JPhys Materials

PURPOSE-LED PUBLISHING™

ACCEPTED MANUSCRIPT • OPEN ACCESS

Artificial intelligence for advanced functional materials: Exploring current and future directions

To cite this article before publication: Cristiano Malica *et al* 2025 *J. Phys. Mater.* in press https://doi.org/10.1088/2515-7639/adc29d

Artificial Intelligence for Advanced Functional Materials: Exploring Current and Future Directions

Author list

Cristiano Malica (1,*), Kostya Novoselov (2,3), Amanda S Barnard (4), Sergei V. Kalinin (5,6), Steven R. Spurgeon (7,8), Karsten Reuter (9), Maite Alducin (10,11), Volker L. Deringer (12), Gabor Csanyi (13), Nicola Marzari (1,14), Shirong Huang (15), Gianaurelio Cuniberti (15), Qiushi Deng (16), Pablo Ordejón (17), Ivan Cole (16), Kamal Choudhary (18), Kedar Hippalgaonkar (19,20), Ruiming Zhu (19,20), O. Anatole von Lilienfeld (21,22,23), Mohamed Hibat-Allah(24,22), Juan Carrasquilla (25), Giulia Cisotto (26), Alberto Zancanaro (27), Wolfgang Wenzel (28), Andrea C. Ferrari (29), Andrey Ustyuzhanin (30, 2), Stephan Roche (31, 32,*).

*Corresponding authors <cmalica@uni-bremen.de>; <stephan.roche@icn2.cat>

(1) U Bremen Excellence Chair, Bremen Center for Computational Materials Science, and MAPEX Center for Materials and Processes, University of Bremen, D-28359 Bremen, Germany
(2) Institute for Functional Intelligent Materials, National University of Singapore, Singapore 117544, Singapore
(3) Department of Materials Science and Engineering, National University of Singapore, Singapore 117575, Singapore
(4) Computational Science, School of Computing, Australian National University, Australia
(5) University of Tennessee, Knoxville
(6) Pacific Northwest National Laboratory
(7) National Renewable Energy Laboratory
(8) University of Colorado, Boulder
(9) Fritz-Haber-Institut der Max-Planck-Gesellschaft, Berlin, Germany
(10) Centro de Física de Materiales (CFM/MPC), Donostia-San Sebastián, Spain
(11) Donostia International Physics Center, Donostia-San Sebastián, Spain
(12) Inorganic Chemistry Laboratory, Department of Chemistry, University of Oxford, UK
(13) University of Cambridge, Engineering Laboratory
(14) Theory and Simulation of Materials (THEOS), and National Centre for Computational Design and Discovery of Novel Materials (MARVEL), École Polytechnique Fédérale de Lausanne (EPFL), CH-1015 Lausanne, Switzerland
(15) Institute for Materials Science and Max Bergmann Center for Biomaterials TUD Dresden University of Technology Dresden, Germany
(16) School of Engineering, RMIT University, Melbourne VIC 3000
(17) Catalan Institute of Nanoscience and Nanotechnology ICN2 (CSIC and BIST), Campus UAB, Bellaterra, 08193, Spain
(18) Materials Science and Engineering Division, National Institute of Standards and Technology, Gaithersburg, MD 20899, USA
(19) School of Materials Science and Engineering, Nanyang Technological University, Singapore 639798, Singapore
(20) Institute of Materials Research and Engineering, Agency for Science, Technology and Research (A*STAR), Singapore 138634, Singapore.
(21) Departments of Chemistry, Materials Science & Engineering, Physics, and the Acceleration Consortium, University of Toronto, Canada
(22) Vector Institute, MaRS Centre, Toronto, Ontario, M5G 1M1, Canada
(23) ML Group, Technische Universität Berlin and Institute for the Foundations of Learning and Data, 10587 Berlin, Germany

(24) Department of Applied Mathematics, University of Waterloo, Waterloo, ON N2L 3G1, Canada
(25) Institute for Theoretical Physics, ETH Zürich, 8093, Switzerland
(26) Department of Mathematics, Informatics and Geosciences, University of Trieste, Italy
(27) Department of Information Engineering, University of Padova, Italy
(28) Karlsruhe Institute of Technology, Hermann-von-Helmholtz-Platz 1, Eggenstein-Leopoldshafen, Germany
(29) Cambridge Graphene Centre, University of Cambridge, Cambridge, CB3 0FA UK
(30) Constructor University, Bremen, Campus Ring 1, 28759, Germany
(31) Catalan Institute of Nanoscience and Nanotechnology (ICN2), CSIC and BIST, Campus UAB, 08193, Bellaterra, Barcelona, Spain
(32) ICREA – Institució Catalana de Recerca i Estudis Avançats, 08010, Barcelona, Spain

Abstract

This perspective addresses the topic of harnessing the tools of Artificial Intelligence (AI) for boosting innovation in functional materials design and engineering as well as discovering new materials for targeted applications in biomedicine, composites, nanoelectronics or quantum technologies. It gives a current view of experts in the field, insisting on challenges and opportunities provided by the development of large materials databases, novel schemes for implementing AI into materials production and characterization as well as progress in the quest of simulating physical and chemical properties of realistic atomic models reaching the trillion atoms scale and with near *ab initio* accuracy.

Introduction

Innovative material design and engineering

The design and engineering of innovative advanced materials is facing a variety of challenges in today´s industries, including the access to proper design strategies for reaching upper performances of materials for targeted applications, the discovery of alternatives and the search for more functional intelligent materials which can help solving health, energy or environmental issues. Such new functionalities often require complex material structure, such as complex alloys, composites, heterostructures, etc. Traditional methods of material modelling cannot cope with such demands. In this context the use of Artificial Intelligence (AI) tools has become cornerstone for boosting innovation strategies and ensuring sustainability and safe-by-design approaches.

The development and availability of material databases is a fundamental part for further training AI models (machine learning and so on) able to cope with diversity and complexity and extract hidden information which could ultimately offer further intelligent guidance in optimisation and also materials (property) discovery. However, the multiplicity of databases also calls for efforts in improving universality in development languages, interoperability as well as integrated workflows which can connect information concerning the structure to the end physical or chemical properties of materials of concern. More, the needs for more and more predictive modelling and capability of simulation tools to cope with systems reaching the trillion atoms-scale limit (while keeping a near *ab initio* accuracy) presents grand challenges and demands for novel workflows to be developed and more synergies between academic research and industrial developments. To this end, property and functions-oriented databases are required, especially when aiming at solving the inverse problem of finding the material with predetermined properties.

On the other side, Intelligent materials are defined as structural materials with advanced functionalities and can be classified as structure-mimetic (mimicking the structure of organisms) and function-mimetic (mimicking the function of organisms). Intelligent materials target self-sensing of the material during its use (e.g., damage, loads, shape, temperature, pressure, etc.) and/or target adaptive actuation (e.g., changing deformation, colour, shape, inner stresses, stiffness, temperature, etc.) which depends on the biological or environmental conditions (humidity, pH, temperature, etc). The quest for more innovative intelligent materials depends on the capability of AI tools to provide proper booster in benchmarking, fast and precise analysis and extrapolation for materials design.

As a result, the synergy between the activities of computer scientists, material engineers and AI developers with the experimental activities and elaboration of novel types of functional intelligent materials has become key for advanced development in innovative materials design and engineering.

In this perspective, we provide snapshots about efforts made in a variety of different fields and visions of international experts, searching for the same common objective, that is the

deployment of AI tools for an accelerated development of materials design and deeper access to hidden dimensions of materials growth, structure-property correlations and reverse engineering strategies. Such a vast field of research calls for structuring the exploding amount of information and also for implementing chains of tools able to communicate information and extract essential parameters that can be ultimately accessible to the largest possible public. In that perspective, international events such as AI4AM (www.ai4am.net) are enabling platforms to gather communities, facilitate networking, roadmapping and ultimately enhance our knowledge and methodologies.

1. Higher-Order Pattern Recognition for Materials Informatics using Explainable Artificial Intelligence

Explainable artificial intelligence (XAI) is an emerging field in computer science based in statistics that can augment materials informatics workflows. XAI can be used as a forensic analysis technique to understand the consequences of data, model, and application decisions, or as a model refinement method capable of distinguishing important information **[1,2].** This approach is often used to explain the how the structural characteristics of materials (features) contribute to a target property prediction using tools such as feature importance rankings that highlight useful or nuisance variables. However, an alternative approach is to apply similar methods to the instance space and identify influential or unproductive data instances (materials). Data sets contain a range of special cases such as outliers (unusual types), archetypes (pure types), stereotypes (those assumed to be representative) and prototypes (those that actually are representative), and groups of data instances (clusters) that are similar in the high dimensional feature space. The amount of influence these special types of data instances have on a pattern, cluster or prediction is rarely explored or quantified, but they can also have a profound effect on model architecture and predictive ability.

Recent work has shown that is it possible to decompose the residuals of machine learning loss functions to better understand how individual materials contribute to model predictions **[3].** This has been used to explain how including certain materials in a data set can improve the ability to accurately predict the properties of others **[4].** This research has now been expanded to explain unsupervised patterns in data and identify special subsets of materials worthy of detailed consideration **[5].** The first step is to represent materials using Shapley values, which are a solution in cooperative game theory where each game is assigned a unique distribution of a total surplus generated by the coalition of all players **[1]**. A popular tool for studying cost-sharing, market analytics and voting, in materials informatics the game is usually the model, and the players are the materials. By testing the impact of removing individual instances or features, and aggregating across the feature space or instance space, respectively, Shapley values quantify how much the inclusion or exclusion of a particular material (or a structural feature) affects the result. The second step is to transform the data, represented by its Shapley values, in different ways to reveal hidden groups or patterns. This two-step process aids the data analysis process, and acts as a precursor to the residual decomposition; simultaneously finding influential materials in the data set and quantifying how they are impacting the prediction of other materials.

The novelty in this new model-agnostic approach is that the cooperative game is the underlying data distribution, not a model, which opens up the opportunity for explainable unsupervised learning. This enables researchers to better understand how a machine learning methods use the latent information captured in the data, informing better decisions about what kind of materials to make or simulate, what kind of characterisation or analysis to perform, and how these choices impact the outcome.

