Predicting the Electronic Structure of Matter on Ultra-Large Scales

Authors:

Lenz Fiedler,[1] Normand Modine,[2] Steve Schmerler,[3]
Dayton J. Vogel,[2] Gabriel A. Popoola,[4] Aidan Thompson,[5]
Sivasankaran Rajamanickam,[5*] Attila Cangi[1*]

Affiliations:

[1] Center for Advanced Systems Understanding, Helmholtz-Zentrum Dresden-Rossendorf, Untermarkt 20, Görlitz, 02826, Saxony, Germany

[2] Computational Materials and Data Science, Sandia National Laboratories, 1515, Eubank Blvd, Albuquerque, 87123, NM, USA

[3] Information Services and Computing, Helmholtz-Zentrum Dresden-Rossendorf, Bautzner Landstraße 400, Dresden, 01328, Saxony, Germany

[4] Elder Research, Inc. 300 West Main Street, Charlottesville, 22903 VA, USA

[5] Center for Computing Research, Sandia National Laboratories, 1515, Eubank Blvd, Albuquerque, 87123, NM, USA

[*]To whom correspondence should be addressed; E-mail: srajama@sandia.gov, a.cangi@hzdr.de

October 20, 2022

Abstract: The electronic structure of matter is of fundamental importance for chemistry and materials science. Modeling and simulation rely primarily on density functional theory (DFT), which has become the principal method for predicting electronic structures. While DFT calculations have proven to be incredibly useful, their computational scaling limits them to small systems. We have developed a machine-learning framework for predicting the electronic structure of a given atomic configuration at a much lower computational cost than DFT. Our model demonstrates up to three orders of magnitude

Sandia National Laboratories is a multimission laboratory managed and operated by National Technology & Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.

speedup on systems where DFT is tractable. By leveraging mappings within local atomic environments, our framework also enables robust electronic structure calculations at yet unattainable length scales. Our work demonstrates how machine learning circumvents the long-standing computational bottleneck of DFT.

One-sentence summary: The long-standing problem of predicting the electronic structure of matter on ultra-large scales (beyond 100,000 atoms) is solved with machine learning.

Main text: Electrons are elementary particles of fundamental importance. Their quantum mechanical interactions with each other and with atomic nuclei give rise to the plethora of phenomena we observe in chemistry and materials science. Understanding the probability distribution of electrons in molecules and materials − the so-called electronic structure − propels novel technologies impacting both industry and society. In light of the global challenges related to climate change, green energy, and energy efficiency, the most notable applications that require an explicit insight into the electronic structure of matter include the search for better batteries (*1*) and the identification of more efficient catalysts (*2*) to name just a few. The electronic structure is furthermore of great interest to fundamental physics as it determines the Hamiltonian of an interacting many-body quantum system (*3*) and is observable using experimental techniques (*4*).

The quest for predicting the electronic structure of matter dates back to Thomas (*5*), Fermi (*6*), and also Dirac (*7*) who formulated the very first theory in terms of electron density distributions. While computationally cheap, their theory was not useful for chemistry or materials science due to its lack of accuracy, as pointed out by Teller (*8*). Subsequently, based on a mathematical existence proof (*3*), the seminal work of Kohn and Sham (*9*) provided a smart reformulation of the electronic structure problem in terms of modern density functional theory (DFT) that has led to a paradigm shift. DFT is nowadays by far the most widely used method for computing the electronic structure of matter. Due to its balance of accuracy and computational cost, DFT has revolutionized chemistry − with the Nobel Prize in 1998 to Kohn (*10*) and Pople (*11*) marking its breakthrough. With the advent of exascale high-performance computing systems, DFT continues reshaping computational materials science at an even bigger scale (*12*). Consequently, DFT also constitutes one of the largest computational loads on high-performance computing platforms.

While DFT is in principle exact, in practice the exchange-correlation functional needs to be approximated (*13*). Sufficiently accurate approximations do exist for useful applications, but the search for ever more accurate functionals that extend the scope of DFT is an active area of research (*14*). In the past decade, artificial intelligence methods have led to great advances in DFT (*15,16*). Most notably, machine learning (ML) has proven to be very useful in constructing highly accurate exchange-correlation functionals (*17,18*).

Despite this success, DFT calculations are hampered inherently due to their computational cost. The standard algorithm scales as the cube of system size, limiting routine calculations to

problems comprised of only a few hundred atoms. This is a fundamental limitation that has hampered computational studies in chemistry and materials science so far. Lifting the curse of their computational cost has been a long-standing challenge. Prior works have followed in the footsteps of Thomas and Fermi in attempting to find an orbital-free formulation of DFT (*19*). While large-scale calculations have been realized (*20*), further progress is hampered by the need for an accurate approximation to the kinetic energy density functional (*21*). Hitherto developed approximations work well only for systems with nearly-free electrons, and a systematic approach for finding a universal functional is not in sight. While still useful, this limits the scope of orbital-free DFT to a subclass of problems. Similarly, prior work has attempted to speed up conventional DFT calculations by algorithmic development, an approach known as linear-scaling DFT. This includes divide-and-conquer methods (*22*) and techniques that rely on the Fermi operator expansion (*23*). Linear-scaling methods are ill-defined for metals, and their implementations suffer from computational instabilities (*24*). Despite some progress, this route has not led to large-scale electronic structure calculations that are generally applicable as well. More recently, other works have explored leveraging ML techniques to circumvent the inherent bottleneck of the DFT algorithm. These have used kernel-ridge regression (*25*) or neural networks (*26*), but remained on the conceptual level. Despite all these efforts, scaling DFT calculations to very large numbers of atoms while maintaining high accuracy has remained an elusive goal so far.

Ultra-large scale electronic structure predictions with neural networks

In this work, we circumvent the computational bottleneck of DFT calculations by utilizing neural networks in local atomic environments to predict the local electronic structure (*27, 28*). Thereby, we achieve the ability to compute the electronic structure of matter at ultra-large scales with minimal computational effort and at the accuracy of conventional DFT. Our workflow that achieves this is shown in Fig. 1**A**.

To this end, we train a feed-forward neural network M that performs a simple mapping

$$\tilde{d}(\epsilon, \boldsymbol{r}) = M(B(J, \boldsymbol{r})) \ , \tag{1}$$

where the bispectrum coefficients B of order J serve as *descriptors* that encode the positions of atoms relative to every point in real space $\boldsymbol{r}$, while $\tilde{d}$ approximates $d = \sum_j \phi_j^*(\boldsymbol{r})\phi_j(\boldsymbol{r})\,\delta(\epsilon-\epsilon_j)$, the local density of states (LDOS) at energy ϵ with ϕ_j denoting the Kohn-Sham eigenstates and ϵ_j the Kohn-Sham eigenenergies of DFT. The LDOS encodes the local electronic structure at each point in real space and energy. The key point is that the neural network is trained locally on a given point in real space and therefore has no awareness of the system size. Our underlying working assumption relies on the nearsightedness of the electronic structure (*29*). It sets a characteristic length scale beyond which effects on the electronic structure decay rapidly with distance. If this corresponds to the length scale of the training data, highly accurate ML predictions can be expected on ultra-large scales. Since the mapping defined in Eq. (1) is purely local, i.e., performed individually for each point in real space, the resulting workflow is scalable across the real-space grid, highly parallel, and transferable to different system sizes.

We illustrate our workflow by computing the electronic structure of sample material that contains more than 100,000 atoms. The employed ML model is a feed-forward neural network that is trained on simulation cells containing 256 Beryllium atoms. The network architecture is chosen similarly to earlier work (*28*). The underlying open-source software framework is developed as the Materials Learning Algorithms (MALA) package (*30*). In Fig. 1**B**-**D**, we showcase how our framework predicts multiple observables at previously unattainable scales. The corresponding atomic snapshots containing 131,072 Beryllium atoms are thermalized to 298K using the LAMMPS code (*31*) and the spectral neighborhood analysis potential (*32*) given in Ref. (*33*). Afterward, a stacking fault is introduced into the cell. It shifts three atomic layers and their local geometries from the hcp to the fcc crystal structure (see Fig. 1**B**). Our MALA model is then used to predict both the electronic densities and energies of this simulation cell with and without the stacking fault. As expected, MALA predictions reflect the changes to the electronic density due to the changes in the atomic geometry (see Fig. 1**C**). This allows us to investigate the energetic difference induced by the changes in the electronic structure. The energy differences are supposed to follow a $\sim N^{-\frac{1}{3}}$ behavior, where N is the number of atoms. We assess this by computing these energy differences for similarly generated structures of 256, 2048, and 16,384 atoms. We illustrate in Fig. 1**D** that this power law in system size is indeed satisfied. It is important to emphasize that the absolute energies reported here are not competitive with models or potentials specifically trained to reproduce stacking faults since no stacking fault structures were included in our training set. Instead, the results shown in Fig. 1**B**-**D** exemplify that MALA models are capable of predicting the spatially resolved electronic structure and related physical phenomena on a previously unattainable spatial scale. They open up the possibility to train models for specific applications, such as crystal structures that are perturbed on a large spatial scale, which was far beyond the reach of any electronic structure method. Our ML predictions on the 131,072 atom system take 48 minutes on 150 standard CPUs; the resulting computational cost of roughly 121 CPU hours (CPUh) is comparable to a conventional DFT calculation that, however, would be able to provide results for a system with only a few hundred atoms.

The computational cost of our ML workflow is orders of magnitude below currently existing linear-scaling DFT codes, i.e., codes scaling with $\sim N$ (*34*). Such calculations enable first-principles calculations of similarly extended systems by drawing on approximations in terms of the density matrix. Their computational cost is $\sim 10^4$ CPUh, i.e., two orders of magnitude above our approach. Standard DFT codes scale even more unfavorably as $\sim N^3$, which renders simulations like the one presented here completely infeasible.

