8-21-2020

Machine Learning Discovery of Computational Model Efficacy Boundaries

Michael S. Murillo
Michigan State University

Mathieu Marciante
CEA Dam

Liam G. Stanton
San Jose State University, liam.stanton@sjsu.edu

Recommended Citation

Michael S. Murillo, Mathieu Marciante, and Liam G. Stanton. "Machine Learning Discovery of Computational Model Efficacy Boundaries" *Physical Review Letters* (2020). https://doi.org/10.1103/PhysRevLett.125.085503

Machine Learning Discovery of Computational Model Efficacy Boundaries

Michael S. Murillo,[1,*] Mathieu Marciante,[2,†] and Liam G. Stanton[3,‡]
[1]Department of Computational Mathematics, Science and Engineering, Michigan State University, East Lansing, Michigan 48824, USA
[2]CEA-DAM, DIF F-91297 Arpajon, France
[3]Department of Mathematics and Statistics, San José State University, San José, California 95192, USA

(Received 28 December 2019; accepted 30 July 2020; published 20 August 2020)

Computational models are formulated in hierarchies of variable fidelity, often with no quantitative rule for defining the fidelity boundaries. We have constructed a dataset from a wide range of atomistic computational models to reveal the accuracy boundary between higher-fidelity models and a simple, lower-fidelity model. The symbolic decision boundary is discovered by optimizing a support vector machine on the data through iterative feature engineering. This data-driven approach reveals two important results: (i) a symbolic rule emerges that is independent of the algorithm, and (ii) the symbolic rule provides a deeper understanding of the fidelity boundary. Specifically, our dataset is composed of radial distribution functions from seven high-fidelity methods that cover wide ranges in the features (element, density, and temperature); high-fidelity results are compared with a simple pair-potential model to discover the nonlinear combination of the features, and the machine learning approach directly reveals the central role of atomic physics in determining accuracy.

DOI: 10.1103/PhysRevLett.125.085503

Computational models of physical systems vary markedly in accuracy and attainable scales. The costs associated with high-fidelity (HF) models drive the need for accurate surrogate models as well as methods that combine fidelities [1–3]. Unfortunately, there are no simple rules that determine the "fidelity boundary" among all available models. Here, we construct a symbolic machine-learning framework with the goal of discovering the fidelity boundary between HF and low-fidelity (LF) computational models. For our purposes, we employ HF models that resolve atomic scales and include electronic-structure methods that generate on-the-fly potentials. Such HF models incur costs associated with shorter timescales and length scales, reduced statistical convergence, and fewer cases, among other difficulties. Choosing the optimal fidelity level allows these costs to be minimized; in some cases, the accessible physics phenomena can be qualitatively different when using a LF model. For example, the number of particles used in HF models [4,5] is typically many orders of magnitude lower than that of LF models [6,7], and compromises can often be made [8] to access important heterogeneous, nonequilibrium mesoscale phenomena.

Machine learning (ML) offers a set of tools that potentially provide novel approaches to solving such problems. Increasingly, ML is being used to tackle a wide range of problems in physics, including predicting disruptions in burning plasmas [9], modeling ionization energies [10], accelerating molecular dynamics (MD) [11], enhancing many-body sampling techniques [12], coarse-graining molecular force fields [13], learning coherent structure from spatiotemporal data [14], and aiding inertial-confinement-fusion experimental design [15], among many others. Here, we propose to use ML not as a deployable algorithm that can be used to make predictions, but as a data-driven discovery framework that assigns accuracy scores to our hypotheses, allowing us to discover symbolic rules that are then independent of the specific ML algorithms employed.