2. Machine Learning for autonomous microscopy: from physics discovery to atomic fabrication

Electron and scanning probe microscopies are now one of the foundational methods for characterization of structure and functional properties of matter on the nanometer and atomic scales. Scanning probe microscopy (SPM) enables rapid characterization of surface topography and mechanical, magnetic, ferroelectric, and electrochemical properties. Electron microscopy now provides comprehensive probe of structure, chemical composition, and vibrational properties at nanometer and atomic scales.

For most domain areas, microscopies traditionally represent downstream characterization methods in materials discovery cycle yielding the qualitative data. Recent progress in quantitative SPMs and scanning transmission electron microscopy (STEM) is challenging this paradigm, delivering large volumes of quantitative structural and high-veracity property data. However, the sheer volume of data has necessitated very complex analyses, minimizing the impact. The recent progress in machine learning (ML) and rapid data analytics for postacquisition analyses and particularly active learning methods that can be operationalized on active microscopes offer to change this paradigm **[6]**. On the data analytic side, ML provides the flexibility and speed necessary to analyze large volumes of multidimensional imaging and spectroscopy data for building low-dimensional representations and, in many, cases extraction of relevant materials parameters.

A fundamentally new spectrum of opportunities emerges in the context of active learning, where ML based workflows not only inform human-based decision making, but directly return control commands to the instrument. Operationalized on the SPM and STEM machines, these methods can be used for rapid mapping of the structure-property relationships. This knowledge can further be used for the discovery of generative physical models such as microstructure evolution equations, free energy functionals and Hamiltonians, and learning processing mechanisms. By combining zero-shot **[9]** and predictive **[10]** ML models with *insitu* particle beam, heating, or other processing, it is possible to learn materials responses and impart desired metastable states. These models are especially well suited to discovery scenarios, where they can reveal latent features to scientists, informing synthesis or degradation mechanisms.

These approaches create new opportunities for materials discovery. The last 20 years have seen exponential growth of the theoretical predictive capability for crystalline materials and small molecules. The last 5 years have seen the exponential growth in the capability to accelerate materials synthesis via laboratory robotics and microfluidic synthesis. However, the

lesson of the past two decades is that scaling computation or synthesis individually by many orders of magnitude is insufficient to expedite materials discovery. Rather, the key is accelerating the feedback loop between theory and hypothesis making, experiment planning, synthesis, and characterization with subsequent update of theoretical models. Currently, characterization is the bottleneck – while synthesis can be scaled to 1000s compositions per day, the sequential structural, functional, and chemical probing outside of fast optical/photoluminescent methods still require hours and days. Closing these characterization loops requires scaling down the probing volume and reducing measurement times, tasks ideally matched to microscopy capabilities. Here, microscopy offers the natural tool for exploration of multidimensional composition and processing spaces via strong (i.e. matching target macroscopic functionalities) and weak proxies **[7]**. There is also an opportunity to leverage ML-based adaptive sampling and intelligently select modalities based on uncertainty metrics, shortcutting the time to discovery.

A fundamentally new space of opportunities for materials discovery emerges based on controlled interventions in microscopy. In SPM, these include local polarization switching and electrochemical reactions that can now be studied at the time- and length scale well outside of conventional characterization methods, but very close to the intrinsic length scales of these phenomena. For electron microscopy, unique opportunities are the result of the electron beam's power to break local chemical bonds, enabling controlled fabrication of atomic defects **[10]**, beam controlled atomic motion, and building homo-and heteroatomic artificial molecules atom by atom **[8]**. The rapid exploration of materials synthesis and degradation pathways at spatial, chemical, and temporal scales commensurate with fundamental physical interactions is now more viable than ever before.

Incorporation of ML methods both in real time and post-acquisition data analysis offers the compelling case to greatly increase the efficiency of instrument utilization by orders of magnitude and close the materials characterization gap, ushering the new era of materials and physics discovery and atomic fabrication.

3. Beyond Crystallinity and Throughput: AI for Working Interfaces in Energy Conversion Technologies

The urgency with which mankind needs to accomplish the transition to a sustainable energy economy dictates a drastic acceleration of established research and development cycles toward ever improved energy conversion devices like solar cells, catalysts, electrolyzers or batteries. With respect to materials discovery much prospect to this end is seen in datacentric approaches, which harness the powerful algorithms of machine learning (ML) or artificial intelligence (AI). In many areas of materials science, corresponding techniques ranging from high-throughput screening to inverse design are already most successfully employed to search the vast materials spaces for promising candidates at unprecedented efficiency **[11]**. The discovery is thereby often conducted entirely *in silico*, exploiting the predictive quality of first-principles computational data.

Unfortunately, such developments are at present still largely stalled for the energy conversion context as the functionality of corresponding devices is generally limited by interfacial problems and interfacial data is much more difficult to come by than the bulk data that suffices for many other application areas. This relates to the involved (experimental or computational) costs for the generation of such data, but even more so to the lack of bestpractice protocols to do this reliably and reproducibly. A crucial component here is the strong structural, compositional and morphological evolution that the interfaces in energy conversion devices undergo *operando* **[12]**. These working surfaces or interfaces are thus anything but simple truncations or ideal junctions of known bulk materials, respectively. Instead, they extend over a finite width, and exhibit novel purely interface-stabilized phases with often a low degree of crystalline order. For the experimental data generation, this *operando* evolution dictates not only stringent protocols for the initial synthesis, but a seamless and exhaustive documentation of the entire history of environmental operation conditions to which the interfaces were subject to. As this is rarely reached, data is not comparable and interoperable, preventing a community-wide build-up of large-scale data bases. At the same time, the *operando* evolution also precludes the generation of pertinent first-principles data. There are essentially no established *operando*-aware descriptors, and even if there were, there are in general no established structural models for working interfaces that could be used to compute them.

In this situation, there are two major strands in which AI and ML is presently employed toward an accelerated discovery. On the computational side, data-centric approaches are used to gain a deeper mechanistic understanding into working interfaces, with the long-term goal to use this insight to formulate *operando*-aware descriptors that could then be used for an efficient exploration of materials spaces **[13]**. A dominant development to this end is ML surrogate models, and there in particular ML interatomic potentials (MLIPs), which allow to generate first-principles quality data at orders of magnitude reduced computational costs. Appropriate for the data-scarce regime, the MLIP training is thereby done in agile active learning loops, with automated workflows being developed that ideally fully interlace this with the actual simulations to ensure consistent reliability **[14]**. In cutting-edge studies, the unprecedented capabilities are presently used to conduct the simulations in much larger simulation cells (therefore also allowing to address disorder) or perform significantly increased and therewith powerful samplings.

On the experimental side, AI and ML is increasingly employed to reach a deeper analysis of (*operando*) characterization data, either to also reach an improved mechanistic understanding or to identify structure and correlations in the data that would enable improved workflows (proxy experiments, multi-fidelity experiments, …). Notably, AI and ML is employed within emerging self-driving laboratories (SDLs). Here they complement lab automation and robotics to reach higher throughputs, but foremost they take over the experiment planning. SDLs operate in active-learning loops, in which data from executed experiments is fed back into the ML model to refine it and design subsequent experiments. Current methodological frontiers in employed Bayesian optimization or adaptive Design of Experiment approaches concern significant or varying noise levels (e.g. in case of multi-fidelity measurements), the design of larger numbers of data points (to meet batch-type operation in increasingly parallelized

workflows), or agility to either autonomously adapt the shape and dimensions of the search spaces across loops or react to corresponding changes imposed by human scientists **[15]**.

4. Accessing photoinduced reaction dynamics on surfaces with neural networks

Laser-induced reactions at surfaces are particularly interesting because this kind of excitation mechanism can increase significantly the reaction probability with respect to ordinary thermal activation and, importantly, even open new reaction channels. Still, despite the impressive technical advances, experiments alone cannot fully determine with atomistic space and time resolution all the elementary steps involved in the reaction as well as the properties determining each of these steps. It is at this point that molecular-dynamics simulations become crucial to dissect the reaction dynamics.

Modelling the ultrafast photo-induced dynamics and reactivity of adsorbates on metals requires including the effect of the laser-excited electrons and also the effect of the concomitantly excited surface lattice. All these features can be effectively achieved by solving Langevin equations of motion, in which the coupling of each adsorbate and surface atom nuclei to the excited electrons is modelled in terms of electronic friction and stochastic forces that depend on the time-dependent electronic temperature that characterizes the excited Fermi-Dirac distribution, while the rest of interactions are described with the adiabatic potential energy surface (PES) that must account for all the system degrees of freedom. In spite of the apparent simplicity of the model, such simulations are highly demanding. Low energy molecules/atoms are particularly sensitive to energy differences of tens of meV that they experience in the proximity of a solid surface. And this sensitivity is even amplified when measuring, for instance, photoinduced desorption and reactivity probabilities and final-state distributions of the scattered gas species, such as, kinetic energies, scattering angles, and rovibrational quantum state distributions to cite some. Thus, any reliable description of gas/surface interactions requires the knowledge of the corresponding accurate first-principles multidimensional PES. First-principles molecular dynamics with electronic friction and thermostats (T_e,T_l)-AIMDEF, which calculate on-the-fly the adiabatic forces with density-functional theory (DFT) **[16,17]**, do enable such a complex modelling **[18]**, but, unfortunately, these simulations come with a very large computational expense that severely limits any statistical analysis of the reaction and, also, it restricts the simulation time to just a few picoseconds that might be insufficient to guarantee well-converged reaction yields.

In the last years, the use of neural network (NN)-generated multidimensional PESs and, in particular, the use of atomistic neural networks (AtNNs), is becoming the accurate alternative to first-principles molecular dynamics studies of diverse gas-surface reactions **[19, 20]**. Aside AtNN methods, the newer message passing neural network potentials, which are also discussed in forthcoming sections, are certainly promising in terms of accuracy and efficiency when using the capabilities of GPUs. However, it must be emphasized that the requirements imposed to a NN-PES capable of describing photoinduced reactions are even more demanding than those required in usual elementary gas-surface processes. A reliable NN-PES must be able to model large and out-of-phase movements of multiple and different adsorbates and also surface atoms and it must describe accurately the very distinct and changing adsorbate

coverages that may exist during the photoinduced dynamics because of desorption events. This means that it is necessary to assure a precise description of both adsorbate-substrate and interadsorbate interactions under very different and changing conditions, including local variations in the number of neighbor adsorbates and strong lattice distortions, since the lattice temperature may vary rapidly in the range of tens to thousands Kelvin.