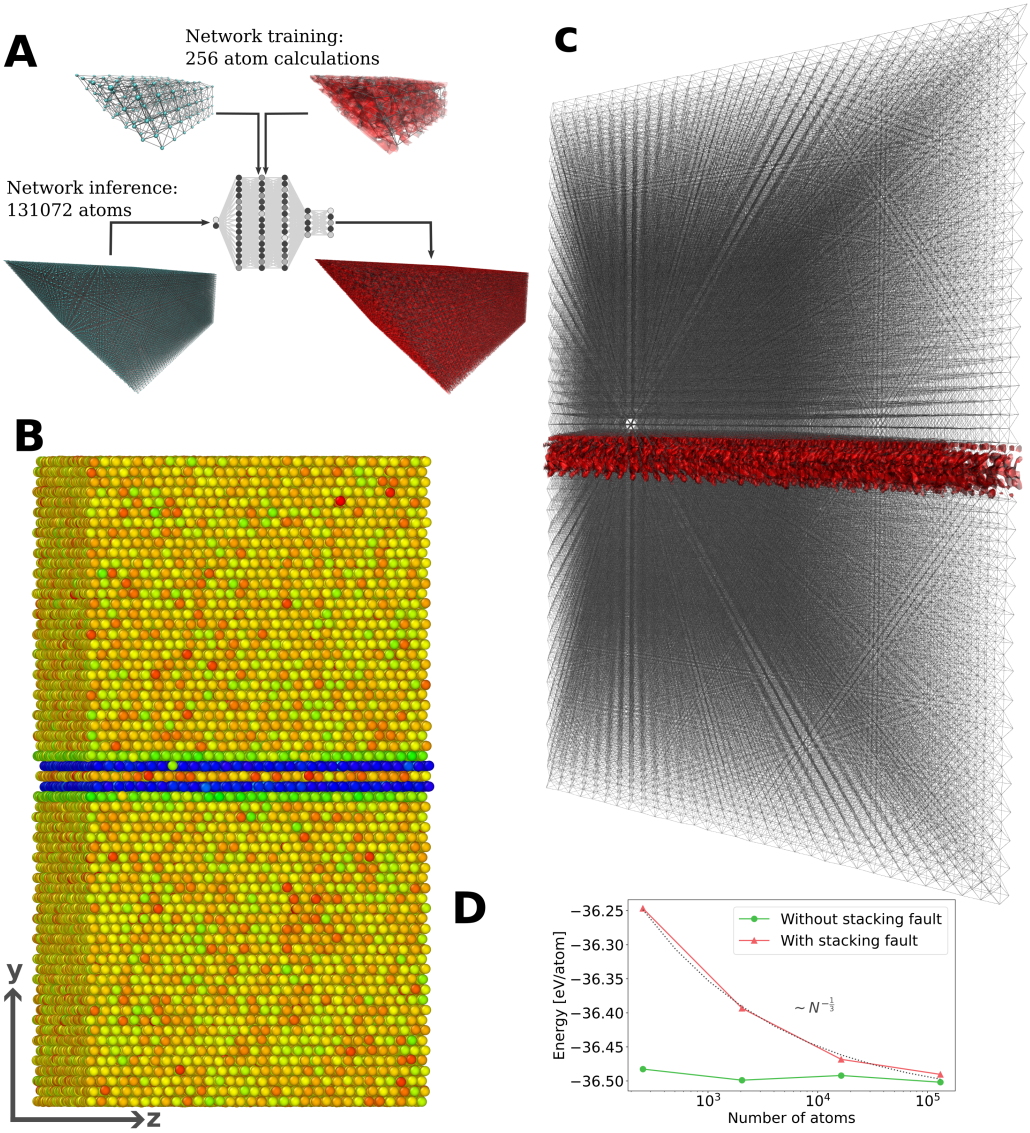

Figure 1: Visualization of electronic structures on the ultra-large scale and at DFT accuracy. **A**: General workflow. **B**: Stacking fault created in a cell with 131,072 Beryllium atoms. The stacking fault is generated by shifting three layers along the y-axis creating a local fcc geometry, as opposed to the hcp crystal structure of Beryllium. The colors correspond to the centrosymmetry parameter calculated by OVITO (*35*), which was also used for the visualization, where blue corresponds to fcc and red-to-light-green to hcp local geometries. Pictured here is the y-z plane. **C**: Difference in the electronic density for 131,072 Beryllium atoms with and without a stacking fault, pictured roughly along the y-z plane. **D**: Energy differences due to introducing a stacking fault into Beryllium cells of differing sizes. The calculated energies reproduce the expected scaling behavior with respect to the number of atoms.

In comparison to atomistic molecular dynamics simulations based on ML interatomic potentials or property mappings, our MALA models are not limited to single observables. ML models for predicting energies and forces of extended systems are well-established in the computational materials science community and are used frequently (*33*), yet they are limited to specific predictions. The same applies to models that predict application-specific observables such as polarizabilities (*36*) across a vast chemical space. Our model is able to perform predictions that give insight into most importantly the electronic structure, from which the energetics follow. In addition to the electronic density and the total energy showcased in Fig. 1, our models provide access to the density of states, the band energy, and the LDOS directly. All of them are important quantities for investigating the electronic structure of a material and are used to calculate relevant observables. Being able to calculate the total free energy also enables Monte-Carlo simulations for sampling thermodynamic observables. Furthermore, with the aid of (automatic) differentiation, our ML model can yield atomic forces for molecular dynamics simulations; in either case, finite-size effects can be minimized through the sheer size of the simulations that are now possible with our ML workflow.

The utility of our ML framework for chemistry and materials science relies on two key aspects. It needs to scale well with system size up to the 100,000 atom scale and beyond. Furthermore, it also needs to maintain accuracy as we run inferences on increasingly large systems. Both of these issues are addressed in Fig. 2.

Computational scaling

The computational cost of conventional DFT calculations is well known to scale proportional to N^3, where N is the system size, i.e., the number of atoms. Improved algorithms can enable an effective $\sim N^2$ scaling in certain cases over certain size ranges (*37*). In either case, one is faced with an increasingly insurmountable computational cost for systems involving more than a few thousand atoms. As can be seen in Fig. 2**A**, Quantum ESPRESSO (*38*), a popular DFT software package we employed for training data generation, is subject to this scaling behavior. The cost of MALA models on the other hand grows linearly with the number of atoms and has a significantly smaller computational overhead, to begin with. Up to atom numbers for which DFT simulations were computationally possible, we observe speed-ups of up to three orders of magnitude.

MALA model inference consists of three steps (*39*). First, the descriptor vectors are calculated on a real-space grid, then the LDOS is computed using a pre-trained neural network for given input descriptors, and finally, the LDOS is post-processed to compute electronic densities, total energies, and other observables. The first two parts of this workflow scale with $\sim N$, since they strictly perform operations per grid point, and the real space simulation grid grows linearly with N. Obtaining linear scaling for the last part of the workflow, which includes processing the electronic density to the total free energy, is less trivial and requires a few custom routines (*39*).

Accuracy and transferability to large scales

When assessing the transferability of our workflow, we are faced with the problem that we

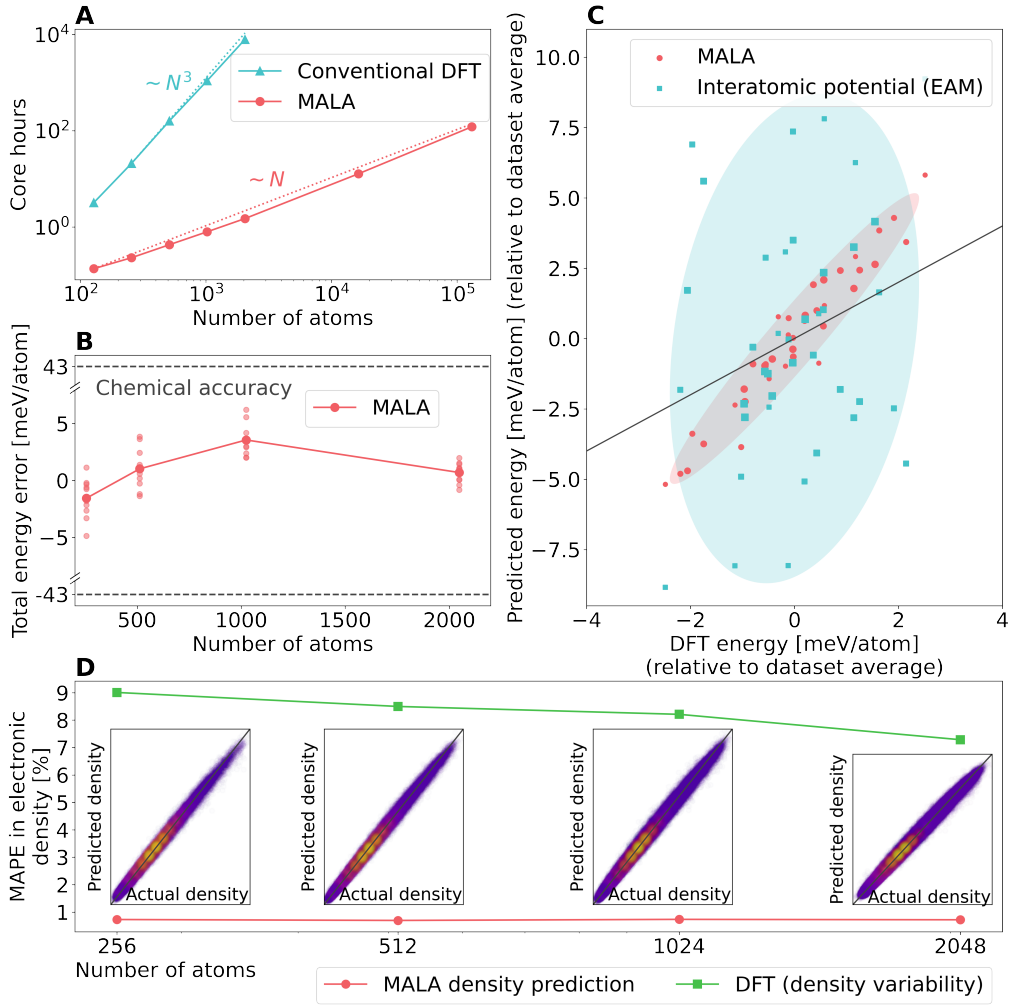

Figure 2: Assessment of MALA computational properties with respect to transferability of system size. **A**: Comparison of the scaling behavior of conventional DFT (QuantumESPRESSO) and the MALA framework with the number of atoms. Please note that for the sake of consistency, slightly different computational parameters have been used for the DFT calculations here compared to the DFT reference calculations in plots B-D. **B**: Total energy differences on the respective test sets (10 structures per number of atoms/element) when comparing DFT and MALA. **C**: Correlation between DFT and predicted total energies for MALA and an EAM type interatomic potential for Beryllium (across all system sizes). All energy values are given relative to the average in the respective data set. The colored surfaces visualize the 95% confidence interval for the respective distributions. Marker size correlates with system size and the grey line corresponds to ideal prediction. **D**: Mean Absolute Percentage Error (MAPE) of MALA density prediction (red), in relation to the general density variability across atomic configurations (green). The insets show the correlation between the predicted and actual density at different system sizes, with lighter colors indicating larger occurrences.

cannot compute conventional DFT results beyond about 2,000 atoms due to the high cost of these calculations. We, therefore, split the task of evaluating the predictive performance of our model into two achievable steps. We first show that our model does not lose accuracy when performing inference above system sizes the model was trained on, but still within an atom count for which we can compute DFT reference results. In the second step, we provide evidence for transferability to large systems by analyzing the resulting radial distribution function for atom counts beyond the reach of the DFT reference data. In this manner, we demonstrate that our model is consistently used in a strictly interpolative fashion and that it yields accurate results even for hundreds of thousands of atoms.