To date, most computational physics communities do not generate and gather results with data science in mind. For this reason, we constructed a dataset from the extant literature, focusing on methods from the high energy-density community because of the range of features available, which are the element studied, the density, and the temperature; in thermodynamic equilibrium for a single species, these are the only three quantities needed. The most commonly reported quantity is the equilibrium ion-ion radial distribution function (RDF) $g(r)$; $g(r)$ values were digitized, and the height of the first peak was used as our metric for accuracy, as this is where the largest deviation between the RDFs of two models will typically occur. While other quantities could have been chosen, $g(r)$ plays a central role in determining most equilibrium quantities, and its peak position and height are well studied, with the height being the more sensitive of the two quantities [16] for most materials. (The complete dataset is available at GitHub [17].) One-hot encoding is used to map the ratio of the peak heights into binary form, with 0 for inaccurate and 1 for accurate, for an accuracy target,

which was taken to vary in the range 5%–15% in this work, unless otherwise specified; this process converts the physical data into a classification problem. RDFs were obtained from Kohn-Sham density functional theory molecular dynamics (KS-DFT-MD) [18–24], orbital-free density functional theory (DFT) [25–27], classical-map hypernetted chain [28,29], linear-response effective ions [30], quantum Langevin MD [31], dynamically screened ion-ion interactions [32], and quantum-statistical-potential MD [33]. An initial exploration of the data revealed several cases in which either no LF model would suffice (e.g., the presence of molecular states) or there was an obvious error (e.g., the RDF did not tend to unity), and these cases were removed to leave 34 RDFs in our dataset. Our final database reflected the diversity we desired to mitigate inaccuracies in the data and fidelity variations among the HF models.

Assessing fidelity requires a LF model, the simplest of which is the Yukawa model, which is defined in terms of a two-step process [8]. First, the physical domain of N nuclei is decomposed into N spheres, each with the ion-sphere radius $a = (3/4\pi n)^{1/3}$. An all-electron electronic structure calculation is then performed around each central nucleus, where, using a suitable definition, the electrons are decomposed into separate densities that are either strongly or weakly interacting with the nucleus. The strongly interacting electrons are assumed to be localized near the nucleus, and their impact is to convert the nuclear charge Ze to an ionic charge $\langle Z\rangle e$. Conversely, the weakly interacting electrons are treated in a long-wavelength linear response model to obtain the electronic screening cloud, with screening length λ, around the ionic core. This procedure yields the Yukawa ion-ion *pair* interaction energy between ions

$$u_Y(r) = \frac{\langle Z\rangle^2 e^2}{r}\exp\left(-r/\lambda\right), \tag{1}$$

which we take as our LF model. In this work, we employed the simplest choices for the Yukawa parameters, which are the Thomas-Fermi values of $\langle Z\rangle$ and λ [8]; our goal here is *not* to develop a new pair potential, but to examine how to establish a physical accuracy rule from data using the most widely used LF model. Yukawa RDFs were computed using standard pair-potential MD simulations.

Two examples from the dataset are shown in Fig. 1. Here, the HF methods KSMD [20] and QLMD [31] were each used for two densities and temperatures. Note that the hydrogen case is accurate for a very low temperature, but is at an elevated density. In contrast, at much higher temperatures, the Yukawa models fail to reproduce the iron results, with moderate improvement at 10 eV. (More examples are shown in the Supplemental Material [34].)

An alternative view of the dataset is visualized in Fig. 2. Points are labeled as either accurate (red), where the LF model agrees with the HF model (peak heights are within 5%), or inaccurate (blue), where the LF model does not agree with the HF model. The upper left panel indicates that our dataset has good coverage across temperature and density, and that, perhaps surprisingly, no accuracy trend is found in this plane. The next three panels reinforce this conclusion by revealing that there is no trend in accuracy versus temperature, density, or nuclear charge; therefore, it is not possible to know the accuracy of the LF (Yukawa) model based on any of these features alone.