The AtNN-like embedded atom neural network (EANN) method, which uses descriptors inspired in the embedded-atom electron density, demonstrates to be impressively accurate and flexible to account for all these requirements **[21]**. EANN PESs allowed us to reproduce and understand the experimental strong coverage dependence of CO phodesorption in Pd(111) **[22]**, the large branching ratio between CO photo-desorption and CO photo-oxidation in Ru(0001) **[23]**, and reveal the dynamical nature of the CO physisorption well that so far was only found in XPS experiments. But there are additional open questions that we can now think in treating by exploiting NN capabilities. Besides the general challenges faced in gas/surface dynamics **[19, 20]**, specific challenges for photoinduced surface chemistry are related to performing nonadiabatic dynamics, either by advancing in orbital-based electronic friction coefficients adapted to the highly dynamic surface or even more challenging by developing excited state NN-PESs that could contribute to clarify the role on the initial nonthermal distribution of excited states.

5. Data-driven advances in modelling and understanding amorphous materials

Machine learning has transformed atomic-scale materials modelling: rather than building simplified models of reality, we can now describe “the real thing” in increasingly accurate simulations **[24]**. Machine-learned interatomic potentials (MLIPs) are trained to reproduce quantum-mechanical energy and force data for this very purpose. In the domain of inorganic materials, MLIPs are typically based on density-functional theory (DFT) ground-truth data; once they have been fitted and properly validated, they therefore enable very-large-scale molecular-dynamics (MD) simulations of bulk and nanostructured materials, all while retaining DFT-like accuracy. MLIPs have evolved from specialised tools to increasingly widely available (and visibly popular) simulation methods, and their development has been documented in numerous review articles **[19, 25]**. The architectures used to train MLIPs have been advanced over many years and, thanks to these efforts, have now reached impressive accuracy. There are still many future research directions: among them are improved strategies for dataset construction, and for MLIPs that can be distilled for downstream tasks **[26]**.

Looking from the development of MLIPs onwards to their (current and expected) impact on materials chemistry research, MLIPs are particularly promising tools in the area of amorphous materials—non-crystalline solids, whose complex atomic structures and structure–property relationships are now increasingly exploited for practical applications. Indeed, amorphous materials are of growing interest for energy storage, computing, catalysis, and many other fields (see Ref. **[27]** and references therein). Accordingly, amorphous materials are a frontier research challenge in computational materials design, and MLIPs are well placed to help to address this challenge **[27]**.

A recent ML-driven study of graphene oxide (GO) exemplifies several aspects related to MLIPs and their applications to disordered and amorphous materials **[28]**. Formally, GO is a sheet of graphene modified by the presence of various functional groups (say, hydroxyl, carbonyl, and so on). In laboratory experiments, this modification is achieved using chemical reactions; in simulations, one can now quickly construct atomistic structural models over a wide-ranging parameter space of compositions and functional groups, and yet only the subsequent comparison with experiment will ultimately validate a given structural model. The study in Ref. **[28]** takes a two-step approach: first, exploring structures with ML-accelerated first-principles molecular dynamics; second, using a graph-neural-network architecture for fitting increasingly accurate MLIPs that iteratively "learn" about 2D extended and subsequently about 1D edge structures—details and methodological references may be found in Ref. **[28]**. With the final MLIP model available, MD simulations were carried out, exploring the gradual thermal reduction of a GO sheet.

ML-driven simulations have already begun to have a major impact in materials chemistry and related fields. In the future, together with other emerging AI/ML approaches, they might enable the discovery and design of amorphous functional materials for a variety of practical applications **[27]**.

6. Foundation models for atomistic materials chemistry

Density-functional theory (DFT) and its associated methods have become the standard toolkit of computational materials science and also to a large extent computational chemistry. As such, DFT constitutes the pinnacle of the *Dirac programme* of first-principles modelling **[29]**: start with the fundamental equations of quantum mechanics that describe the electrons and atomic nuclei (the latter represented as point charges to an exceedingly good approximation), and derive the consequences for the behaviour of crystals, molecules, currents, etc. The resulting sequence of approximations over the past ~ 50 years have enabled the description of known - and prediction of new material properties and underpins our understanding of the material world at the microscopic scale.

The computational cost and scaling of DFT in practice limits is general usefulness to the treatment of hundreds of atoms and picoseconds of time scale. While these limitations are being challenged and push all the time by the progress in computational hardware and also algorithmic efficiency, but the extension of first-principles modelling to significantly larger length scales require a change in the modelling framework. Just as DFT eliminates the degrees of freedom inherent in the full many-body wave function and just retains the one-particle operator corresponding to the electron density, we can go further and eliminate electronic degrees of freedom altogether by writing the total energy as a function of just the atomic coordinates: a force field. This function is very complicated, but advances in parametrising functions using a very large number (typically millions) of parameters based on fits to a large amount of data (widely known as *machine learning*) has enabled useful approximations that

allow the simulation of tens of thousands of atoms for millions of time steps, i.e. *nanometers* of material for *nanoseconds.*

The past decade or so has seen incredible progress, and was spent mostly understanding how to build datasets for fitting ML force fields for particular systems, how achieve the accuracy that is required for the model to be usefully predictive of interesting properties **[25]**. Indeed even just characterising the relationship between the "pointwise" accuracy of the potential energy of a force field to the error in its prediction of any particular material property is highly nontrivial, and turns out to be critical for success. Simulations of phase transitions both under equilibrium and nonequilibrium conditions, heterogeneous catalysis, study of diffusion and spatiotemporal correlation are now routinely possible for complex materials.

Just very recently it was discovered that when the training set is diverse enough, a force field can be made that covers most of periodic table, and despite only having been fitted to DFT calculations of small inorganic periodic crystals, is capable of running stable molecular dynamics on essentially any chemical system **[30]**. Such extreme generalisation goes somewhat counter to the conventional wisdom in machine learning research, which has made tremendous progress recently by using ever larger data sets. There is currently little understanding of what gives rise to such generalisation, but it raises the tantalising possibility of a *universal* force field. There is no doubt that further accuracy for a wide range of systems will be gained by training on large databases, and the construction of numerically consistent DFT data is currently the limiting factor.

7. Machine learning electrochemistry

The accurate description of redox reactions from the perspective of first-principles calculations still represents a challenge. Standard density-functional theory (DFT) approximations to the exchange-correlation functionals suffers from the so-called selfinteraction errors (SIE), leading to an unphysical delocalization of electrons and thereby limiting its ability to accurately study processes where changes in oxidation states are critical. Hybrid functionals, and even more extended Hubbard functionals (DFT+U+V) can provide a successful solution to this challenge **[31]**. As recently shown, DFT+U+V provides a robust framework to mitigate SIE in materials with strongly localized *d* or *f* electrons, especially for systems where the electronic localization occurs alongside with substantial hybridization. Recently, it has been shown how the use of DFT+U+V along with first-principles molecular dynamics (FPMD) is capable to follow the adiabatic evolution of oxidation states over time in representative cathode materials for Li-ion batteries **[32]**. In addition, this opens the door to incorporating the concept of redox-aware into machine-learned potentials. Starting from the physical rationale that atoms with different oxidation states behave like distinct species, it has been shown that a neural-network training that considers atoms with different oxidation states (obtained through DFT+U+V FPMD) as distinct species can identify the correct ground state and pattern of oxidation states for the redox elements present **[32]**. This can be achieved through a combinatorial search for the lowest-energy configuration, among all possible patterns, and is shown to recover correctly

the DFT+U+V ground state. This brings the time-scale and length-scale advantages of machine-learned potentials to central technological applications (e.g., rechargeable batteries), which require the correct description of redox states.

The predictive accuracy of DFT+U+V heavily depends on the precise determination of the onsite U and inter-site V Hubbard parameters, which describe localization and hybridization, respectively. While in the simplest cases these parameters could be obtained through semiempirical tuning (but then negating the predictive power of the approach, and the capability to deal with complex and very diverse local environments, that require atom-specific U and V), unbiased predictions identify Hubbard parameters self-consistently through linearresponse calculations, particularly efficient when density-functional perturbation theory (DFPT) is deployed. This approach has now been fully automated **[33]**, enabling highthroughput calculations of Hubbard parameters that can provide extensive datasets for further investigations.

In particular, it becomes even possible to build a machine learning model to predict these bypassing the DFPT step. For example, the machine learning method of Ref. **[34]** has been recently devised to this goal. The model is based on equivariant neural networks, and uses electronic occupation matrices as descriptors, capturing the electronic structure, local chemical environment, and oxidation states of the system in question. The model significantly speeds up the prediction of Hubbard parameters, while approaching the accuracy of DFPT. The model uses two DFT-based calculations: first, a DFT+U+V ground-state calculation with initial guesses for U and V (which can be set to zero) to obtain atomic occupation matrices; second, a structural optimization using the model-predicted self-consistent (SC) Hubbard parameters to obtain the SC structural-electronic ground state. Furthermore, thanks to its strong transferability, it enables accelerated materials discovery and design via high-throughput calculations, with relevance for various technological applications.

Another key topic in computational electrochemistry is the accurate calculation of molecular ion solvation energies, crucial for controlling electrochemical reactions. In particular, this information is essential for the characterization of of relative phase stability in different environments, and thus of major interest to advanced materials and manufacturing. First and foremost, first-principles accuracy is needed to determine the solvation energies of ions and small molecules in arbitrary solvents. The neural network potentials discussed in the previous Sections make these calculations viable, and overcome the computational bottlenecks of FPMD. A recently developed neural network-based workflow **[35]** has shown the capability to compute ion solvation energies for alkaline(-earth) cations with chemical accuracy. Future directions will involve developing active learning schemes to automate the calculations' workflows. Moreover, electrostatic interactions that have been treated directly through the neural network for the short range and analytically for the long range might need to take into account the complex nature of the electrochemical potential across all length scales.

8. Machine learning for molecular sensing

Machine learning is fundamentally transforming molecular sensing, particularly in gas sensing field, by revolutionizing the screening of sensing materials and the enhancement of sensor

performance through advanced signal processing techniques. By integrating machine learning with theoretical tools (*e.g.,* density-functional theory, DFT), researchers have unlocked a powerful methodology for designing selective gas sensing materials and decoding complex sensor signals. This synergistic approach accelerates material discovery and sensor optimization, paving the way for molecular sensing devices that are more sensitive, selective, and reliable.