Benchmarks at DFT scales ($\sim 10^3$ atoms)

We tackle the first part of this problem by investigating a system of Beryllium atoms at room temperature and ambient mass density (1.896 g/cc). Neural networks are trained on LDOS data generated for 256 atoms. After training, the inference was performed for an increasing number of atoms, namely 256, 512, 1024, and 2048 atoms.

For 10 atomic configurations per system and simulation size, the LDOS predictions have been used to calculate total free energies. Naively, one simply assesses the (mean) errors observed which are shown Fig. 2**B**. It is evident that the errors stay roughly constant across system size and are well within chemical accuracy (below 43 meV/atom). They are in fact even below 10 meV/atom which is the gold standard of ML interatomic potentials. Yet, such an analysis fails to separate systematic from unsystematic errors. The total free energy does not hold much physical virtue by itself; in practice, one is interested in energy differences. Thus, even errors exceeding 10 meV/atom may be unproblematic, given that they are related to a systematic shift in energy, i.e., a model that consistently over- or underestimates energies.

To this end, Fig. 2**C** relates the predicted total free energy to the DFT reference data set. The data points are drawn across all system sizes but are given relative to the respective means per system size for the sake of readability. Ideally, the resulting distribution would lie along a straight line. In practice, both a certain spread around this line as well as a tilt of the line can be expected. The latter is related to a systematic shift in energy, as observed in Fig. 2**B**, whereas the former represents unsystematic errors. To quantify the impact of such a shift, we compare MALA (red circles) with an embedded-atom-method (EAM) interatomic potential (blue squares) (*40*). EAM potentials are commonly used for running classical molecular dynamics simulations. The EAM reference data is computed based on the interatomic potential from Ref. (*41*). As illustrated in Fig. 2**C**, our MALA model overall exhibits both a smaller shift and a smaller spread than this commonly used potential. Thus, MALA models not only exhibit smaller random errors, they further experience a smaller constant shift in energy. MALA models, therefore, deliver physically correct energies and are suitable for extended simulations at ultra-large scales.

Equally important is the transferable accuracy of the electronic structure predictions with our MALA model. The most important quantity is the spatially resolved electronic density. It

is calculated by integrating the predicted LDOS over the energy as

$$n(\boldsymbol{r}) = \int d\epsilon \; f^{\tau}(\epsilon) d(\epsilon, \boldsymbol{r}) \; , \tag{2}$$

where $f^{\tau}(\epsilon)$ denotes the Fermi-Dirac distribution at temperature τ. In an ideal DFT calculation, this density would be identical to the density that stems from the Hamiltonian of the interacting many-body problem. We benchmark the predictions of our MALA model in Fig. 2**D**. Analogous to the total energy, reference data can only be calculated using DFT up to 2,048 Beryllium atoms. We compute the mean absolute percentage error (MAPE) for a single configuration per simulation size. Across all system sizes, the MAPE lies below 1% (red curve). We put this value into perspective by showing the density variability (green curve) which is obtained from calculating the MAPE for densities of two different atomic configurations at the same temperature. The MAPEs calculated in this way correspond to roughly 8%, i.e., one order of magnitude above the inference error MALA incurs. The density predicted by MALA can therefore be assumed to be correct, which becomes even more evident when directly visualizing the correlation of actual and predicted density values on each grid point in the insets of Fig. 2**D**. It is evident that erroneous predictions mostly occur only for fringe regions of the density, and the overall density is predicted rather accurately on the real space grid.

Accuracy at ultra-large scales ($\sim 10^5$ atoms)

Finally, we tackle the second step of providing evidence that MALA predictions on the ultra-large scale are expected to be as accurate as conventional DFT calculations. This analysis is grounded in the local nature of our workflow. Given that the local environments are similar to those observed in training, predictions for arbitrarily large cells boil down to interpolation, a task at which neural networks excel. Accordingly, our ML model performs a perceived size extrapolation by actually performing local interpolations.

Determining whether a model operates in an interpolative regime requires the quantification of uncertainties, which is a very active topic of research in the ML community. The task at hand is to verify whether a prediction is based on input data that is drawn from the same distribution as the training set. Here, we rely on a physics-informed approach by comparing the radial distribution functions of atomic snapshots that stem from the training, inference, and ultra-large prediction data sets. The atomic radial distribution function is a useful quantity that distinguishes between different phases of a material. It gives insight into how likely it is to find an atom at a given distance from a reference point (*39*). We use it to confirm that the input data for the ultra-large system is drawn from the same distribution the model was trained on. Since the input to our workflow, B, is calculated based on atomic densities drawn from a certain cutoff radius (*39*), a matching radial distribution function up to this point indicates that the individual input vectors B should on average be similar between simulation cells. This comparison is shown in Fig. 3**A** where the radial distribution functions $g(r)$ of the training (256 atoms, green), inference test (2,048 atoms, blue), and ultra-large prediction (131,072 atoms, red and orange) data sets are illustrated. Note that for better illustration the inference test and prediction curves

are shifted along the y-axis by a constant value of 0.2. Furthermore, the absolute difference $\Delta g(r)$ with respect to the training data set is illustrated in Fig. 3**B**. As expected, slight deviations are apparent in the simulation cell containing the stacking fault (orange), but generally, all radial distribution functions agree very well up to the cutoff radius (dotted black). This analysis hence provides evidence that training (256 atoms), inference test (up to 2,048 atoms), and the ultra-large (131,072 atoms) simulation cells possess local environments sampled from the same distribution. It indicates that our MALA predictions of the electronic structure and energy are expected to be accurate at ultra-large scales.

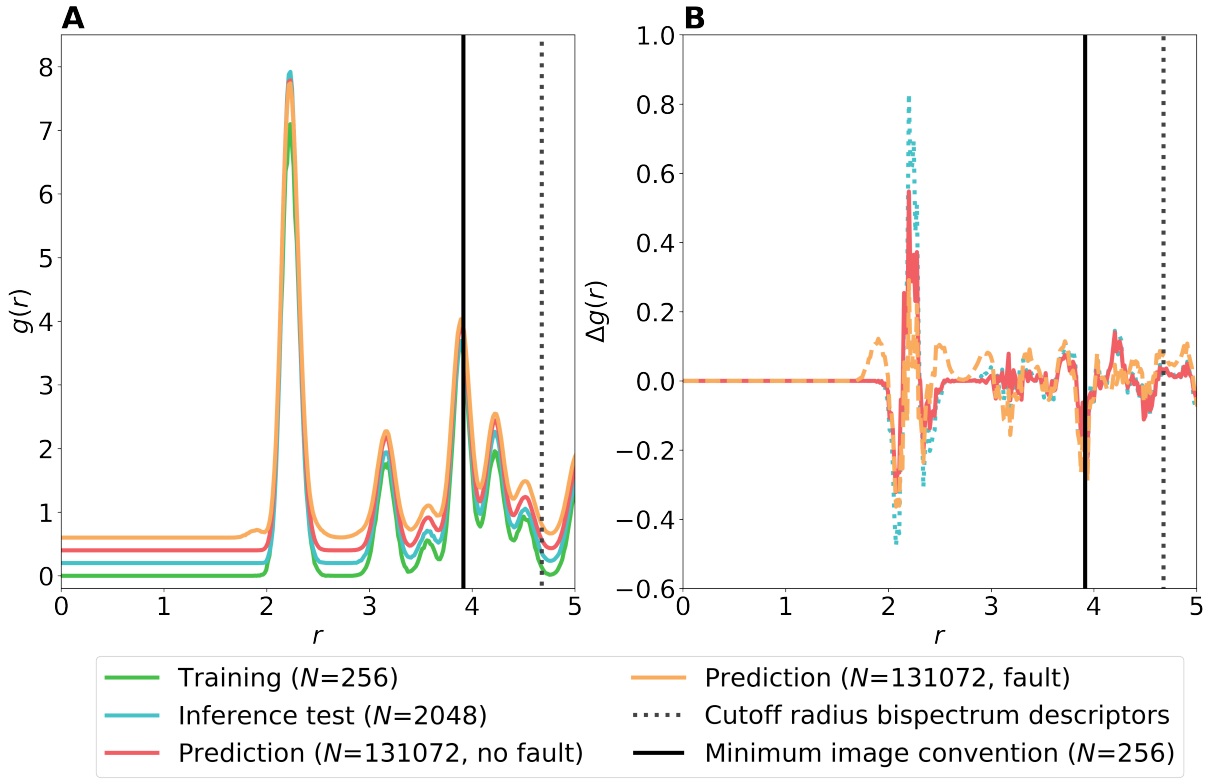

Figure 3: Analysis of size transferability of presented input data. **A**: Radial distribution functions for Beryllium cells of differing sizes, within the radius in which information is incorporated into bispectrum descriptors. The minimum image convention is the maximum radius for which the radial distribution function can be physically correct in a certain cell. It is added here because, for 256 atoms, the chosen cutoff radius slightly exceeds this radius. Note that the curves of the inference test (blue) and prediction (red, orange) data sets have been shifted along the y-axis by a constant value of 0.2 to better illustrate how similar they are. **B**: Absolute difference $\Delta g(r)$ of the radial distribution functions with respect to the training data set. It demonstrates the similarity of the ultra-large data set (red) with the training data set.

Conclusion

We have introduced an ML model that avoids the computational bottleneck of DFT calculations. Our ML model scales linearly with system size as opposed to conventional DFT calculations that follow a cubic scaling. Our ML model enables efficient electronic structure predictions at scales far beyond what is tractable with conventional DFT. In contrast to interatomic potentials, our workflow is not limited to certain observables but reproduces the functionality of DFT calculations to a large capacity by providing direct access to the electronic structure. At system sizes where DFT benchmarks are still available, we demonstrate that our ML model is capable of reproducing energies and electronic densities of extended systems at virtually no loss in accuracy. We, furthermore, show that our ML workflow enables predicting the electronic structure for systems with more than 100,000 atoms at a very low computational cost. We underpin its accuracy at these ultra-large scales by analyzing the radial distribution function.