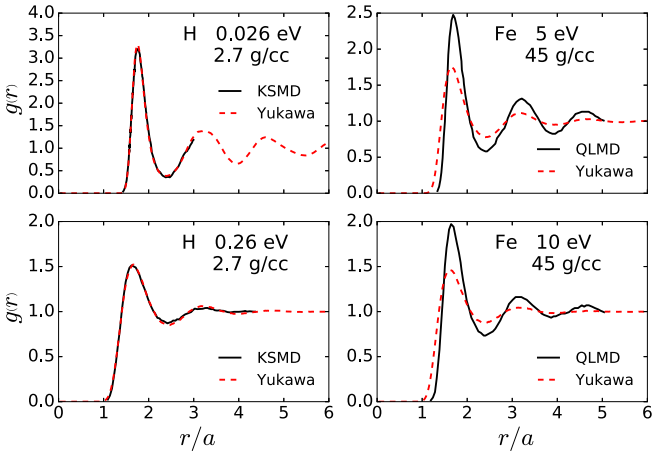

FIG. 1. Example RDFs from the dataset: Representative RDFs are shown for hydrogen [20] and iron [31] at various densities and temperatures. Two curves are shown in each panel, corresponding to the HF method (solid or black curve) and our base LF Yukawa model (dashed or red curve).

Any ML classifier employing linear separability (a vertical line for this 1D example) would fail; a better approach would be to seek probability distributions using logistic regression (LR); the LR predictions are shown as

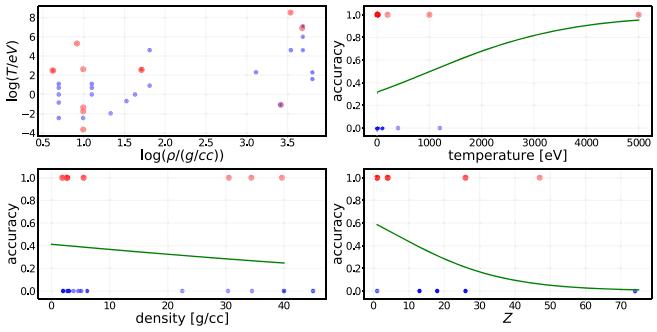

FIG. 2. Trends in the dataset: Data points in the $T-\rho$ plane are shown in the upper left plot, revealing good coverage within the dataset. Red (larger) points and blue (smaller) points are accurate and inaccurate, respectively, with accuracy defined here as agreement in peak height within 5%. In the next three panels, accuracy is plotted versus temperature, density, and nuclear charge, showing that no simple rule for assessing accuracy exists. The green curves show the results of a 1D (single-feature) logistic regression. Note that some of the points overlap, which is indicated through the intensity of the color.

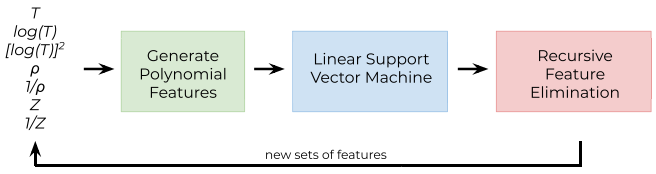

FIG. 3. Machine-learning workflow: Our symbolic machine-learning workflow is an iterative procedure that constructs the best features from physical features (possibly scaled), their inverses, and polynomial combinations. Recursive feature elimination is used to sort the quality of the features, which leads to a new set of features.

solid green lines in Fig. 2. Because of the dearth of data, these results are only notional, but they reveal the following rough trends. The LR curve obtained using only the temperature feature is moderately flat, and its trend is dominated by a single data point. The density feature yields a very flat probability distribution, indicating no predictive power. Finally, the nuclear-charge feature is also moderately flat, with a rough trend towards increased accuracy for lower-Z elements. (Alternate visualizations, and an application to transport [31], are given in the Supplemental Material [34].) We conclude that none of these three features alone can predict the fidelity boundary and that simple ML approaches are not particularly useful. Similar studies were carried out in two dimensions, using pairs of features, and in three dimensions, with similar results.