Traditional methods for designing and screening gas sensing materials rely on trial-and-error experimentation, which is often labor-intensive and time-consuming. By contrast, machine learning, when combined with computational tools, facilitates the efficient prediction of material properties, significantly streamlining the process. For instance, machine learning models can correlate key material descriptors, such as adsorption energy, surface reactivity and electronic properties, with their responses to specific gases. This enables rapid screening and selection of materials without the need for exhaustive experimental validation. For instance, in a recent study, machine learning combined with DFT successfully predicted the sensitivity of $Cs_3Cu_2I_5$ to hydrogen sulfide, achieving a remarkable 92% accuracy in predictions, which were later validated experimentally **[36]**. Beyond accelerating material discovery, this integration also provides valuable mechanism insights into gas adsorption and sensitivity, enabling a deeper understanding of the materials' functionality. Similarly, for metal oxide materials-based sensors, machine learning models have been instrumental in identifying critical descriptors that dictate their sensing capabilities, guiding the targeted selection of materials for diverse applications ranging from industrial safety to environmental monitoring **[37]**.

Following the selection of sensing materials, machine learning continues to play a pivotal role in optimizing sensor performance by fine tuning critical parameters such as sensitivity, selectivity, and response time. It's reported that machine learning techniques have been applied to gas-sensing platforms based on copper phthalocyanine functionalized graphene, enhancing their ability to detect trace amounts of gases like ammonia and phosphine **[38]**. By analysing the sensor's responses, machine learning improves accuracy and specificity, even in complex gas mixtures. Furthermore, machine learning has proven invaluable in the design of sensor arrays capable of detecting multiple gases simultaneously. Algorithms are employed to analyse interactions between sensor elements and their collective responses, enabling the identification of the most effective configurations. This optimization is particularly critical for applications like air quality monitoring, where the simultaneous and accurate identification of various pollutants is essential.

In molecular sensing, interpreting gas sensing signals is crucial. Machine learning techniques are extensively utilized in signal processing to extract meaningful features from raw sensor data while minimizing noise, a common challenge in real-world sensing environments. For instance, machine learning has been used in electronic noses to extract transient kinetic features from sensor response profiles. These features act as distinct fingerprints of odorants, enabling the accurate classification of volatile organic compounds and addressing one of the field's primary challenges **[39]**.

Despite its transformative potential, the application of machine learning in molecular sensing faces several challenges, including improving the interpretability of machine learning models, reducing dependence on large datasets, and enhancing the real-time performance of sensing systems, as well as energy-consuming. Overcoming these hurdles will require continued advancements in machine learning techniques, such as the integration of deep learning and reinforcement learning, development of more accurate adaptive sensing systems, as well as development of brain-inspired neuromorphic computing system **[40]**. Future developments could enable gas sensors that not only detect and classify gases but also predict environmental changes or potential hazards. Such advancements will pave the way for smart, autonomous sensing systems across diverse domains, including healthcare, environmental monitoring, and industrial safety.

To sum up, the integration of machine learning with theoretical tools is revolutionizing the design and optimization of molecular sensors. By expediting material discovery, refining sensor configurations, and enhancing signal processing, machine learning stands at the forefront of developing next-generation molecular sensing technologies. These innovations promise enhanced sensitivity, selectivity, and performance, ensuring their pivotal role in addressing the challenges of modern sensing applications.

9. Refining Molecular Characterization for Robust Machine-guided Corrosion Inhibitor Discovery

A particularly urgent and industrially significant case where machine learning is being applied to discover effective molecular materials is in the discovery of corrosion inhibitors. Such inhibitors would be embedded in the primer of a paint system or used in initial metal passivation. Traditionally inhibitors have been based on chromate or other toxic compounds that are being banned by legislation worldwide. Small hetero-cyclic compounds are a promising alterative, yet to determine the exact molecular structure with high-efficient corrosion inhibition from tens of thousands of possibilities remains a challenge. Various methods including high-throughput experimentation and computational modelling have been developed to select or design the optimal molecular structure. A very promising approach is the use of inverse design, in which high throughput experiments defining electrochemical performance and computation methods deriving inhibitor characteristics and attributes are linked by a machine learning (ML) method to define the molecular attributes essentially important for inhibition performance. These critical features are then used to search molecular databases and select promising candidate inhibitors that are then subject to testing for verification.

Early work was able to use quantitative structural activity relationships (QSAR) methods based on a neural network approach to obtain reasonable models of the features controlling inhibition **[41].** However, while these models represent successfully the existing dataset, further generalization ability was still to be enhanced. Challenges may come in twofold: (i) the relevance of molecular characterization and attribute definitions; and (ii) the deficiency of computationally/experimentally generated datasets. Since then, large programs have been undertaken, where the databases were significantly enhanced, and great care was cast to

ensure that both experimental and computational data were accurate and reproducible. Further the molecular attributes were refined to better represent molecular interactions with solvent and metal surfaces. Ranges of statistical and QSAR techniques have been used to define the relationships between the molecular attributes and electrochemical performance. These models demonstrated an enhanced ability to represent existing data, but their predictive ability could still be enhanced **[42]**. Recent experimental work reveals the complexity of corrosion inhibition process, of which peak performance (both electrochemically and mechanically) was reached by short-term inhibitor treatment, but subsequential voids appeared within the inhibitor film with time expansion **[43]**. Molecular-dynamics models indicated that the inhibitor film may be subject to electroporation where charge at the metal surface causes the inhibitors to clump together allowing water to again reach the metal surface **[44]**.

Thus, it is evident that additional factors involved in the inhibitor adsorption control the overall performance and stability of inhibitor layer, entailing a deeper understanding of inhibitor layer formation and lifetime. A recent review **[45]** highlighted the limitations of previous models: inadequate or no representation of solvent, lack of potential effects and relatively small models. New methodology was proposed based on a combined quantum mechanics/molecular mechanics/non equilibrium greens function (QM/MM/NEGF) approach. This approach enables the simulation of larger models that include both solvent and voltage effects. The system has been applied to the inhibition study of both copper and zinc surfaces by 2- mercaptobenzimidazole (MBI). A major result of the study is that, when MBI binds to the surface, a major electronic re-alignment across the inhibitor assembly rearranges the dipole moment at the exterior of the molecule. The traditional theory of inhibitors is that they form a barrier to both water and solvents against charge transfer. This study proposed that while MBI acts as an effective barrier against water, it cannot be regarded as a charge barrier. In fact, charge realignment and the formation of the dipole will have a profound influence on the deposition of the subsequent inhibitor layer. The relevant molecular attributes contributing to the dynamics of corrosion inhibition processes are potentially important descriptors that were previously overlooked for ML development. As highlighted in our studies, the molecular attributes that can represent these processes are quite different from those that reflect surface bonding and thus our datasets used in ML approaches need significant redesign.

10. Exploring New Frontiers in Inverse Materials Design through Graph Neural Networks and Large Language Models

Finding new materials with suitable properties has been a challenging task due to the computational and experimental costs. AI/ML techniques have been successfully used for both forward (structure to property) and inverse (property to structure) tasks in materials design **[46]**. Inverse design approaches can surpass traditional funnel-like materials screening methods and facilitate the computational discovery of next-generation materials. Since no explicitly available physics-based methods exist for inverse design tasks, AI/ML is an obvious choice.

To accomplish inverse materials design tasks, we require the following: 1) a well-curated and diverse dataset, 2) an AI/ML model and architecture that can establish a mapping between the properties of materials and material structures, and 3) suitable metrics and a benchmarking strategy to guide the design process. While there are numerous material properties—such as electronic bandgap, bulk modulus, refractive index, etc., or their combinations—that can be used as target properties, we can start with a specific property, such as superconducting transition temperature (Tc). Superconductors are one of the most celebrated classes of materials in materials science, but there are very few such materials known experimentally. As mentioned above, we require a dataset for superconductors. While many materials databases exist, they lacked superconducting properties until JARVIS-DFT.

JARVIS-DFT **[47]** consists of more than 80,000 materials and millions of material properties, with around 1,000 superconducting materials in the dataset. Note that predicting Tc is computationally expensive compared to other properties, such as formation energy, when using DFT. As the next step for inverse design, we require AI/ML methods suitable for this task. There are a variety of AI/ML methods, such as fingerprint-based traditional methods, deep learning techniques like convolutional neural networks (CNNs), graph neural networks (GNNs), and generative pre-trained transformers (GPTs).

GNNs, in particular, have been successful recently for atomistic materials design tasks. In these models, atoms are represented as nodes, bonds as edges, and angles as edges of the corresponding line graphs, for instance. GNNs such as Atomistic Line Graph Neural Networks (ALIGNN), combined with diffusion models like the Crystal Diffusion Variational Autoencoder (CDVAE), have enabled the generation of superconducting atomic structures **[48]**. The dataset was split into training and testing sets, and the metric for performance was the interatomic distances between target and predicted structures in the test dataset. After model development, more than 50 candidate superconductors were computationally discovered and later characterized with density-functional theory (DFT) to validate AI predictions. Another AI approach used was GPT models.

In GPT models such as AtomGPT, both the atomic structure and the target property can be represented as text **[49]**. These texts are converted into tokens, and GPT models establish the relationship between the atomic structure and property/prompt tokens. Such GPT models have shown remarkable promise for both forward and inverse materials design tasks. For the superconducting dataset, we followed similar train-test splits as in GNN methods and measured performance based on the interatomic bond distance comparison metric between target and predicted materials in the test dataset. We found that GPT-based models surpass GNN models in terms of this metric, and new candidate superconductors were computationally discovered and later validated with DFT. Additionally, GPT models are much faster and easier to implement than GNN models. These comparisons are hosted on the JARVIS-Leaderboard **[50]** open-source platform to enhance reproducibility, transparency, and allow others to contribute their models as well.

11. Property directed generative design of inorganic materials

Property-directed generative design presents a unique opportunity in modern materials discovery, shifting from large-scale data-driven screening to precise, generative approaches aimed at discovering novel compounds with tailored properties. Recent advancements in computational science for materials discovery have made significant progress in addressing one of the key challenges in crystal structure prediction (CSP) — identifying stable and metastable structures efficiently **[51]**. Traditional CSP methods, such as evolutionary algorithms and particle swarm optimization, however, are computationally expensive and limited in their ability to explore vast chemical spaces. Generative models, by contrast, provide a promising alternative by efficiently targeting structures that are near ground-state configurations when trained on existing data. These generative frameworks also enable inverse design, where the desired targets guide the generation of materials, making them particularly valuable for property-directed generative design.