We expect our ML model to set new standards in materials science and computational chemistry in a number of ways. Using such ML models either directly or in conjunction with other ML workflows, such as ML interatomic potentials for pre-sampling of atomic configurations, will enable first-principles modeling of a range of materials without finite-size errors. Combined with Monte-Carlo sampling and atomic forces from automatic differentiation, our ML model can furthermore replace ML interatomic potentials and yield thermodynamic observables at much higher accuracy. Another promising application domain our ML models enable is the simulation of electronic densities in semiconductor devices, for which a sufficiently accurate modeling capability at the device scale has been notoriously lacking. Finally, we also expect our ML model to pave the way to predicting electronic phase transitions on a quantitative level as it resolves changes in the electronic structure at hitherto unattainable length scales.

References and Notes

1. K. Kang, Y. S. Meng, J. Bréger, C. P. Grey, G. Ceder, *Science* **311**, 977 (2006).
2. R. T. Hannagan, *et al., Science* **372**, 1444 (2021).
3. P. Hohenberg, W. Kohn, *Phys. Rev.* **136**, B864 (1964).
4. P. N. H. Nakashima, A. E. Smith, J. Etheridge, B. C. Muddle, *Science* **331**, 1583 (2011)
5. L. H. Thomas, *Math. Proc. Camb. Philos. Soc.* **23**, 542 (1927).
6. E. Fermi, *Z. Phys.* **36**, 902 (1926).
7. P. A. M. Dirac, *Math. Proc. Camb. Philos. Soc.* **26**, 376 (1930).
8. E. Teller, *Rev. Mod. Phys.* **34**, 627 (1962).
9. W. Kohn, L. J. Sham, *Phys. Rev.* **140**, A1133 (1965).

10. W. Kohn, *Rev. Mod. Phys.* **71**, 1253 (1999).

11. J. A. Pople, *Rev. Mod. Phys.* **71**, 1267 (1999).

12. R. O. Jones, *Rev. Mod. Phys.* **87**, 897 (2015).

13. K. Lejaeghere, *et al., Science* **351**, aad3000 (2016).

14. M. G. Medvedev, I. S. Bushmarinov, J. Sun, J. P. Perdew, K. A. Lyssenko, *Science* **355**, 49 (2017).

15. L. Fiedler, K. Shah, M. Bussmann, A. Cangi, *Phys. Rev. Mater.* **6**, 040301 (2022).

16. R. Pederson, B. Kalita, K. Burke, *Nat. Rev. Phys.* **4**, 357 (2022).

17. J. C. Snyder, M. Rupp, K. Hansen, K.-R. Müller, K. Burke, *Phys. Rev. Lett.* **108**, 253002 (2012).

18. J. Kirkpatrick, *et al., Science* **374**, 1385 (2021).

19. V. L. Lignères, E. A. Carter, "An Introduction to Orbital-Free Density Functional Theory" in *Handbook of Materials Modeling* (Springer Netherlands, Dordrecht, 2005), pp. 137–148.

20. L. Hung, E. A. Carter, *Chem. Phys. Lett.* **475**, 163 (2009).

21. Y. A. Wang, E. A. Carter, "Orbital-Free Kinetic-Energy Density Functional Theory" in *Handbook of Materials Modeling* (Springer Netherlands, Dordrecht, 2005), pp. 117–184.

22. W. Yang, *Chem. Phys. Lett.* **66**, 1438 (1991).

23. S. Goedecker, L. Colombo, *Chem. Phys. Lett.* **73**, 122 (1994).

24. D. R. Bowler, T. Miyazaki, *Rep. Prog. Phys.* **75**, 036503 (2012).

25. F. Brockherde, *et. al Nat. Commun.* **8** (2017).

26. M. Tsubaki, T. Mizoguchi, *Chem. Phys. Lett.* **125**, 206401 (2020).

27. A. Chandrasekaran, *et al., NPJ Comput. Mater.* **5**, 22 (2019).

28. J. A. Ellis, *et al., Phys. Rev. B* **104**, 035120 (2021).

29. W. Kohn, *Chem. Phys. Lett.* **76**, 3168 (1996).

30. A. Cangi, *et al.*, Software publication 'MALA', DOI: 10.5281/zenodo.5557254 (2021).

31. A. P. Thompson, *et al., Comp. Phys. Comm.* **271**, 108171 (2022).

32. A. P. Thompson, L. P. Swiler, C. R. Trott, S. M. Foiles, G. J. Tucker, *J. Comput. Phys.* **285**, 316 (2015).

33. M. A. Wood, M. A. Cusentino, B. D. Wirth, A. P. Thompson, *Phys. Rev. B* **99**, 184305 (2019).

34. A. Nakata, *et al.*, *J. Chem. Phys.* **152**, 164112 (2020).

35. A. Stukowski, *Model. Simul. Mat. Sci. Eng.* **18**, 015012 (2009).

36. D. M. Wilkins, *et al., Proc. Natl. Acad. Sci. U.S.A.* **116**, 3401 (2019).

37. G. Kresse, J. Hafner, *Phys. Rev. B* **47**, 558 (1993).

38. P. Giannozzi, *et al.*, *J. Condens. Matter Phys.* **21**, 395502 (2009).

39. *Materials and methods are available as supplementary materials at the Science website*

40. M. S. Daw, M. I. Baskes, *Phys. Rev. B* **29**, 6443 (1984).

41. A. Agrawal, R. Mishra, L. Ward, K. M. Flores, W. Windl, *Model. Simul. Mat. Sci. Eng.* **21**, 085001 (2013).

42. M. P. Allen, D. J. Tildesley, *Computer simulation of liquids*, (Oxford University Press, 2017).

43. L. Fiedler, A. Cangi, Data set 'LDOS/SNAP data for MALA: Beryllium at 298K', DOI: 10.14278/rodare.1834 (2022).

44. L. Fiedler, *et al.*, Data set 'Scripts and Models for "Predicting the Electronic Structure of Matter on Ultra-Large Scales"', DOI: 10.14278/rodare.1851 (2022).

45. N. D. Mermin, *Phys. Rev.* **137**, A1441 (1965).

46. M. Born, R. Oppenheimer, *Ann. Phys.* **389**, 457 (1927).

47. M. Toda, R. Kubo, N. Saitō, *Statistical Physics: Equilibrium statistical mechanics*, (Springer, Berlin, 1983).

48. D. M. Ceperley, B. J. Alder, *Phys. Rev. Lett.* **45**, 566 (1980).

49. J. P. Perdew, W. Yue, *Phys. Rev. B* **33**, 8800 (1986).

50. J. P. Perdew, Y. Wang, *Phys. Rev. B* **45**, 13244 (1992).

51. J. P. Perdew, K. Burke, M. Ernzerhof, *Chem. Phys. Lett.* **77**, 3865 (1996).

52. J. Sun, A. Ruzsinszky, J. P. Perdew, *Phys. Rev. Lett.* **115**, 036402 (2015).

53. V. V. Karasiev, D. Chakraborty, O. A. Shukruto, S. B. Trickey, *Phys. Rev. B* **88**, 161108 (2013).

54. V. V. Karasiev, T. Sjostrom, J. Dufty, S. B. Trickey, *Chem. Phys. Lett.* **112**, 076403 (2014).

55. S. Groth, *et al., Chem. Phys. Lett.* **119**, 135001 (2017).

56. E. W. Brown, J. L. DuBois, M. Holzmann, D. M. Ceperley, *Phys. Rev. B* **88**, 081102 (2013).

57. R. Iftimie, P. Minary, M. E. Tuckerman, *Proc. Natl. Acad. Sci. U.S.A.* **102**, 6654 (2005).

58. M. A. Wood, A. P. Thompson, *J. Chem. Phys.* **148**, 241721 (2018).

59. M. A. Cusentino, M. A. Wood, A. P. Thompson, *J. Phys. Chem. A* **124**, 5456 (2020).

60. L. Fiedler, *et al., arXiv preprint arXiv:2202.09186* (2022).

61. K. Hornik, *Neural. Netw.* **4**, 251 (1991).

62. M. Minsky, S. A. Papert, *Perceptrons. An Introduction to Computational Geometry* (MIT Press, Cambridge, MA, 2017)

63. F. Rosenblatt, "The Perceptron: A Perceiving and Recognizing Automaton (Project PARA).", *Tech. Rep. 85-460-1*, (Cornell Aeronautical Laboratory, 1957).

64. D. E. Rumelhart, G. E. Hinton, R. J. Williams, *Nature* **323**, 533 (1986).

65. I. Goodfellow, Y. Bengio, A. Courville, *Deep Learning* MIT Press, Cambridge, MA, (2016)

66. A. Paszke, *et al.,* paper presented at the Advances in Neural Information Processing Systems 32 (NeurIPS 2019), Vancouver, Canada, 08 December 2019

67. P. Giannozzi, *et al.*, *J. Chem. Phys.* **152**, 154105 (2020).

68. P. Giannozzi, *et. al J. Condens. Matter Phys.* **29**, 465901 (2017).

69. W. Smith, *CCP5 Info. Quart.* **30**, 35 (1989).

70. G. Kresse, J. Furthmüller, *Phys. Rev. B* **54**, 11169 (1996).

71. G. Kresse, J. Furthmüller, *Comput. Mater. Sci.* **6**, 15 (1996).

72. A. Dal Corso, *Comput. Mater. Sci.* **95**, 337 (2014).

73. P. E. Blöchl, *Phys. Rev. B* **50**, 17953 (1994).

74. G. Kresse, D. Joubert, *Phys. Rev. B* **59**, 1758 (1999).

75. L. Fiedler, *et al., arXiv preprint arXiv:2206.03754* (2022).

76. H. J. Monkhorst, J. D. Pack, *Phys. Rev. B* **13**, 5188 (1976).

Acknowledgments: The authors are grateful to the Center for Information Services and High Performance Computing [Zentrum für Informationsdienste und Hochleistungsrechnen (ZIH)] at TU Dresden for providing its facilities for high throughput calculations. We also gratefully acknowledge Alexander Debus for providing a CPU allocation on the taurus HPC system of ZIH at TU Dresden.

Funding: Sandia National Laboratories is a multimission laboratory managed and operated by National Technology & Engineering Solutions of Sandia, LLC, a wholly-owned subsidiary of Honeywell International Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525. This paper describes objective technical results and analysis. Any subjective views or opinions that might be expressed in the paper do not necessarily represent the views of the U.S. Department of Energy or the United States Government.

This work was in part supported by the Center for Advanced Systems Understanding (CASUS) which is financed by Germany's Federal Ministry of Education and Research (BMBF) and by the Saxon state government out of the State budget approved by the Saxon State Parliament.

Author contributions: L.F. performed all Beryllium-related calculations (DFT-MD, DFT, and MALA), code integration into the MALA code, and data visualization. N.M. and D.V. implemented the parallelization of the total energy evaluation, and N.M. eliminated scaling bottlenecks in the total energy evaluation. S.S. carried out the transferability analysis. A.T. developed the parallelization of the descriptor calculation. S.R. and A.C. contributed to the theory and development of the MALA framework, supported data analysis, and supervised the overall project. All authors contributed to writing the manuscript.