We developed a workflow to build new features in higher-dimensional spaces. Our ML workflow is shown in Fig. 3. The goal is to engineer features that yield human-interpretable accuracy boundaries. We employ a combination of feature engineering [35], feature selection [36,37], and a linear classifier (see below) to create a symbolic result [38]. To generate a physically meaningful symbolic representation of the decision boundary, we begin with the three basic features of temperature T, mass density ρ, and nuclear charge Z to form our basic feature set $\mathcal{F}_0 = \{T, \rho, Z\}$. As no additional physics information exists beyond $\mathcal{F}_0$, we engineer new features from $\mathcal{F}_0$. These new features are nonlinear combinations of those in $\mathcal{F}_0$, much like those generated in kernel methods. Note that we employ only the three most obvious and most basic features so as not to bias the method toward requiring specific domain knowledge of this example application.

Because our goal is a symbolic classifier, we do not employ nonlinear ML algorithms (e.g., kernel methods, neural networks) [39]. Rather, we employ a linear support vector machine (lSVM) to create a linear separability boundary in the high-dimensional space of our engineered features. The lSVM hyperparameter C was optimized. The coefficients are the weights of the nonlinear features that we use to assign importance to. The lSVM is used in a workflow that uses cross validation (CV) and recursive feature elimination (RFE). RFE ranks the importance of each feature, and CV informs us of the quality of the prediction. This scheme is an adaptation of the use of lSVM with RFE to down-select feature spaces as a preprocessing step for an expensive ML algorithm; here, by adding new nonlinear features, this scheme is essentially reversed to create additional features that have better performance. CV guards against overfitting by learning from various subsets of the data and predicting the remaining data, thereby quantifying generalizabilty as part of the workflow.

It is difficult to represent division in ML algorithms [40,41], so we augment $\mathcal{F}$ with inverses to extend our feature set to $\mathcal{F}_{\text{base}} = \{T, \rho, \rho^{-1}, Z, Z^{-1}\}$. Feature scaling was examined with no noticeable improvement except for the replacement $T \to \log(T)$, yielding $\mathcal{F} = \{\log(T), \rho, \rho^{-1}, Z, Z^{-1}\}$. Because the logarithm of T^{-1} is trivially $-\log(T)$, we did not include T^{-1} in the feature set; thus, the three physical dimensions inherent in $\mathcal{F}_0$ are transformed into a 5D feature space. Next, we construct all second-order polynomials from this feature set to project into a much higher-dimensional feature space containing all bilinear combinations of the features and squares of the basic features; for example, for the simplest case of $\mathcal{F}_0$ we obtain $\mathcal{F}_{\text{poly}} = \{1, T, \rho, Z, T^2, T\rho, TZ, \rho^2, \rho Z, Z^2\}$; importantly, note that constants are included. Polynomial terms constructed from the feature vector $\mathcal{F}$ can be itemized according to importance through RFE, which yields the symbolic result we seek.

In practice, an iterative approach was used to find the best combination of the basic features by updating the feature vector based on the current best features: $\mathcal{F}^n \to \mathcal{F}^{n+1}$. For example, RFE revealed that the square of $\log(T)$ was a strong feature, and thus, the feature space $\mathcal{F}$ was updated to include this feature. This iterative procedure, which we call "recursive feature updating" (RFU), allows for higher-order powers to appear, retains the best features, and forces new feature rankings. Eventually, products such as $\log(T)/Z$ were identified as strong features, and RFU led to the inequality

$$\xi = \log^2(\mathrm{T/eV})\,\frac{(\rho + 10)/(\mathrm{g/cm^3})}{\mathrm{Z}} > 2.0, \qquad (2)$$

which gave $> 90\%$ accuracy on our dataset. The ratio of peak heights is shown versus (2) in Fig. 4, which reveals that there is a clear boundary that separates inaccurate predictions for small values of ξ and accurate predictions for larger values of ξ.