However, a major challenge in applying generative models is in ensuring that generated structures obey the symmetry and periodicity essential for physical plausibility. Symmetry considerations are fundamental for determining key properties of inorganic materials, including electronic band structures, optical behaviour, and mechanical strength. We summarize some recent frameworks, such as DiffSCP++ **[52]**, CrystalFormer **[53]**, WyCryst **[54]**, MatterGen **[55]**, and PGCGM **[56]**, which have demonstrated the importance of integrating symmetry constraints into a generative framework to ensure the physical plausibility of generated materials. These models utilize a variety of AI-driven models, such as symmetry-based representations, diffusion-based methods, and graph neural networks, to generate stable and diverse crystal structures that satisfy specific property requirements. By embedding symmetry into the generative process, these frameworks enhance the efficiency of materials discovery, reduce reliance on trial-and-error experimentation, and open new avenues for the design of materials with applications in energy, electronics, and catalysis.

Among these models, WyCryst enables symmetry-constrained structure generation through three key components: a Wyckoff position-based representation to enforce symmetry constraints, a property-directed variational autoencoder (PVAE) for generating novel crystal structures, and an automated density-functional theory (DFT) workflow for validating the stability and properties of the generated materials. By embedding symmetry constraints, WyCryst efficiently generates materials that adhere to space group symmetries while meeting desired property criteria. Similarly, DiffSCP++ employs a symmetry-constrained diffusion model to refine atom types, positions, and lattice parameters, ensuring that the generated structures maintain realistic symmetry and periodicity. This approach enhances the diversity and stability of generated materials, opening new possibilities for discovering synthesizable inorganic compounds. CrystalFormer employs a transformer-based architecture to generate crystal structures by predicting symmetry-inequivalent Wyckoff positions in the unit cell ensuring compliance with space group symmetries: this produces thermodynamically stable materials with various symmetries. CrystalFormer is also capable of performing propertyguided exploration with probabilistic modelling, facilitating the discovery of inorganic compounds with targeted properties. MatterGen employs an SE(3) equivariant diffusion approach to generate crystal structures by iteratively refining random initial configurations until they conform to a targeted distribution. MatterGen, as a base pre-trained model, can be

finetuned towards stability or functional properties, facilitating the discovery of materials tailored to specific applications. The PGCGM (Physics Guided Crystal Generative Model) achieves symmetry-based generative design by incorporating physics-oriented losses related to physics and space group symmetry. The model training emphasizes thermodynamic stability ensuring the generation of low energy compounds. PGCGM also enables the generation of crystal structures with specific space group symmetries, allowing further discovery of functional materials.

In conclusion, property-directed generative design frameworks represent a significant advancement in the field of materials science. By embedding symmetry-based constrains into the generative process, these models enhance the validity and stability of predicted materials, thereby accelerating the discovery of inorganic compounds with desired properties. A key bottleneck is the generation of experimental or high-quality computational data to train such the generative models. This approach, however promises not only a streamlined materials design process but also new avenues for the development of advanced materials tailored for specific applications. The synergy between AI models and physics-based property-directed design holds immense promise for revolutionizing the way materials are discovered and optimized for real-world use.

12. Physics based machine learning for materials and compound space

The virtual navigation of chemical compound space has been significantly constrained by the prohibitive computational demand associated with numerically solving approximations to Schrödinger's equation with satisfying accuracy for an exponentially growing number of possible systems. Over the last decade, considerable progress has been realized thanks to the application of statistical techniques commonly referred to as artificial intelligence, as recently documented in an entire issue in *Chemical Reviews* dedicated to machine learning at the atomic scale **[57]**. Due to the colossal number of potential and costly training compounds, the central inquiry has been on how to improve training efficiency — as quantified by scaling laws (or learning curves). This question has persisted ever since it was first demonstrated that machine learning models of quantum properties can be applied throughout chemical compound space, i.e. for out-of-sample systems (not part of training), with prediction errors that decay systematically with training set size **[58]**. Subsequent applications have highlighted the promise of machine learning for the atomistic sciences by systematically surpassing the accuracy of hybrid density-functional theory (DFT) approximations for various quantum properties **[59]**, estimating formation energies for millions of quaternary crystals **[60]** or reaching the accuracy of explicitly correlated electronic structure theory methods through Δlearning **[61]**, or multi-level learning **[62]**. Further breakthroughs in training efficiency, scalability, and transferability were achieved by virtue of similarity-based query aware models, trained on the fly, and decomposition of training and testing systems into fragments, based on Atoms-in-Molecule-ONs (AMONs) **[63]**.

Most recent contributions indicate that meaningful combinations of these techniques are possible, often via intimate combinations with DFT. As such, DFT has assumed an outstanding role for the use of artificial intelligence in chemistry and materials not only for merely

generating data sets for training and testing but also for informing superior machine learning model architectures and workflows **[64]**. Specific examples include the combination of Δlearning and AMONs to enable quantum Monte Carlo level of accuracy **[65]**, similarity-based learning and ridge regression identifying potentially superconducting candidates **[66]**, or adaptive hybrid DFT which reaches superior accuracies when it comes to singlet-triplet spingaps or other quantum observables **[67]**.
The work mentioned only represents a small glimpse of recent activities in the entire and rapidly growing field. Overall, remarkable progress has been made towards the generic goal of reaching EAST, i.e. Efficiency, Accuracy, Scalability, and Transferability **[64]**. Remaining challenges include the generation of more and sufficient data that is universally representative not only for minima but also for barriers, foundational machine learning models that can be used to estimate any quantum mechanical observable in any electronic state, and the possibility to account for multi-reference, as well as nuclear quantum and relativistic effects.

13. Language models for many-body physics

Now is an exciting time for research on quantum physics due to the opportunities and significant advances in the application of machine learning (ML) and artificial intelligence (AI) to fundamental problems in physics, chemistry, and materials science. In particular, the transformative power of language models like Recurrent Neural Networks (RNNs) and Transformers **[68]**, originally designed for natural language processing (NLP), has opened a new frontier across a wide array of technologically and scientifically relevant disciplines, including classical and quantum many-body physics.

Although historically these models originally demonstrated breakthrough performances in NLP, such as in ChatGPT **[68]**, they have in principle little to do with ”language” itself. From a broader perspective, these models constitute powerful statistical modelling and information processing machines that can process a wide array of data types exhibiting correlations of different nature not limited to language. Tokens traditionally understood as pieces of words or phrases could also represent physical or chemical degrees of freedom, namely, spins in a lattice, lattice occupation numbers, atomic coordinates, or generally any sequence of inputs that are statistically mutually dependent. By expanding the token universe to encompass states from any other degrees of freedom relevant to a physical system, large language models (LLMs) can allow physicists to simulate many-body interactions with unprecedented precision and efficiency.

While the origin of these token streams is disparate, the statistical correlations in datasets commonly used in NLP, computer vision, and other popular tasks in ML, display striking similarities with data from physical systems. Key similarities include symmetries, high dimensionality, and correlation functions. For example, spatial symmetries in present in natural datasets and classical and quantum systems simultaneously improve the sample complexity and learnability of models in computer vision as well as enriches our understanding of physical systems in classical and quantum mechanics. The behaviour of the correlations among the constituent elements in the token streams in computer vision and NLP display strikingly analogous behaviour to classical and quantum systems in thermal

equilibrium near a critical point **[69]**. These commonalities make it natural to attempt to use these models to study classical and quantum many body systems and are an important reason behind their rise and success in quantum many-body physics research.

Recent research has begun to capitalize on this potential. Techniques such as RNN wave functions **[70]** and language model-based quantum state tomography offer flexible and powerful representations of quantum states than conventional approaches. These studies have been extended to the task of finding ground states of quantum many-body systems, e.g., ground states of frustrated magnets, Rydberg atoms arrays, and fermionic systems, as well as to simulate the time evolution of quantum states and to solve combinatorial optimization problems **[71]**. Quantum chemistry, a field crucial for understanding molecular interactions and reactions, has also benefited from these advances. Transformer-based models can predict molecular ground state energies with comparable accuracy to traditional methods. Such advancements hold immense promise for rapid simulations in quantum chemistry, offering a pathway toward scalable tools that handle large basis sets and dense electron correlations that are challenging for standard quantum chemistry methods.

One potential application of LLMs in physics is the simulation of many-body fermion systems, such as those encountered in Rydberg arrays, exotic material phases, or molecules. Models such as ”RydbergGPT” **[72]** could enable simulations, potentially influencing quantum computing and materials science in the long term. By offering a scalable and adaptable approach to many-body physics, LLMs could present a complementary method to state-ofthe-art algorithms such as Quantum Monte Carlo and Density Matrix Renormalization Group, especially when dealing with high-dimensional frustrated or out-of-equilibrium systems.

Looking ahead, the development of efficient and environmentally conscious LLMs is critical. The computational costs of training these models using Graphical Processing Units (GPUs) are significant, and reducing the environmental impact of large-scale simulations remains a pressing concern even in physics simulations based on language models. Innovations in model architecture design could help address this issue, aligning with the broader push for sustainable computing. In conclusion, language models can become versatile tools for scientists by bridging the gap between language processing and physical simulations, impacting fields beyond NLP **[71]**. As research in this space advances, LLMs may catalyze breakthroughs across many-body physics, quantum chemistry, and beyond, unlocking a new era of data- and physics-driven, efficient, and scalable many-body systems simulations.

14. Variational autoencoders-enabled high-fidelity reconstruction and effective anomaly detection in time-series data

Robust modelling of multi-channel biological time-series data, such as EEG, across different individuals is crucial in numerous applications. Most often, identifying common patterns (*biomarkers*) is as relevant as distinguishing them from individual behaviors (*fingerprints*). However, achieving accurate modelling involves tackling three primary challenges:

intersubject variability, intra-subject variability, and ensuring data quality and fairness, including the automatic detection of artifacts **[73]**.