Competing interests: There are no competing interests.

Data and materials availability: All calculations described within this work have been carried out with the freely available MALA code (*30*) version 1.1.0. Training data and benchmark models of the Beryllium system are provided in Ref. (*43*). The corresponding input scripts can be obtained via Ref. (*44*). This includes data discussed in the supplementary materials. Please note that Ref. (*43*) includes a larger data set of the Beryllium system as it has been used in multiple publications. In this manuscript, only a subset of Beryllium configurations has been used (256 atoms: 0,2,10-19; 512 atoms: 5-14; 1024 atoms: 0-9; 2048 atoms: 0-9).

Supplementary materials

Materials and Methods

Figures S1-S2

Table S1

Supplementary Materials for "Predicting the Electronic Structure of Matter on Ultra-Large Scales"

Lenz Fiedler,[1] Normand Modine,[2] Steve Schmerler,[3]
Dayton J. Vogel,[2] Gabriel A. Popoola,[4] Aidan Thompson,[5]
Sivasankaran Rajamanickam,[5*] Attila Cangi[1*]

[1] Center for Advanced Systems Understanding, Helmholtz-Zentrum Dresden-Rossendorf,
Untermarkt 20, Görlitz, 02826, Saxony, Germany
[2] Computational Materials and Data Science, Sandia National Laboratories,
1515, Eubank Blvd, Albuquerque, 87123, NM, USA
[3] Information Services and Computing, Helmholtz-Zentrum Dresden-Rossendorf,
Bautzner Landstraße 400, Dresden, 01328, Saxony, Germany
[4] Elder Research, Inc.
300 West Main Street, Charlottesville, 22903 VA, USA
[5] Center for Computing Research, Sandia National Laboratories,
1515, Eubank Blvd, Albuquerque, 87123, NM, USA
*To whom correspondence should be addressed; E-mail: srajama@sandia.gov, a.cangi@hzdr.de

October 20, 2022

This PDF file includes:

Materials and Methods

1 Density Functional Theory

Density functional theory (DFT) is the most widely used method for computing (thermodynamic) materials properties in chemistry and materials science because it strikes a balance between computational cost and accuracy. Within DFT, one commonly seeks to describe a coupled system of N ions of charge Z_α at collective positions $\underline{\boldsymbol{R}} = (\boldsymbol{R}_1, \boldsymbol{R}_2, ..., \boldsymbol{R}_N)$ and L electrons at collective positions $\underline{\boldsymbol{r}} = (\boldsymbol{r}_1, \boldsymbol{r}_2, ..., \boldsymbol{r}_L)$ on a quantum statistical mechanical level (*3,45*). Within the commonly assumed Born-Oppenheimer approximation (*46*), the Hamiltonian

$$\hat{H} = \hat{T} + \hat{V}^{ee} + \hat{V}^{ei} \,, \tag{1}$$

represents a system of interacting electrons in the external field of the ions that are simplified to classical point particles. Here, $\hat{T} = \sum_j^L -\nabla_j^2/2$ denotes the kinetic energy operator of the electrons, $\hat{V}^{ee} = \sum_i^L \sum_{j\neq i}^L 1/(2|\boldsymbol{r}_i - \boldsymbol{r}_j|)$ the electron-electron interaction, and $\hat{V}^{ei} = -\sum_j^L \sum_\alpha^N Z_\alpha/|\boldsymbol{r}_j - \boldsymbol{R}_\alpha|$ the external potential, i.e., the electron-ion interaction. The Born-Oppenheimer Hamiltonian separates the electronic and ionic problems into a quantum mechanical and classical mechanical problem. Such an assumption is feasible since ionic masses far exceed the electronic mass, leading to vastly different time scales for movement and equilibration.

At finite temperatures $\tau > 0$, the theoretical description is extended to the grand canonical operator (*47*)

$$\hat{\Omega} = \hat{H} - \tau\hat{S} - \mu\hat{N} \,, \tag{2}$$

where the $\hat{S} = -k_B \ln\hat{\Gamma}$ denotes the entropy operator, $\hat{N}$ the particle-number operator, and μ the chemical potential. Here, we introduced the statistical density operator $\hat{\Gamma} = \sum_{L,m} w_{L,m} |\Psi_{L,m}\rangle\langle\Psi_{L,m}|$ with the L-electron eigenstates $\Psi_{L,m}$ of the Hamiltonian $\hat{H}$ and $w_{L,m}$ as the normalized statical weights that obey $\sum_{L,m} w_{L,m} = 1$. Any observable A is then computed as an average

$$A[\hat{\Gamma}] = \mathrm{Tr}\{\hat{\Gamma}\hat{A}\} = \sum_{L,m} w_{L,m} \langle\Psi_{L,m}|\hat{A}|\Psi_{L,m}\rangle \,. \tag{3}$$

Most importantly, finding the grand potential

$$\Omega[\hat{\Gamma}] = \mathrm{Tr}\{\hat{\Gamma}\hat{\Omega}\} \tag{4}$$

amounts to finding a $\hat{\Gamma}$ that minimizes this expression.

The exact solution to this problem evades numerical treatment even with modern hardware and software due to the electron-electron interaction in the Born-Oppenheimer Hamiltonian of Eq. (1). It dictates an exponential growth of complexity with the number of electrons L, i.e., e^L.

Based on the theorems of Hohenberg-Kohn (*3*) and Mermin (*45*), DFT makes solving this problem tractable by employing the electronic density n as the central quantity. The formal scaling reduces to L^3 due to the Kohn-Sham approach (*9*). Within DFT, all quantities of interest are formally defined as functionals of the electronic density via a one-to-one correspondence with the external (here, electron-ion) potential. In conjunction with the Kohn-Sham scheme, which introduces an auxiliary system of non-interacting fermions restricted to reproduce the density of the interacting system practical calculations become feasible. Rather than evaluating Eq. 4 using many-body wave functions Ψ_L, the grand potential is expressed in terms of a functional of n as

$$\begin{aligned}\Omega[n] =& T_\mathrm{S}[\phi_j] - k_\mathrm{B}\tau S_\mathrm{S}[\epsilon_j] + E_\mathrm{H}[n] \\ &+ E_\mathrm{XC}[n] + E^{ei}[n] - \mu L \ ,\end{aligned} \tag{5}$$

with the Kohn-Sham wave functions and energy eigenvalues ϕ_j and ϵ_j, the kinetic energy of the Kohn-Sham system $T_\mathrm{S}[\phi_j]$, the entropy of the Kohn-Sham system $S_\mathrm{S}[\epsilon_j]$, the classical electrostatic interaction energy $E_\mathrm{H}[n]$ (Hartree energy), the electrostatic interaction energy of the electronic density with the ions $V^{ei}[n]$, and the exchange-correlation (free) energy $E_\mathrm{XC}[n]$. The Kohn-Sham system serves as an auxiliary system used to calculate the kinetic energy and entropy terms in Eq. (5). The Kohn-Sham equations are defined as a system of one-electron Schrödinger-like equations

$$\left[-\frac{1}{2}\nabla^2 + v_\mathrm{S}(\boldsymbol{r})\right]\phi_j(\boldsymbol{r}) = \epsilon_j\phi_j(\boldsymbol{r}) \ , \tag{6}$$

with an effective potential, the Kohn-Sham potential $v_\mathrm{S}(\boldsymbol{r})$, that yields the electronic density of the interacting system via

$$n(\boldsymbol{r}) = \sum_j f^\tau(\epsilon_j) \ \mid \phi_j(\boldsymbol{r}) \mid^2 \ , \tag{7}$$

with the Fermi-Dirac distribution $f^\tau(\epsilon_j)$ at temperature τ. Note that within the Kohn-Sham framework at finite temperatures, several quantities, namely, $v_\mathrm{S}(\boldsymbol{r})$, ϵ_j, ϕ_j, n, T_S, S_S and E_XC are technically temperature dependent; we omit to label this temperature dependency explicitly in the following for the sake of brevity. The Kohn-Sham formalism of DFT is formally exact if the correct form of the exchange-correlation functional $E_\mathrm{XC}[n]$ was known. In practice, approximations of the exchange-correlation functional are employed. There exists a plethora of useful functionals, both for the ground state (such as the LDA (*9, 48*), PBE (*49–51*), and SCAN (*52*) functionals) and at finite temperature (*53–56*). Such functionals draw on different ingredients for approximating the exchange-correlation energy. Some rely only on the electronic density (e.g. LDA), while others incorporate density gradients (e.g. PBE) or even the kinetic energy density (e.g. SCAN). Consequently, functionals differ in their provided accuracy and application domain like molecules or solids.

The calculation of dynamical properties is enabled in this framework via the estimation of the atomic forces, which are then used to time-propagate the ions in a process called DFT molecular dynamics (DFT-MD). The forces are evaluated via the total free energy, which is obtained

from Eq. (5), as $-\partial A[n](\underline{\boldsymbol{R}})/\partial\underline{\boldsymbol{R}}$ where $A^{\mathrm{BO}}_{\mathrm{total}}[n](\underline{\boldsymbol{R}}) = \Omega[n] + \mu L$. While this framework can be employed to calculate a number of (thermodynamic) materials properties (*57*), the treatment of systems of more than roughly a thousand atoms becomes computationally intractable due to the L^2 to L^3 scaling typically observed when running DFT calculations for systems in this size range. Therefore, current research efforts are increasingly focused on the combination of machine learning (ML) and DFT methods (*15*).

2 DFT surrogate models

ML comprises a number of powerful algorithms, that are capable of learning, i.e., improving through data provided to them. Within DFT and computational materials science in general, ML is often applied in one of two settings, as shown in Ref. (*15*). Firstly, ML algorithms learn to predict specific properties of interest (e.g. structural or electronic properties) and thus bypass the need to perform first-principles simulations for investigations across vast chemical parameter spaces. Secondly, ML algorithms may provide direct access to atomic forces and energies, and thus accelerate dynamical first-principles simulations drastically, resulting in ML interatomic potentials (ML-IAPs) for MD simulations.

We have recently introduced an ML framework that does not fall in either category, as it comprises a DFT surrogate model that replaces DFT for predicting a range of useful properties (*28*). Our framework directly predicts the electronic structure of materials and is therefore not restricted to singular observables.

