluth, N. Molinari, T. E. Smidt, and B. Kozinsky, "E(3)-equivariant graph neural networks for data-efficient and accurate interatomic potentials", Nature Communications **13**, 2453 (2022).

[25] T. Frank, O. Unke, and K.-R. Müller, "So3krates: equivariant attention for interactions on arbitrary length-scales in molecular systems", in Advances in neural information processing systems, Vol. 35, edited by S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (2022), pp. 29400–29413.

[26] M. Todorović, M. U. Gutmann, J. Corander, and P. Rinke, "Bayesian inference of atomistic structure in functional materials", npj Computational Materials **5**, 10.1038/s41524-019-0175-2 (2019).

[27] *Boss public website*, https://sites.utu.fi/boss (visited on 2024).

[28] F. Oviedo, J. L. Ferres, T. Buonassisi, and K. T. Butler, "Interpretable and explainable machine learning for materials science and chemistry", Accounts of Materials Research **3**, 597–607 (2022).

[29] R. Ouyang, S. Curtarolo, E. Ahmetcik, M. Scheffler, and L. M. Ghiringhelli, "Sisso: a compressed-sensing method for identifying the best low-dimensional descriptor in an immensity of offered candidates", Physical Review Materials **2**, 083802 (2018).

[30] T. A. R. Purcell, M. Scheffler, and L. M. Ghiringhelli, "Recent advances in the SISSO method and their implementation in the SISSO++ code", The Journal of Chemical Physics **159**, 114110 (2023).

[31] *Taskblaster*, https://gitlab.com/taskblaster/taskblaster (visited on 04/06/2024).

[32] K. Li, D. Persaud, K. Choudhary, B. DeCost, M. Greenwood, and J. Hattrick-Simpers, "Exploiting redundancy in large materials datasets for efficient machine learning with less data", Nature Communications **14**, 7283 (2023).

[33] L. Fang, E. Makkonen, M. Todorović, P. Rinke, and X. Chen, "Efficient Amino Acid Conformer Search with Bayesian Optimization", Journal of Chemical Theory and Computation **17**, 1955–1966 (2021).

[34] J. Järvi, P. Rinke, and M. Todorović, "Detecting stable adsorbates of (1S)-camphor on Cu(111) with Bayesian optimization", Beilstein Journal of Nanotechnology **11**, 1577–1589 (2020).

[35] J. Järvi, B. Alldritt, O. Krejčí, M. Todorović, P. Liljeroth, and P. Rinke, "Integrating Bayesian Inference with Scanning Probe Experiments for Robust Identification of Surface Adsorbate Configurations", Advanced Functional Materials **31**, 10.1002/adfm.202010853 (2021).

[36] A. T. Egger, L. Hörmann, A. Jeindl, M. Scherbela, V. Obersteiner, M. Todorović, P. Rinke, and O. T. Hofmann, "Charge Transfer into Organic Thin Films: A Deeper Insight through Machine-Learning-Assisted Structure Search", Advanced Science **7**, 10.1002/advs.202000992 (2020).

[37] A. Fangnon, M. Dvorak, V. Havu, M. Todorović, J. Li, and P. Rinke, "Protective Coating Interfaces for Perovskite Solar Cell Materials: A First-Principles Study", ACS Applied Materials & Interfaces **14**, 12758–12765 (2022).

[38] *Boss code*, https://gitlab.com/cest-group/boss (visited on 2024).

[39] *Boss documentation*, https://cest-group.gitlab.io/boss (visited on 2024).

[40] *Sequential learning with sisso*, https://gitlab.com/FHI-aims-club/tutorials/tutorial-sl-sisso.

[41] K. T. Schütt, M. Gastegger, A. Tkatchenko, K.-R. Müller, and R. J. Maurer, "Unifying machine learning and quantum chemistry with a deep neural network for molecular wavefunctions", Nature communications **10**, 5024 (2019).

[42] O. Unke, M. Bogojeski, M. Gastegger, M. Geiger, T. Smidt, and K.-R. Müller, "Se (3)-equivariant prediction of molecular wavefunctions and electronic densities", Advances in Neural Information Processing Systems **34**, 14434–14447 (2021).

[43] X. Fu, A. Rosen, K. Bystrom, R. Wang, A. Musaelian, B. Kozinsky, T. Smidt, and T. Jaakkola, "A recipe for charge density prediction", arXiv preprint arXiv:2405.19276 (2024).

[44] J. C. Snyder, M. Rupp, K. Hansen, K.-R. Müller, and K. Burke, “Finding density functionals with machine learning”, Physical review letters **108**, 253002 (2012).

[45] J. Kirkpatrick, B. McMorrow, D. H. Turban, A. L. Gaunt, J. S. Spencer, A. G. Matthews, A. Obika, L. Thiry, M. Fortunato, D. Pfau, et al., “Pushing the frontiers of density functionals by solving the fractional electron problem”, Science **374**, 1385–1389 (2021).