We used the well-known *BCI dataset 2a*, a very popular EEG dataset collected from nine subjects performing motor imagery of hand and feet movements, to test both classification and reconstruction using various deep learning models **[74]**. First, *vEEGNet-ver1* served as the baseline model upon which we built subsequent versions. *vEEGNet-ver1* is a variational autoencoder with the encoder inspired by the popular EEGNet architecture, with three main convolutional layers. The decoder is the mirrored version of the encoder. By enhancing its encoder architecture, we developed *vEEGNet-ver2*, which offered improved performance over the first version in terms of reconstruction. Then, we decided to focus on the reconstruction task, in line with previous literature indicating a trade-off between classification and reconstruction learning abilities **[75]**. This led to the creation of vEEGNetver3, which targets a single task, i.e., the reconstruction. In *vEEGNet-ver3*, we defined the reconstruction loss as the (soft) dynamic time warping distance between the original and the reconstructed time-series. This approach significantly improved the model's performance, suggesting the importance of concentrating on specific tasks to achieve better results. Finally, by employing a hierarchical variational autoencoder architecture **[76]**, we transformed *vEEGNet-ver3* into the *hvEEGNet* model. This advanced architecture demonstrated highfidelity reconstruction performance and provides three distinct latent representations, extracted from the three latent spaces of the model. As the reconstruction performance on the *dataset 2a* were very high, we tested *hvEEGNet* as an automatic artifact detector, enabling the identification of artifacts that had not been previously detected in the wellknown public dataset.

One of the key insights from our work is the crucial role of domain knowledge that allowed us to recognize that poor reconstruction results were linked to acquisition problems, such as signal saturation, or physiological artefacts, such as eye blinking. Ensuring high data quality is essential for the successful and reliable learning of machine learning models. Without highquality data, even the most advanced algorithms can produce misleading or suboptimal results.

Moreover, the latent representations extracted by *hvEEGNet* can be further investigated to develop new physics-informed smaller and more effective latent space structures **[77]**. Such advancements could pave the way for more robust and informative deep learning models for time-series modelling and anomaly detection. By improving the effectiveness and interpretability of latent representations, future research could address the challenge of distinguishing common patterns from individual ones and better quantify inter- and intrasubject variability. Also, improved interpretability will enable a higher degree of interaction with domain experts, who can help drive the development of deep learning models tailored to their research and clinical questions.

The significance of this research extends beyond this immediate application, as the above challenges are common to other domains where complex living systems are under investigation. Moreover, *hvEEGNet* is a versatile model which can be adjusted to other types time-series data, with different dynamics, and different applications.

15. Multiscale Materials Science: Tasks, Challenges, and Cross-Domain Synergies

Materials science is a field driven by its multiscale nature, where phenomena at vastly different spatial and temporal scales interact to define the properties and behaviors of materials. From atomic vibrations that dictate thermal conductivity to macroscopic structures determining mechanical strength, understanding and predicting material behavior requires bridging these scales.

Traditionally, physics has provided a robust set of mathematical tools to address multiscale problems. Methods such as renormalization groups, effective field theories, and closure coordinates have been used to study specific properties like critical points or ground states. While these approaches have been immensely successful in understanding phase transitions and other fundamental phenomena, they were typically designed to address narrowly focused problems.

Today, the scope of problems in materials science has expanded significantly. Researchers are not only interested in understanding ground states or critical phenomena but also in exploring broader challenges like finding meta-stable states, analyzing mechanical properties under various conditions, and even generating entirely new structures.
Addressing these challenges requires a paradigm shift from traditional analytical methods to new data-driven approaches. Machine learning and artificial intelligence (AI) have emerged as powerful tools to augment classical methods, enabling scientists to model, predict, and design materials across multiple scales with unprecedented efficiency.

This shift toward data-driven methodologies is transforming materials science, creating opportunities to solve problems that were previously intractable and broadening the field's potential impact across domains.

Integrated Multiscale Tasks in Materials Science

In materials science, tasks such as property prediction, conditional structure generation, automated synthesis, and physical law discovery are inherently multiscale. Each of these tasks requires understanding the interplay between small-scale phenomena and large-scale outcomes.

Predicting material properties often involves connecting atomistic interactions to macroscopic behaviors. For example, multiscale deep learning models can predict the elastic properties of woven composites by analyzing data from simulations at the microstructural level **[78]**. These models provide valuable insights into how small changes at the microscale influence the overall performance of a material.

Generating structures with specific properties is a complex inverse problem. Advanced generative models, like those used to predict domain boundaries in potassium sodium niobate thin films, can reveal previously unobserved structural motifs that emerge from simple local rules **[79]**. This work highlights how structural complexity arises naturally from underlying physical principles.

Automating synthesis processes accelerates material discovery by optimizing experimental conditions. For instance, the LeapFrog framework combines adaptive mesh refinement with machine learning to simulate the solidification of alloys, offering insights into how synthesis parameters affect microstructure formation **[80]**.

Discovering physical laws and principles requires connecting diverse scales of phenomena. Compression theory, for example, identifies relevant degrees of freedom in complex systems like quasicrystals, uncovering new critical behaviors that were previously hidden **[81, 82]**.

Universality of Multiscale Methods

One of the most exciting aspects of multiscale methods is their universality. Once developed for a specific domain, these methods can often be applied to entirely different fields. For example, techniques for coarse-graining molecular dynamics with graph neural networks reduce computational costs while generating transferable representations applicable across molecular systems **[83]**. Similarly, data-driven models used to study DNA methylation patterns have uncovered thermodynamic variables that govern healthspan and lifespan across species, demonstrating the potential for cross-domain applications **[84]**.

By leveraging the universality of multiscale approaches, we can accelerate discoveries not only in materials science but also in fields like biology, chemistry, and even social systems.

The Core Challenge: Balancing Detail and Holism

A central challenge in multiscale modeling is determining the appropriate level of detail. Too much detail can make models computationally infeasible, while too little can lead to inaccuracies. For example, the $\mathrm{FE}^{\mathrm{ANN}}$ framework balances accuracy and computational efficiency by using physics-constrained neural networks to model fibre-reinforced composites **[85]**.

When a single level of detail is insufficient, integrating multiple scales into a cohesive framework becomes essential. Flow-matching, a novel method for coarse-grained molecular dynamics, combines generative modeling with force-matching to efficiently capture key interactions across scales **[86]**. Such approaches demonstrate how multiscale frameworks can provide a holistic view without overwhelming computational resources.

Outlook: Toward a Unified Multiscale Ecosystem

The future of materials science lies in creating a unified ecosystem that integrates multiscale simulations, AI, and experimental data. Graph-enhanced deep material networks, for example, unify the modeling of diverse microstructures, enabling predictions across families of materials **[87]**. These tools not only improve accuracy but also pave the way for entirely new material designs.

Furthermore, integrating theoretical principles with data-driven models offers powerful opportunities. For instance, a platform based on the Onsager principle creates reduced thermodynamic coordinates for stochastic systems, allowing for a more profound understanding of complex material behaviours **[88]**.

By developing interoperable, scalable, and transferable tools, we can accelerate innovation, enabling faster discoveries and broader applications of multiscale methodologies.

Multiscale materials science stands at the intersection of computation, experimentation, and theory. From predicting properties to discovering universal principles, multiscale methods allow us to tackle challenges across domains. By balancing detail and holism and leveraging the universality of these approaches, we can push the boundaries of what is possible, not only in materials science but in many other fields as well.

Conclusion and perspective

Digitalization of materials is a strategic action in the frame of emerging twin green & digital transition which aim at a more sustainable and resilient world economy. New digitalization solutions are needed covering the whole materials value chain and interconnecting all phases of the materials life cycle, from materials design and development, production, optimal usage, to maintenance and to re-use, and recycling. Principally these efforts can be broadly categorized into two domains: “digital twins” and materials models in a very broad sense, provided digital representations of materials in the context of their application independent of a specific measurement. Secondly, curated dataspaces provide access to experimental data that forms the basis for the generation of model-based digital representations.

Digital twins and materials models need to extrapolate beyond the limited number of available data points. Given the vast dimension of the materials space this problem will not be solved by high-throughput experiments. For this reason, one of the major challenges of developing accurate and functional modelling of innovative advanced materials (IAM) is to reach the accuracy of first-principles approaches over a very large volume of systems. Moreover, a generic modelling procedure at an experimental scale where material imperfections such as defects, disordered chemical composition, rough interfaces, etc, play a major role, is hardly possible with conventional first-principles techniques. The development of machine-learned (ML) strategies that allow obtaining atomistic models from a large dataset of small and accurate first-principles calculations could enable achieving unprecedented time and length scales. The elaboration of novel types of datasets required to train ML models is also of major concern, but while open-access material databases offer valuable information on thousands of crystalline materials, they overlook the nature and impact of the variety of possible atomic imperfections as usually observed in experiments. Finally, a substantial challenge persists in extracting meaningful physical insights from the vast amount of data (raw images and spectra) generated during experimental analysis. The use of AI-driven methodologies associated with (S)TEM data analysis for instance could boost the automation of experiments and data analysis of IAMs.

It is therefore urgent to develop AI and ML based models embedded into workflows that can accelerate the design of IAMs, and optimize their compositions and structures for enhancing their application performances. Efforts need to be focused on the generalization of AI-driven ML techniques to cope with realistic modelling of IAMs, as well as on the development of

workflows connecting the generation of atomistic models to the simulation of their electronic, transport, thermal and optical properties. Critically, the impact of disorder, interface symmetries or chemical composition on their physical properties (electronic, optical, magnetic, etc), in limiting the use of IAMs for optimization of devices and achieving device metrics' upper limits need to be considered. Additionally, AI-enhanced characterization workflow should be developed to facilitate breakthroughs in data analysis methodologies of IAMs.

Developing physically informed AI-based models can allow material scientists, engineers, and companies to determine the physical properties (electronic, optical, transport, magnetic...) of IAMs in significantly less time than through conventional modelling, hence accelerating the path to innovation and new discoveries. These approaches will boost the exploration of complex structures relevant to energy, electronic, photonic quantum and composites applications. Moreover, properly trained models will enable to quickly test multiple experimental conditions with minimal modelling effort, which will help conventional fab and lab metrologist to access a comprehensive analysis of intricate architectures and compositions of IAMs, serving as a solution to the lack of sufficient statistical sampling for understanding performance variability among individual devices.