Within this framework, the central variable becomes the local density of states (LDOS), defined via

$$d(\epsilon, \boldsymbol{r}) = \sum_j \mid \phi_j(\boldsymbol{r}) \mid^2 \delta(\epsilon - \epsilon_j) \ . \tag{8}$$

The merit of using the LDOS as a central variable is that it determines both the electronic density and the density of states (DOS) D via

$$D(\epsilon) = \sum_j \delta(\epsilon - \epsilon_j) = \int d\boldsymbol{r}\ d(\epsilon, \boldsymbol{r}) \ , \tag{9}$$

$$n(\boldsymbol{r}) = \sum_j f^\tau(\epsilon_j) \ \mid \phi_j(\boldsymbol{r}) \mid^2 = \int d\epsilon\ f^\tau(\epsilon) d(\epsilon, \boldsymbol{r}) \ . \tag{10}$$

These two quantities can be used to calculate the total free energy drawing on a reformulation of Eq. (5), which expresses all energy terms dependent on the KS wave functions and eigenvalues in terms of the DOS. More precisely, by employing the band energy

$$E_b = \int d\epsilon\ f^\tau(\epsilon) \epsilon D(\epsilon) \ , \tag{11}$$

and reformulating the electronic entropy in terms of D, i.e.,

$$\begin{aligned} S_{\mathrm{S}} &= -\sum_j \left[f_j^\tau(\epsilon_j) \ln f_j^\tau(\epsilon_j) + \left(1 - f_j^\tau(\epsilon_j)\right) \ln \left(1 - f_j^\tau(\epsilon_j)\right) \right] \ , \\ &= -\int d\epsilon \, \left(f^\tau(\epsilon) \ln \left[f^\tau(\epsilon) \right] + \left[1 - f^\tau(\epsilon)\right] \ln \left[1 - f^\tau(\epsilon)\right] \right) D(\epsilon) \ , \end{aligned} \tag{12}$$

the total free energy $A_{\mathrm{total}}^{\mathrm{BO}}$ can be expressed as

$$\begin{aligned} A_{\mathrm{total}}^{\mathrm{BO}}[d] =& E_b[D[d]] - \tau S_{\mathrm{S}}[D[d]] - E_{\mathrm{H}}[n[d](\boldsymbol{r})] \\ &+ E_{\mathrm{XC}}[n[d](\boldsymbol{r})] - \int d\boldsymbol{r} \, v_{\mathrm{XC}}(\boldsymbol{r}) n[d](\boldsymbol{r}) \ , \end{aligned} \tag{13}$$

where D and n are functionals of the LDOS and $v_{\mathrm{XC}}(\boldsymbol{r}) = \delta E_{\mathrm{XC}}[n[d](\boldsymbol{r})]/\delta n[d](\boldsymbol{r})$.

In our framework, the LDOS is learned *locally*. For each point in real space, the respective LDOS (a vector in the energy domain) is predicted separately from adjacent points. Non-locality enters this prediction through the *descriptors* that serve as input to the ML algorithm. Here, we chose descriptors based on the Spectral Neighborhood Analysis Potential (SNAP) (*32, 33, 58, 59*) method, that we will refer to as bispectrum descriptors B . In contrast to the usual application of SNAP, i.e., building interatomic potentials, these descriptors are employed to encode local information at each point in space. This is done by evaluating the total density of neighbor atoms via a sum of delta functions

$$\rho(\boldsymbol{r}) = \delta(\mathbf{0}) + \sum_{r_k < R_{cut}^{\nu_k}} f_c(|\boldsymbol{r}_k|, R_{cut}^{\nu_k}) w_{\nu_k} \delta(\mathbf{r}_k) \, . \tag{14}$$

In Eq. (14), the sum is performed over all k atoms within a cutoff distance $R_{cut}^{\nu_k}$ using a switching function f_c that ensures smoothness of atomic contributions at the edges of the sphere with radius $R_{cut}^{\nu_k}$. These atoms are located at position $\boldsymbol{r}_k$ relative to the grid point $\boldsymbol{r}$, while the chemical species ν_k enters the equation via the dimensionless weights w_{ν_k}. The thusly defined density is then expanded into a basis of 4D hyperspherical harmonic functions, eventually yielding the descriptors $B(J, \boldsymbol{r})$, with the feature dimension J (see Ref. (*28, 32*)). Constructing descriptors in such a way introduces two hyperparameters, $R_{cut}^{\nu_k}$, which determines the radius from which information is incorporated into the descriptors and $J_{\max}$, which determines the number of hyperspherical harmonics used for the expansion, and therefore the dimensionality of the descriptor vectors. As we have shown recently in Ref. (*60*), they can be chosen accurately without the need for ML model training and inference based on similarity measures that agree with physical intuition.

A mapping from $B(J, \boldsymbol{r})$ to $d(\epsilon, \boldsymbol{r})$ is now performed via a neural network (NN), M, i.e.,

$$\tilde{d}(\epsilon, \boldsymbol{r}) = M(B(J, \boldsymbol{r})) \ , \tag{15}$$

where $\tilde{d}$ is the approximate LDOS. After performing such a network pass for each point in space, the resulting approximate LDOS can be post-processed into the observables mentioned above.

We employ NNs, because they are, in principle, capable of approximating any given function (*61*). In the present case, we employ feed-forward NNs (*62*), which consist of a sequence of layers containing individual artificial neurons (*63*) that are fully connected to each neuron in subsequent layers. Each layer is a transformation of the form

$$\boldsymbol{x}^{\ell+1} = \varphi(\mathbf{W}^{\ell}\boldsymbol{x}^{\ell} + \boldsymbol{b}^{\ell}) \tag{16}$$

that maps $\boldsymbol{x}$ from layer ℓ to $\ell + 1$ by addition of a bias vector $\boldsymbol{b}$, matrix multiplication with a weight matrix $\mathbf{W}$, and an *activation function* φ. For the DFT surrogate models discussed here, the input to the first transformation $\boldsymbol{x}^0$ is $B(J, \boldsymbol{r})$ for a specific point in space $\boldsymbol{r}$; the output of the last layer $\boldsymbol{x}^{\mathcal{L}}$ is $d(\epsilon, \boldsymbol{r})$ for the same $\boldsymbol{r}$. The number of layers $\mathcal{L}$ and activation function φ have to be determined through prior *hyperparameter optimization*, among other such *hyperparameters*, such as the width of the individual layers. In Ref. (*60*) we show how such a hyperparameter optimization can be drastically improved upon in terms of computational effort, while the hyperparameters employed for this study are detailed in Sec. 4. For each architecture of the NN, the weights and biases have to be optimized using gradient-based updates in a process called *training* based on a technique called *backpropagation* (*64*), which in practice is done using gradients averaged over portions of the data (so-called *mini-batches*). Other technical parameters include stopping criteria for the *early stopping* of the model optimization and the *learning rate* for the gradient-based updates. For more details on NNs see, e.g., Ref. (*65*).

The steps of combining the calculation of bispectrum descriptors to encode the atomic density, training and evaluation of NNs to predict from them the LDOS, and finally the post-processing of the LDOS to physical observables, is implemented as a DFT surrogate model workflow called Materials Learning Algorithms (MALA). The full workflow is visualized in Fig. S.1. We employ interfaces to popular open-source software packages, namely LAMMPS (*31*) (descriptor calculation), PyTorch (*66*) (NN training and inference), and Quantum ESPRESSO (*38, 67, 68*) (post-processing of the electronic density).

Unlike DFT, the great majority of operations in our DFT surrogate model have a computational cost that naturally scales linearly with system size: (1) the descriptors are evaluated independently at each point on the computational grid using algorithms in LAMMPS that take advantage of the local dependence of the descriptors on the atomic positions in order to maintain linear scaling; (2) the NN is evaluated independently at each grid point in order to obtain the LDOS at each point; (3) the DOS is evaluated by a reduction over grid points, the Fermi level is found, and E_b and S_{S} are evaluated; (4) the density is calculated independently at each grid point; and (5) three-dimensional Fast Fourier transforms (FFTs), which are implemented efficiently in Quantum ESPRESSO, are used to evaluate E_{H} from the density. The remaining terms are E_{XC}, v_{XC}, and the ion-ion interaction energy. The exchange-correlation terms can almost be evaluated independently at each point (using FFTs to evaluate gradients if necessary), but the pseudo-potentials that we use include non-linear core corrections, which require the addition of a "core density" centered on each atom to the density used to calculate E_{XC} and v_{XC}. Likewise, the ion-ion interaction energy can be evaluated efficiently using Fast Fourier transforms if we can compute the sum of non-overlapping charge distributions containing the appropriate ionic

charge centered on each atom. The key to calculating these terms with a computational cost that scales linearly with system size is an efficient algorithm to evaluate the structure factor.

If $F(\boldsymbol{r})$ is some periodic function represented by its values on the computational grid, its Fast Fourier transform $\tilde{F}(\boldsymbol{G})$ gives its representation in the basis of plane-waves $\exp(i\boldsymbol{G}\cdot\boldsymbol{r})$, where the reciprocal lattice vectors $\boldsymbol{G}$ form a reciprocal-space grid with the same dimensions as the computational grid. The structure factor is defined as

$$\tilde{S}(\boldsymbol{G}) = \sum_{\alpha} \exp(i\boldsymbol{G}\cdot\boldsymbol{R}_{\alpha}) \ , \tag{17}$$

where the summation over atom positions $\boldsymbol{R}_{\alpha}$ runs over all atoms within one copy of the periodically repeated computational cell. The structure factor is very useful since

$$F^{S}(\boldsymbol{r}) \equiv \sum_{\alpha} F(\boldsymbol{r}-\boldsymbol{R}_{\alpha}) \ , \tag{18}$$

can be efficiently evaluated as the inverse Fourier transform of $\tilde{F}^{S} = \tilde{S}(\boldsymbol{G})\tilde{F}(\boldsymbol{G})$. Thus, the structure factor can be used to evaluate the non-linear core correction density and the ion-ion interaction energy when evaluating the DFT total energy. However, the straightforward evaluation of $\tilde{S}(\boldsymbol{G})$ on the grid of $\boldsymbol{G}$ vectors scales as the square of the system size. We circumvent this bottleneck by taking advantage of the real-space localization properties of the Gaussian function $G(\boldsymbol{r})$ in order to efficiently evaluate $G^{S}(\boldsymbol{r})$ within the LAMMPS code (*31*). Then, within Quantum ESPRESSO, we use a fast Fourier transformation to calculate $\tilde{G}^{S}(\boldsymbol{G})$, and the structure factor is obtained as

$$\tilde{S}(\boldsymbol{G}) = \frac{\tilde{G}^{S}(\boldsymbol{G})}{\tilde{G}(\boldsymbol{G})} \ . \tag{19}$$

A suitable choice of the Gaussian width for $G(\boldsymbol{r})$ allows us to minimize aliasing errors due to Fourier components beyond the Nyquist limit of the computational grid, while also maintaining good precision in the above division.