The decision boundary implied by ξ in temperature-density space is shown in Fig. 5. In contrast to other metrics, such as the Coulomb coupling or degeneracy boundaries [42] that imply that very high temperatures are required at high density, the temperature at which a LF model is appropriate occurs at *lower* with densities. This

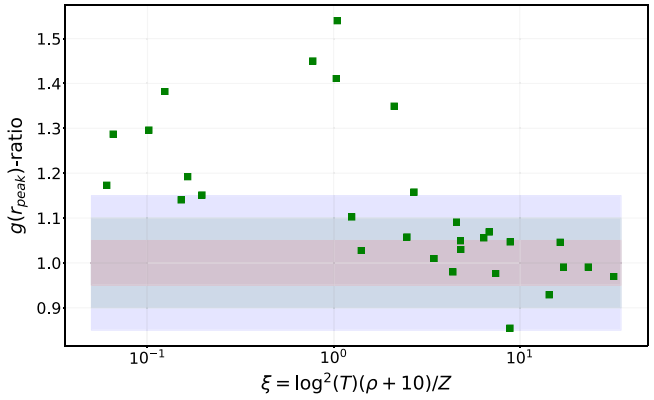

FIG. 4. Machine-learning boundary: The ratio of $g(r)$ peak heights (HF divided by LF) are shown versus the discovered parameter ξ in Eq. (2). The colored bands indicate accuracy ranges of 5%, 10%, and 15%. The inequality for ξ in Eq. (2) arises from drawing a vertical line near the erroneous points on the left.

result can be understood in the context of modern computational methods in which MD simulations of simple properties like $g(r)$ are now ubiquitous: the use of MD "solves" the ionic strongly coupling "problem," which no longer adds to our uncertainty. Similarly, the use of Thomas-Fermi inputs, which are widely available, solves the high-density problem, because the Thomas-Fermi model becomes more accurate at higher density. Our RFU ML approach has naturally found these trends from the data.

While the RFU-based ML approach described above yields a symbolic separation boundary that can be applied independently of the lSVM used to find it, we sought further insight into the physics. The result (2) shows that simpler computational methods can be used when the temperature is high *and* the density is high *and* the nuclear charge is low. This particular combination of features is precisely what controls the mean ionization state (MIS) [43] of the material.

To examine this potential finding, we again form a single feature ζ and plot accuracy versus ζ in Fig. 6, which should be compared with Fig. 4. From this figure, we find an accuracy boundary of

$$\zeta = \frac{\langle Z \rangle}{Z} > 0.35. \tag{3}$$

Note that we use the fairly conservative definition of accuracy of 10% agreement for the first peak height; moreover, this result is conservative because some of the fluctuations in Fig. 6 may be due to imperfect (e.g., finite-size errors) data in the database. Taken together, the two rules (2) and (3) lead to the conclusion that neither temperature nor density alone, nor a combination of the two, leads to an accuracy boundary for the Yukawa model, but rather atomic physics: the rule states that if the material is more than half ionized, a much faster computational model can be used. This result illustrates how the ML found a physical feature that might have been used in the original set of features, thereby empowering the ML with physics guidance based on expert knowledge; here, we made no attempt to bias the learning other than through the three most basic features.

In summary, we have examined a framework in which accuracy scores from ML can be used with feature engineering and extraction to identify a symbolic boundary using easily accessible ML libraries. To illustrate this approach, we constructed a dataset consisting of RDFs obtained using a wide variety of HF computational methods and compared them with predictions from a LF model. Simple analyses, such as LR, showed that the basic physical features $\{Z, \rho, T\}$ are not predictive as unary features or in pairs. More powerful ML approaches,