The transition from traditional data-centric approaches to more sophisticated foundation models is essential to address these challenges. Foundation models, which generalize across diverse datasets and tasks, will help in scaling up the modelling of IAMs and extracting more actionable insights from data. Thus, it is increasingly important to develop AI- and ML-based models embedded in workflows that not only accelerate the design of IAMs but also optimize their compositions and structures for enhanced application performance. These models should be capable of handling a broader range of tasks, from the generation of atomistic models to the simulation of electronic, transport, thermal, and optical properties. Moreover, the impact of material disorder, interface symmetries, and chemical composition on the physical properties (e.g., electronic, optical, magnetic) must be considered, especially when aiming to push the limits of device performance and metrics.

Materials data spaces (such as Material digital, https://www.materialdigital.de/, FAIRmat https://www.fairmat-nfdi.eu/fairmat/, DIADEM, PSDI, CAPeX or NIMS-MPDF) **[89]** constitute an asset for establishing a Materials Commons infrastructure in which federated data repositories with trusted data management, access, and exchange are provided. Building on the experience of several national, European and international initiatives, harmonized semantic data documentation according to FAIR principles should be developed to support interoperability and AI-readiness of produced IAM data. EU and National initiatives provided a huge and ever-increasing body of materials data that needs to be curated and made accessible to ensure maximal exploitation. To this end the data must be findable and accessible independent of its original format, i.e. semantically. Here the development of core and domain-specific ontologies opens significant long-term opportunities. While sharing the meta-data is uncritical for many stakeholders, the efforts to organize data-spaces must take into account the need for data-provenance for certain datasets. Given the enormous increase of the power of foundational models, the data needed for training emerges as the bottleneck and the potential competitive advantage for Europe.

Beyond modelling and data sharing, the role of platforms that facilitate not only the exchange of materials data but also the processing, automated analysis, and on-the-fly literature analysis is crucial. Such platforms would enable seamless integration of various stages of data flow, from acquisition to interpretation, while simultaneously providing relevant, up-to-date research knowledge that can inform and hasten the experimental or computational task at hand. This functionality would enable the acceleration of innovation in IAMs by ensuring that researchers have access to both empirical data and the latest theoretical insights. AIenhanced characterization workflows are particularly important for automating and refining data analysis techniques, allowing for rapid testing of experimental conditions with minimal manual effort. This will help experimental researchers access comprehensive analyses of complex IAM architectures and compositions, solving the problem of insufficient statistical sampling and enabling a better understanding of performance variability across breadth of domains of material science.

Combined efforts in the digitalization of materials and the validation of predictive AI models for materials enable the establishment of materials acceleration platforms, or a self-driving lab have an enormous potential to revolutionize and accelerate the development of IAM both in industrial and academic settings **[89]**. In that sense the emerging initiatives in Europe and elsewhere stand as an opportunity but also great challenge and will demand sustained efforts and funding for the decade to come.

References

[1] T. Liu, A. S. Barnard, *Cell Rep. Phys. Sci.,* **4**, 101630 (2023).

[2] S. Li, R. Wang, Q. Deng, A. S. Barnard, *Proceedings of the 12th International Conference on Learning Representations* (2024)

[3] T. Liu, A. S. Barnard, *Proceedings of the 40th International Conference on Machine Learning*, **202**, 21375 (2023).

[4] T. Liu, Z. Y. Tho, A. S. Barnard, *Digital Disc.,* **3,** 422-435 (2024) [5] T. Liu, A. S. Barnard, *Mach. Learn Sci. Technol.,* in review (2024)

[6] Y. T. Liu, K. P. Kelley, R. K. Vasudevan, H. Funakubo, M. A. Ziatdinov and S. V. Kalinin, Nature Machine Intelligence **4** (4), 341-350 (2022).

[7] M. A. Ziatdinov, Y. Liu, A. N. Morozovska, E. A. Eliseev, X. Zhang, I. Takeuchi and S. V. Kalinin, Adv Mater **34** (20), e2201345 (2022).

[8] O. Dyck, S. Kim, E. Jimenez-Izal, A. N. Alexandrova, S. V. Kalinin and S. Jesse, Small **14** (38), e1801771 (2018).

[9] A.H. Ter-Petrosyan, J.A. Bilbrey, C.M. Doty, B.E. Ma|hews, L. Wang, Y. Du, E. Lang, K. Ha|ar, and S.R. Spurgeon. Proceedings of the Machine Learning and the Physical Sciences Workshop, NeurIPS 2023 (2023), DOI:10.48550/arxiv.2311.08585.

[10] N.R. Lewis, Y. Jin, X. Tang, V. Shah, C. Doty, B.E. Matthews, S. Akers, and S.R. Spurgeon. Npj Comput Mater **8**, 252 (2022).

[11] J. Peng, D. Schwalbe-Koda, K. Akkiraju, T. Xie, L. Giordano, Y. Yu, C.J. Eom, J.R Lunger, D.J. Zheng, R.R. Rao, S. Muy, J.C. Grossman, K. Reuter, R. Gomez-Bombarelli, and Y. Shao Horn, *Nature Review Materials,* **7**, 991 (2022)

[12] K. Reuter, *Catal. Lett*., **146**, 541 (2016).

[13] A. Bruix, J.T. Margraf, M. Andersen, and K. Reuter, *Nature Catal*., **2**, 659 (2019).

[14] J.T. Margraf, H. Jung, C. Scheurer, and K. Reuter, *Nature Catal*., **6**, 112 (2023).

[15] C. Scheurer and K. Reuter, *Nature Catal.* (2024). (arxiv to be added)

[16] P. Hohenberg and W. Kohn. Phys. Rev. **136**, B864 (1964)

[17] W. Kohn and L. J. Sham. Phys. Rev. **140**, A1133 (1965)

[18] M. Alducin, N. Camillone, S.-Y. Hong, J. I. Juaristi, Phys. Rev. Lett., **123** (2019) 246802.

[19] J. Behler Chem. Rev. 121 (2021) 10037.

[20] Amir Omranpour *et al.* arXiv:2411.00720.

[21] A. Serrano Jiménez, A. S. Muzas, Y. Zhang, J. Ovčar, B. Jiang, I. Lončarić, J. I. Juaristi, and M. Alducin, J. Chem. Theory Comput. 17 (2021) 4648

[22] A. S. Muzas, A. Serrano Jiménez, Y. Zhang, B. Jiang, J. I. Juaristi, and M. Alducin, J. Phys. Chem. Lett. 15 (2024) 2587.

[23] I. Žugec, A. Tetenoire, A. S. Muzas, Y. Zhang, B. Jiang, M. Alducin, J. I. Juaristi, J. Amer. Chem. Soc. Au, 4 (2024) 1997.

[24] C. Chang, V. L. Deringer, K. S. Katti, V. Van Speybroeck, C. M. Wolverton, "Simulations in the era of exascale computing", *Nat. Rev. Mater.* **8**, 309 (2023).

[25] Deringer et al. Gaussian Process Regression for Materials and Molecules, *Chemical Reviews*, **121**, pp 10073-10141 (2021)

[26] C. Ben Mahmoud, J. L. A. Gardner, V. L. Deringer, "Data as the next challenge in atomistic machine learning", *Nat. Comput. Sci.* **4**, 384 (2024).

[27] Y. Liu, A. Madanchi, A. S. Anker, L. Simine, V. L. Deringer, "The amorphous state as a frontier in computational materials design", *Nat. Rev. Mater.* (2024), published online at https://doi.org/10.1038/s41578-024-00754-2.

[28] Z. El-Machachi et al., "Accelerated First-Principles Exploration of Structure and Reactivity in Graphene Oxide", *Angew. Chem. Int. Ed.* **63**, e202410088 (2024).

[29] P. A. M. Dirac, Quantum mechanics of many-electron systems, *Proc. R. Soc. Lond. A* **123** pp714–733 (1929)

[30] Batatia et al. A foundation model for atomistic materials chemistry, https://arxiv.org/abs/2401.00096

[31] M. Cococcioni and N. Marzari. *Physical Review Materials*, **3**, 033801 (2019)

[32] C. Malica and N. Marzari. Teaching oxidation states to neural networks, https://arxiv.org/abs/2412.01652

[33] L. Bastonero, C. Malica, E. Macke, M. Bercx, S. Huber, I. Timrov, and N. Marzari. First-principles Hubbard parameters with automated and repodicible workflows, https://arxiv.org/pdf/2503.01590

[34] M. Uhrin, A. Zadoks, L. Binci, N. Marzari, I. Timrov. arXiv preprint arXiv:2406.02457, 2024

[35] N. Bonnet and N. Marzari. *Journal of Chemical Theory and Computation, **20**, 4820 (2023)*

[36] Gao, S., Cheng, Y., Chen, L. & Huang, S. Rapid Discovery of Gas Response in Materials Via Density Functional Theory and Machine Learning. *Energy & Environmental Materials* (2024). https://doi.org:10.1002/eem2.12816

[37] Yang, Z. *et al.* General Model for Predicting Response of Gas-Sensitive Materials to Target Gas Based on Machine Learning. *ACS Sens* **9**, 2509-2519 (2024). https://doi.org:10.1021/acssensors.4c00186

[38] Huang, S. *et al.* Machine Learning-Enabled Smart Gas Sensing Platform for Identification of Industrial Gases. *Advanced Intelligent Systems* **4**, 2200016 (2022). https://doi.org:https://doi.org/10.1002/aisy.202200016

[39] Huang, S. *et al.* Machine learning-enabled graphene-based electronic olfaction sensors and their olfactory performance assessment. *Applied Physics Reviews* **10** (2023). https://doi.org:10.1063/5.0132177

[40] Baek, E. *et al.* Intrinsic plas city of silicon nanowire neurotransistors for dynamic memory and learning func ons. *Nature Electronics* **3**, 398-408 (2020). https://doi.org:10.1038/s41928-020-0412-1

[41] Chen, F. F., Breedon, M., White, P., Chu, C., Mallick, D., Thomas, S., Sapper, E., & Cole, I. (2016). Correla on between molecular features and electrochemical proper es using an ar ficial neural network. *Materials and Design*, *112*. https://doi.org/10.1016/j.matdes.2016.09.084

[42] Winkler, D. A., Breedon, M., White, P., Hughes, A. E., Sapper, E. D., & Cole, I. (2016). Using high throughput experimental data and in silico models to discover alterna ves to toxic chromate corrosion inhibitors. *Corrosion Science*, *106*. https://doi.org/10.1016/j.corsci.2016.02.008

[43] Deng, Q., Rafiuddin Jakeria, M., Elbourne, A., Chen, X.-B., & Cole, I. S. (2025). Revisi ng inhibi on stability of 2-mecaptobenzimidazole as corrosion inhibitor against saline corrosive media: A combined in-situ and ex-situ inves ga on. *Applied Surface Science*, *681*. https://doi.org/10.1016/j.apsusc.2024.161558

[44] Jeschke, S., Eiden, P., Deng, Q., Cole, I. S., & Keil, P. (2024). Structure and Dynamics of Aqueous 2-Aminothiazole/NaCl Electrolytes at Electrified Interfaces. *Journal of Physical Chemistry B*, *128*(25), 6189–6196. https://doi.org/10.1021/acs.jpcb.4c01479.