3 Data analysis

We assess whether our models are employed in an interpolative setting when applied to larger cells.

To that end, we analyze the radial distribution function (RDF), which is defined as

$$g(r) = \frac{\Xi(r)}{4\rho\pi dr\left(r^2 + \frac{dr^2}{12}\right)}\,, \tag{20}$$

where $\Xi(r)$ is the average number of ions in a sphere of radius $r + dr$ around each ion, normalized by the atomic density $\rho = N_\mathrm{i}/V_\mathrm{cell}$ (*42*). The RDF is often used to identify different phases of a material and, in our case, it can be used to verify that simulation cells with differing numbers of atoms are equivalent in their ion distribution up to a certain cutoff radius. For technical purposes, there exists an upper radius of the RDF up to which g is well-defined, referred to as the minimum image convention (MIC) (*69*). For small cells, the employed cutoff radius lies slightly beyond this radius, but this does not affect model inference, since periodic boundary conditions are applied for the calculation of the bispectrum descriptors.

4 Computational details

4.1 Training Data

Increasingly larger DFT-MD simulations at 298K have been performed to acquire atomic configurations for simulation cells of 256 to 2048 beryllium atoms. DFT-MD calculations up to 512 atoms have been carried out using Quantum ESPRESSO, while simulations for 1024 and 2048 atoms have been performed using VASP (*37, 70, 71*). In either case, DFT-MD simulations have been performed at the Γ-point, using a plane-wave basis set with an energy cutoff of 40 Ry (Quantum ESPRESSO) or 248 eV (VASP), and an ultrasoft pseudopotential (*72*) (Quantum ESPRESSO) or a PAW pseudopotential (*73, 74*) (VASP). The resulting trajectories have been analyzed with a method akin to the equilibration algorithm outlined in Ref. (*75*), although here equilibration thresholds have been defined manually. Thereafter, snapshots have been sampled from these trajectories such that the minimal euclidean distance between any two atoms within the last sampled snapshot and potentially next sampled snapshots lies above the empirically determined threshold of 0.125 Å. The resulting data set of beryllium at room temperature includes ten configurations per system size, except for 256 atoms, where a larger number of configurations is needed to enable the training and verification of models. For all of these configurations, DFT calculations have been carried out with Quantum ESPRESSO, using the aforementioned cutoff and pseudopotential. The Brillouin zone has been sampled by Monkhorst-Pack (*76*) sampling, with the number of $\boldsymbol{k}$-points given in Tab. S1.

The employed calculation parameters have been determined via a convergence analysis with a threshold of 1 meV/atom, except for beryllium systems with 2048 atoms, where only Γ point calculations have been performed due to computational constraints.

The values in Tab. S1 refer to those DFT calculations that were performed to gather reference energies and densities. To calculate the LDOS, one has to employ larger $\boldsymbol{k}$-grids, as the discretiatzion of $\boldsymbol{k}$-space with a finite number of points in $\boldsymbol{k}$-space, can introduce errors and features into the (L)DOS that can be unphysical in nature. As has been discussed in Ref. (*28, 60*), such features can be removed by employing a larger number of $\boldsymbol{k}$-points than for typical DFT simulations. The correct $\boldsymbol{k}$-grid has to be determined through a convergence test such that no unphysical oscillations occur in the (L)DOS. By applying an analysis as outlined in Ref. (*28,60*) we have determined $12 \times 6 \times 6$ as a suitable $\boldsymbol{k}$-grid for 256 beryllium atoms. Similar to the DFT calculation for providing reference energies, Monkhorst-Pack sampling has been used.

In order to assess the scaling of DFT for Fig. 2**A** of the main manuscript, we kept a constant $\boldsymbol{k}$-grid were possible, in comparison to the adapted $\boldsymbol{k}$-grids used for the reference data calculation used for Fig. 2**B-D**. More specifically, in order to reflect realistic simulation settings, we employed a $3 \times 3 \times 3$ grid, i.e., a $\boldsymbol{k}$-grid consistent with 1024, the largest number of atoms for which $\boldsymbol{k}$-point converged simulations could be performed. The same number of $\boldsymbol{k}$-points was used for 128, 256 and 512 atoms. For 2048 performing a DFT simulation was impossible with the computational ressources available to us, due to large memory demand. We therefore performed the 2048 atom calculation with a $4 \times 2 \times 2$ $\boldsymbol{k}$-grid, utilizing more $\boldsymbol{k}$-points in the x-direction, since the 2048 atom cells are extended in that direction compared to the 1024 atom cells. Overall, this change in $\boldsymbol{k}$-grid leads to only a small deviation of the observed $\sim N^3$ behavior.

Since we have investigated the size transferability of MALA surrogate models, we have also examined the issue of minimal training data size. Since the main result of our work is that due to the local nature of the presented models, predictions can be made for large systems based on small ones, the question naturally arises how few atoms can included the training data without loss of accuracy. To investigate this, the same beryllium model as discussed for 256 beryllium atoms has also been trained for 128 beryllium atoms.

The result for this model can be seen in Fig. S.2**A**. The results for 256 atoms are the same as in the main manuscript, as is the methodology for either model. It is evident that a model trained on 128 beryllium atoms performs considerably worse. This difference in performance is physically motivated, as Fig. S.2**B** shows. Here, the DOS (as a more accessible representation of the LDOS) is visualized slightly below the Fermi energy for three atomic configurations with differing number of atoms. Despite a large number of $\boldsymbol{k}$-points being used in all cases ($11 \times 11 \times 11$ in case of 128 atoms and $8 \times 8 \times 4$ in case of 512) atoms, distinct features can be observed, especially between 2 and 4 eV. These features are different between different cell sizes, and seem to converge, i.e. be smoothened out as one goes to larger number of atoms. Further analysis shows that they are consistent between different configurations at the same temperature.

We therefore assume them to be physical in nature and present due to finite size effects. Since any ML model can, at its best, only reproduce data it has observed during training, the transferability of MALA models also depends on the correctness of the (L)DOS on which it was trained w.r.t. the (L)DOS of more extended systems. In the case of beryllium, the DOS of

256 atoms is a middleground between the DOS of 128 and 512 atoms, with the sharp double peak of the former slightly washed out towards the single peak of the latter. It is thus similar enough in shape to the DOS of larger systems to perform an accurate size transfer.

Since the analysis of the DOS is trivially possible compared to the repeated training of models of differing training data sizes, we have idenitified a straightforward way to determine the minimum number of atoms to be included in the training set.

4.2 Machine Learning Models

For all ML experiments, the architecture and hyperparameters discussed in Ref. (*28*) has been employed. One training and one validation snapshot has been used in the 256 atom case. All ML experiments have been carried out using the MALA code version 1.1.0 (*30*).

Figure S1

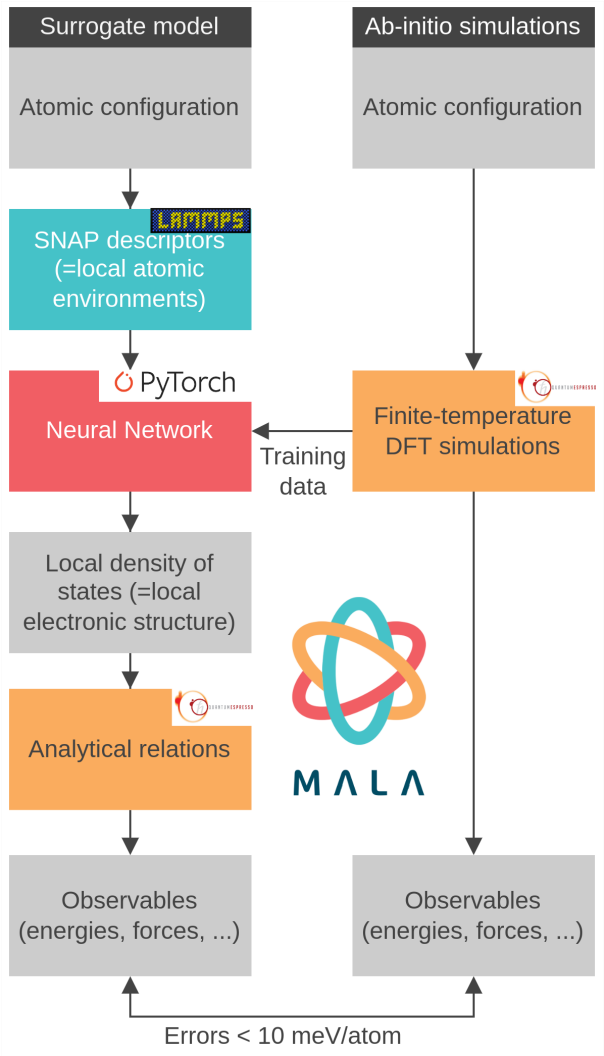

Figure S.1: Overview of the MALA framework.

Figure S2

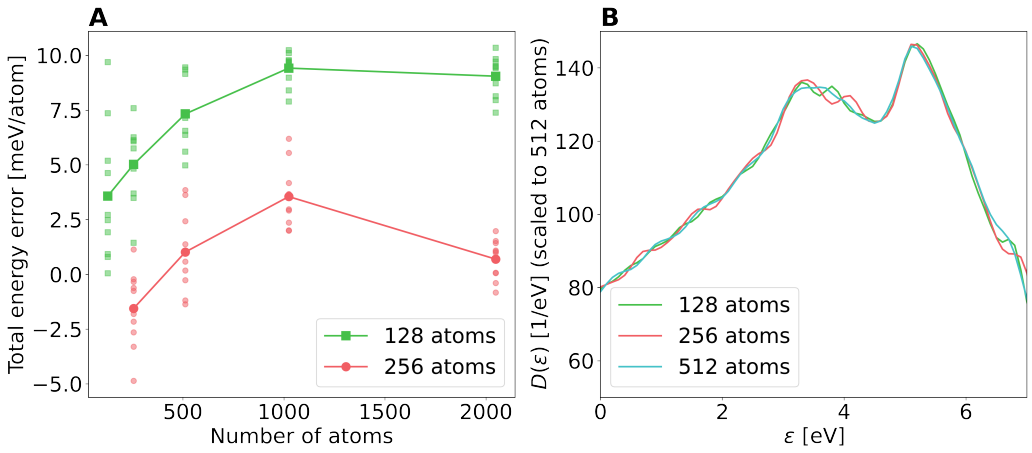

Figure S.2: Investigation of minimal suitable training data size for MALA. **A**: Inference accuracy for models trained on either 128 or 256 beryllium atoms. The inference parameters were kept the same as to the experiments presented in the main manuscript. **B**: Portion of the DOS $D(\epsilon)$ of beryllium configurations with 128, 256 and 512 atoms directly below the Fermi energy.