[45] Castillo-Robles, J. M., de Freitas Martins, E., Ordejón, P., & Cole, I. (2024). Molecular modeling applied to corrosion inhibition: a critical review. *Npj Materials Degradation*, *8*(1). https://doi.org/10.1038/s41529-024-00478-2

[46] Choudhary, K., DeCost, B., Chen, C. *et al. npj Comput Mater* 8, 59 (2022).

[47] Choudhary, K., Garrity, K.F., Reid, A.C.E. *et al. npj Comput Mater* **6**, 173 (2020).

[48] Wines D., Xie T., and Choudhary, K. *J. Phys. Chem. Lett.* 14, 29 (2023).

[49] Choudhary, K. J. *Phys. Chem. Lett.* 15, 27 (2024).

[50] Choudhary, K., Wines, D., Li, K. *et al. npj Comput Mater* **10**, 93 (2024).

[51] D. Yan, A. D. Smith, and C.-C. Chen, "Structure prediction and materials design with generative neural networks," Nat. Comput. Sci., vol. 3, no. 7, pp. 572–574, 2023, doi: 10.1038/s43588-023-00471-w.

[52] R. Jiao et al., "SPACE GROUP CONSTRAINED CRYSTAL GENERATION," pp. 1–18, 2024.

[53] Z. Cao, X. Luo, J. Lv, and L. Wang, "Space Group Informed Transformer for Crystalline Materials Generation," pp. 1–26, 2024, [Online]. Available: http://arxiv.org/abs/2403.15734

[54] R. Zhu, W. Nong, S. Yamazaki, and K. Hippalgaonkar, "WyCryst: Wyckoff inorganic crystal generator framework," Matter, pp. 1–20, 2024, doi: 10.1016/j.matt.2024.05.042.

[55] C. Zeni et al., "MatterGen: a generative model for inorganic materials design," pp. 1–56, 2023, [Online]. Available: http://arxiv.org/abs/2312.03687

[56] Y. Zhao et al., "Physics guided deep learning for generative design of crystal materials with symmetry constraints," npj Comput. Mater., vol. 9, no. 1, p. 38, 2023, doi: 10.1038/s41524-023-00987-9.

[57] M. Ceriotti, C. Clementi, OA. von Lilienfeld. *Chemical Reviews*, **121**, 9719 (2021)

[58] OA. von Lilienfeld, *Angewandte Chemie International Edition*, **57**, 4164 (2018)

[59] F. Faber et al. *Journal of Chemical Theory and Computation*, **13**, 5255 (2017);

[60] F. Faber et al. *Physical Review Letters*, **117**, 135502 (2016);

[61] R. Ramakrishnan et al. *Journal of Chemical Theory and Computation,* **11**, 2087 (2015);

[62] S. Heinen et al. *Machine Learning: Science and Technology*, **5**, 025058 (2024);

[63] OA. von Lilienfeld, *Machine Learning: Science and Technology*, **4**, 045043 (2023);

[64] B. Huang, GF. von Rudorff, OA. von Lilienfeld *Science*, **381**, 170 (2023);

[65] B. Huang et al. *Journal of Chemical Theory and Computation,* **19**, 1711 (2023);

[66] S. Lee et al. arXiv:2406.14524 (2024);

[67] D. Khan et al. *Science Advances*, in print, arXiv:2402.14793 (2024).

[68] Humza Naveed, Asad Ullah Khan, Shi Qiu, Muhammad Saqib, Saeed Anwar, Muhammad Usman, Naveed Akhtar, Nick Barnes, and Ajmal Mian, "A Comprehensive Overview of Large Language Models," (2024), arXiv:2307.06435 [cs].

[69] Greg J. Stephens, Thierry Mora, Gasper Tkacik, and William Bialek, "Statistical Thermodynamics of Natural Images," Physical Review Letters **110**, 018701 (2013).

[70] Mohamed Hibat-Allah, Martin Ganahl, Lauren E. Hayward, Roger G. Melko, and Juan Carrasquilla, "Recurrent neural network wave functions," Phys. Rev. Research **2**, 023358 (2020).

[71] Roger G. Melko and Juan Carrasquilla, "Language models for quantum simulation," Nat Comput Sci **4**, 11–18 (2024).

[70]David Fitzek, Yi Hong Teoh, Hin Pok Fung, Gebremedhin A. Dagnew, Ejaaz Merali, M. Schuyler Moss, Benjamin MacLellan, and Roger G. Melko, "Rydberggpt," (2024), arXiv:2405.21052 [quant-ph].

[71] C. Chang, V. L. Deringer, K. S. Katti, V. Van Speybroeck, C. M. Wolverton, "Simulations in the era of exascale computing", *Nat. Rev. Mater.* **8**, 309 (2023).

[72] C. Ben Mahmoud, J. L. A. Gardner, V. L. Deringer, "Data as the next challenge in atomistic machine learning", *Nat. Comput. Sci.* **4**, 384 (2024).

[73] Gyori, N. G., Palombo, M., Clark, C. A., Zhang, H., and Alexander, D. C. (2022). Training data distribution significantly impacts the estimation of tissue microstructure with machine learning. *Magnetic Resonance in Medicine* 87, 932–947.

[74] G. Cisotto, A. Zancanaro, I. F. Zoppis, and S. L. Manzoni, "hvEEGNet: a novel deep learning model for high-fidelity EEG reconstruction", *Frontiers in Neuroinformatics, vol. 18, 2024* (*to appear*). doi: 10.3389/fninf.2024.1459970.

[75] Rathjens, J., & Wiskott, L. (2024). Classification and Reconstruction Processes in Deep Predictive Coding Networks: Antagonists or Allies. *arXiv preprint arXiv:2401.09237*.

[76] Vahdat, A., & Kautz, J. (2020). NVAE: A deep hierarchical variational autoencoder. *Advances in neural information processing systems*, *33*, 19667-19679.

[77] Karniadakis, G.E., Kevrekidis, I.G., Lu, L. *et al.* Physics-informed machine learning. *Nat Rev Phys* **3**, 422–440 (2021). https://doi.org/10.1038/s42254-021-00314-5.

[78] Ghane, E., M. Fagerström, and S.M. Mirkhalaf. "A Multiscale Deep Learning Model for Elastic Properties of Woven Composites." *International Journal of Solids and Structures* 282 (October 2023): 112452. https://doi.org/10.1016/j.ijsolstr.2023.112452.

[79] Dan, Jiadong, Moaz Waqar, Ivan Erofeev, Kui Yao, John Wang, Stephen J. Pennycook, and N. Duane Loh. "A Multiscale Generative Model to Understand Disorder in Domain Boundaries." *Science Advances* 9, no. 42 (October 20, 2023): eadj0904. https://doi.org/10.1126/sciadv.adj0904.

[80] Pinto, Damien, Michael Greenwood, and Nikolas Provatas. "LeapFrog: Getting the Jump on Multi-Scale Materials Simulations Using Machine Learning." arXiv, August 2, 2024. https://doi.org/10.48550/arXiv.2406.15326.

[81] Gökmen, Doruk Efe, Sounak Biswas, Sebastian D. Huber, Zohar Ringel, Felix Flicker, and Maciej Koch-Janusz. "Compression Theory for Inhomogeneous Systems." arXiv, May 16, 2023. http://arxiv.org/abs/2301.11934.

[82] Gökmen, Doruk Efe, Zohar Ringel, Sebastian D. Huber, and Maciej Koch-Janusz. "Symmetries and Phase Diagrams with Real-Space Mutual Information Neural Estimation." *Physical Review E* 104, no. 6 (December 6, 2021): 064106. https://doi.org/10.1103/PhysRevE.104.064106.

[83] Husic, Brooke E., Nicholas E. Charron, Dominik Lemm, Jiang Wang, Adrià Pérez, Maciej Majewski, Andreas Krämer, et al. "Coarse Graining Molecular Dynamics with Graph Neural Networks." *The Journal of Chemical Physics* 153, no. 19 (November 21, 2020): 194101. https://doi.org/10.1063/5.0026133.

[84] Denisov, Kirill, and Peter Fedichev. "Discovery of Thermodynamic Control Variables That Independently Regulate Healthspan and Maximum Lifespan," 2024. https://www.biorxiv.org/content/10.1101/2024.12.01.626230v1.full.pdf.

[85] Kalina, Karl A., Lennart Linden, Jörg Brummund, and Markus Kästner. "FE^{ANN}: An Efficient Data-Driven Multiscale Approach Based on Physics-Constrained Neural Networks and Automated Data Mining." *Computational Mechanics* 71, no. 5 (May 2023): 827–51. https://doi.org/10.1007/s00466-02202260-0.

[86] Köhler, Jonas, Yaoyi Chen, Andreas Krämer, Cecilia Clementi, and Frank Noé. "FlowMatching -- Efficient Coarse-Graining of Molecular Dynamics without Forces." arXiv, February 5, 2023. http://arxiv.org/abs/2203.11167.

[87] Jean, Jimmy Gaspard, Tung-Huan Su, Szu-Jui Huang, Cheng-Tang Wu, and Chuin-Shan Chen. "Graph-Enhanced Deep Material Network: Multiscale Materials Modeling with Microstructural Informatics." *Computational Mechanics* 75, no. 1 (January 2025): 113–36. https://doi.org/10.1007/s00466-024-02493-1.

[88] Chen, Xiaoli, Beatrice W. Soh, Zi-En Ooi, Eleonore Vissol-Gaudin, Haijun Yu, Kostya S. Novoselov, Kedar Hippalgaonkar, and Qianxiao Li. "Constructing Custom Thermodynamics Using Deep Learning." arXiv, December 22, 2023. https://doi.org/10.48550/arXiv.2308.04119.

[89] Simon. P. Stier et al. Adv. Mater. 2407791 (2024)