Table S1

Number of atoms	***k*-grid**
256	$8 \times 4 \times 4$
512	$4 \times 4 \times 2$
1024	$3 \times 3 \times 3$
2048	Γ-point

Table S1: Overview over the $\boldsymbol{k}$-grids used for the various DFT calculations.

Supplementary references

1. K. Kang, Y. S. Meng, J. Bréger, C. P. Grey, G. Ceder, *Science* **311**, 977 (2006).
2. R. T. Hannagan, *et al., Science* **372**, 1444 (2021).
3. P. Hohenberg, W. Kohn, *Phys. Rev.* **136**, B864 (1964).
4. S. Hüfner, *Photoelectron Spectroscopy: Principles and Applications*, Advanced Texts in Physics (Springer Berlin Heidelberg, 2013).
5. L. H. Thomas, *Mathematical Proceedings of the Cambridge Philosophical Society* **23**, 542–548 (1927).
6. E. Fermi, *Zeitschrift für Physik* **36**, 902 (1926).
7. P. A. M. Dirac, *Mathematical Proceedings of the Cambridge Philosophical Society* **26**, 376–385 (1930).
8. E. Teller, *Rev. Mod. Phys.* **34**, 627 (1962).
9. W. Kohn, L. J. Sham, *Phys. Rev.* **140**, A1133 (1965).
10. W. Kohn, *Rev. Mod. Phys.* **71**, 1253 (1999).
11. J. A. Pople, *Rev. Mod. Phys.* **71**, 1267 (1999).
12. R. O. Jones, *Rev. Mod. Phys.* **87**, 897 (2015).
13. K. Lejaeghere, *et al., Science* **351**, aad3000 (2016).
14. M. G. Medvedev, I. S. Bushmarinov, J. Sun, J. P. Perdew, K. A. Lyssenko, *Science* **355**, 49 (2017).
15. L. Fiedler, K. Shah, M. Bussmann, A. Cangi, *Phys. Rev. Materials* **6**, 040301 (2022).
16. R. Pederson, B. Kalita, K. Burke, *Nature Reviews Physics* **4**, 357 (2022).
17. J. C. Snyder, M. Rupp, K. Hansen, K.-R. Müller, K. Burke, *Phys. Rev. Lett.* **108**, 253002 (2012).
18. J. Kirkpatrick, *et al., Science* **374**, 1385 (2021).
19. V. L. Lignères, E. A. Carter, *An Introduction to Orbital-Free Density Functional Theory* (Springer Netherlands, Dordrecht, 2005), pp. 137–148.
20. L. Hung, E. A. Carter, *Chemical Physics Letters* **475**, 163 (2009).

21. Y. A. Wang, E. A. Carter, *Orbital-Free Kinetic-Energy Density Functional Theory* (Springer Netherlands, Dordrecht, 2002), pp. 117–184.

22. W. Yang, *Phys. Rev. Lett.* **66**, 1438 (1991).

23. S. Goedecker, L. Colombo, *Phys. Rev. Lett.* **73**, 122 (1994).

24. D. R. Bowler, T. Miyazaki, *Reports on Progress in Physics* **75**, 036503 (2012).

25. F. Brockherde, L. Vogt, L. Li, M. E. Tuckerman, K. Burke, K.-R. Müller, *Nature Communications* **8** (2017).

26. M. Tsubaki, T. Mizoguchi, *Phys. Rev. Lett.* **125**, 206401 (2020).

27. A. Chandrasekaran, *et al., npj Computational Materials* **5**, 22 (2019).

28. J. A. Ellis, *et al., Physical Review B* **104**, 035120 (2021).

29. W. Kohn, *Phys. Rev. Lett.* **76**, 3168 (1996).

30. A. Cangi, *et al., Software publication "MALA", DOI: 10.5281/zenodo.5557254* (2021).

31. A. P. Thompson, *et al., Comp. Phys. Comm.* **271**, 108171 (2022).

32. A. P. Thompson, L. P. Swiler, C. R. Trott, S. M. Foiles, G. J. Tucker, *J. Comput. Phys.* **285**, 316 (2015).

33. M. A. Wood, M. A. Cusentino, B. D. Wirth, A. P. Thompson, *Phys. Rev. B* **99**, 184305 (2019).

34. A. Nakata, *et al.*, *J. Chem. Phys.* **152**, 164112 (2020).

35. A. Stukowski, *Modelling and Simulation in Materials Science and Engineering* **18**, 015012 (2009).

36. D. M. Wilkins, *et al., Proc Natl Acad Sci USA* **116**, 3401 (2019).

37. G. Kresse, J. Hafner, *Physical Review B* **47**, 558 (1993).

38. P. Giannozzi, *et al.*, *Journal of Physics: Condensed Matter* **21**, 395502 (2009).

39. *Materials and methods are available as supplementary materials at the Science website*

40. M. S. Daw, M. I. Baskes, *Physical Review B* **29**, 6443 (1984).

41. A. Agrawal, R. Mishra, L. Ward, K. M. Flores, W. Windl, *Modelling and Simulation in Materials Science and Engineering* **21**, 085001 (2013).

42. M. P. Allen, D. J. Tildesley, *Computer simulation of liquids*, (Oxford University Press, 2017).

43. L. Fiedler, A. Cangi, *Data set 'LDOS/SNAP data for MALA: Beryllium at 298K', DOI: 10.14278/rodare.1834* (2022).

44. L. Fiedler, *et al., Data set 'Scripts and Models for "Predicting the Electronic Structure of Matter on Ultra-Large Scales"', DOI: 10.14278/rodare.1851* (2022).

45. N. D. Mermin, *Physical Review* **137**, A1441 (1965).

46. M. Born, R. Oppenheimer, *Annales de Physique* **389**, 457 (1927).

47. M. Toda, R. Kubo, R. Kubo, N. Saitō, N. Hashitsume, *Statistical Physics: Equilibrium statistical mechanics*, Solid-State Sciences Series (Springer-Verlag, 1983).

48. D. M. Ceperley, B. J. Alder, *Physical Review Letters* **45**, 566 (1980).

49. J. P. Perdew, W. Yue, *Physical Review B* **33**, 8800 (1986).

50. J. P. Perdew, Y. Wang, *Physical Review B* **45**, 13244 (1992).

51. J. P. Perdew, K. Burke, M. Ernzerhof, *Phys. Rev. Lett.* **77**, 3865 (1996).

52. J. Sun, A. Ruzsinszky, J. P. Perdew, *Physical Review Letters* **115**, 036402 (2015).

53. V. V. Karasiev, D. Chakraborty, O. A. Shukruto, S. B. Trickey, *Physical Review B* **88**, 161108 (2013).

54. V. V. Karasiev, T. Sjostrom, J. Dufty, S. B. Trickey, *Phys. Rev. Lett.* **112**, 076403 (2014).

55. S. Groth, *et al., Phys. Rev. Lett.* **119**, 135001 (2017).

56. E. W. Brown, J. L. DuBois, M. Holzmann, D. M. Ceperley, *Phys. Rev. B* **88**, 081102 (2013).

57. R. Iftimie, P. Minary, M. E. Tuckerman, *Proceedings of the National Academy of Sciences* **102**, 6654 (2005).

58. M. A. Wood, A. P. Thompson, *J. Chem. Phys* **148**, 241721 (2018).

59. M. A. Cusentino, M. A. Wood, A. P. Thompson, *J. Phys. Chem. A* **124**, 5456 (2020).

60. L. Fiedler, *et al., arXiv preprint arXiv:2202.09186* (2022).

61. K. Hornik, *Neural Networks* **4**, 251 (1991).

62. M. Minsky, S. A. Papert, *Perceptrons. An Introduction to Computational Geometry* (The MIT Press, Cambridge, Mass, 2017), expanded, subsequent edition edn.

63. F. Rosenblatt, The Perceptron: A Perceiving and Recognizing Automaton (Project PARA)., *Tech. Rep. 85-460-1*, Cornell Aeronautical Laboratory (1957).

64. D. E. Rumelhart, G. E. Hinton, R. J. Williams, *Nature* **323**, 533 (1986).

65. I. Goodfellow, Y. Bengio, A. Courville, *Deep Learning* (MIT Press, 2016).

66. A. Paszke, *et al., Advances in Neural Information Processing Systems* (Curran Associates, Inc., 2019), vol. 32.

67. P. Giannozzi, O. Baseggio, P. Bonfà, D. Brunato, R. Car, I. Carnimeo, C. Cavazzoni, S. de Gironcoli, P. Delugas, F. Ferrari Ruffino, *et al.*, *Journal of Chemical Physics* **152**, 154105 (2020).

68. P. Giannozzi, O. Andreussi, T. Brumme, O. Bunau, M. B. Nardelli, M. Calandra, R. Car, C. Cavazzoni, D. Ceresoli, M. Cococcioni, N. Colonna, I. Carnimeo, A. D. Corso, S. de Gironcoli, P. Delugas, R. A. DiStasio, A. Ferretti, A. Floris, G. Fratesi, G. Fugallo, R. Gebauer, U. Gerstmann, F. Giustino, T. Gorni, J. Jia, M. Kawamura, H.-Y. Ko, A. Kokalj, E. Küçükbenli, M. Lazzeri, M. Marsili, N. Marzari, F. Mauri, N. L. Nguyen, H.-V. Nguyen, A. O. de-la Roza, L. Paulatto, S. Poncé, D. Rocca, R. Sabatini, B. Santra, M. Schlipf, A. P. Seitsonen, A. Smogunov, I. Timrov, T. Thonhauser, P. Umari, N. Vast, X. Wu, S. Baroni, *Journal of Physics: Condensed Matter* **29**, 465901 (2017).

69. W. Smith, *CCP5 Information Quarterly for Computer Simulation of Condensed Phases* **30**, 35 (1989).

70. G. Kresse, J. Furthmüller, *Physical Review B* **54**, 11169 (1996).

71. G. Kresse, J. Furthmüller, *Computational Materials Science* **6**, 15 (1996).

72. A. Dal Corso, *Computational Materials Science* **95**, 337 (2014).

73. P. E. Blöchl, *Phys. Rev. B* **50**, 17953 (1994).

74. G. Kresse, D. Joubert, *Phys. Rev. B* **59**, 1758 (1999).

75. L. Fiedler, *et al., arXiv preprint arXiv:2206.03754* (2022).

76. H. J. Monkhorst, J. D. Pack, *Phys. Rev. B* **13**, 5188 (1976).