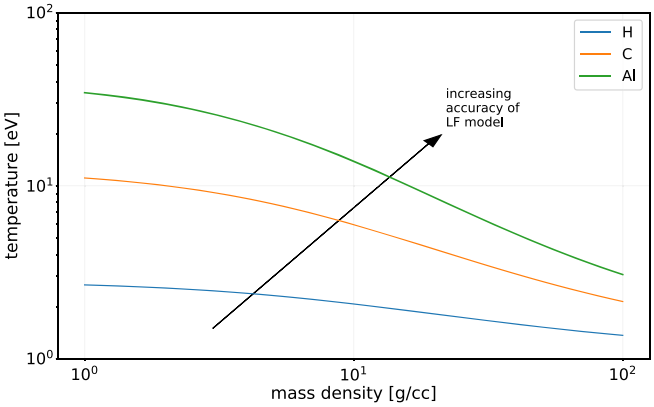

FIG. 5. Boundary in $T - \rho$ space: The decision boundary is shown for three elements, hydrogen, carbon and aluminum, in the temperature-density plane. LF models are expected to be accurate above the line. These curves capture the obvious trends that LF models are applicable for higher densities (Thomas-Fermi limit), lower nuclear charges, and higher temperatures.

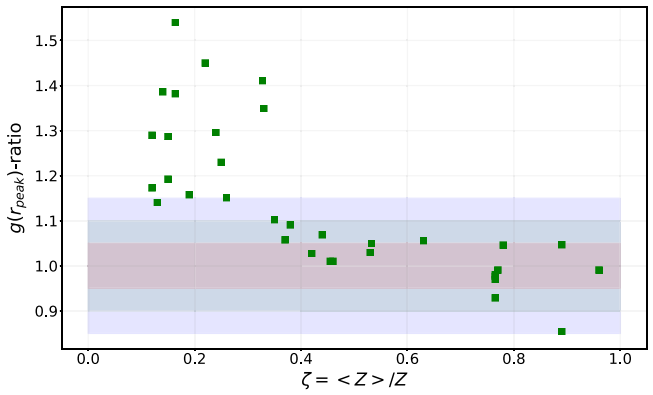

FIG. 6. Mean ionization state boundary: The ratio of $g(r)$ peak heights (HF divided by LF) are shown versus the discovered parameter (3). The colored bands indicate accuracy ranges of 5%, 10%, and 15%.

however, achieved a moderate accuracy in two dimensions (considering pairs of features). In three dimensions, high accuracy can be achieved with nonlinear ML algorithms, although these algorithms do not reveal the decision boundary in an interpretable way.

By considering various polynomial combinations of features, including division, and excising weak features, we find that the decision boundary is given symbolically as $\log^2(T)(\rho + 10)/Z$. We find that this decision boundary is closely connected to the MIS and propose a related criterion $\zeta = \langle Z \rangle / Z$ that is based on atomic physics. The reason that atomic physics (and ionization in particular) is the key physics involved here is that all modern methods naturally capture ionic strong coupling and, at high enough temperature and/or density, the free electrons are captured well in a Thomas-Fermi approximation. This finding suggests that pair potentials that treat the bound electrons with much higher fidelity [28] would potentially greatly expand the Yukawa accuracy regime shown in Fig. 5, allowing for significantly larger simulations with little cost to accuracy; from an uncertainty quantification perspective [44–46], highly converged pair-potential MD could compete with HF methods in some cases. In particular, based on the insensitivity of disparate models to the MIS [43] and to gradient corrections in the screening [47], sensitivity to atomic physics suggests that the most important improvement to Yukawa would be a more refined pseudopotential. For example, our original database was larger than we present here, but many of the HF results were not properly converged (e.g., too noisy to establish a peak height), and we were unable to use such results. Through such improved potentials with orders of magnitude more particles and timesteps, qualitatively different *heterogeneous*, nonequilibrium studies [8] can be performed at the mesoscale.

The results here suggest that a more concerted effort should be made in the computational communities to produce high-quality data. In particular, we found that the density ρ was a generally weak feature, although it appears linearly in our decision boundary. Unfortunately, most results in the literature do not systematically explore wide density variations and report RDFs across those variations. For example, the MIS is not monotonic in ρ [43], although the dataset we employed suggests that it is; the low-density portion of Fig. 5 is likely the most uncertain for these reasons. Ideally, more studies that vary all features in $\mathcal{F}_0$, such as a $\{T, \rho, Z\}$ grid of highly converged HF RDFs and velocity autocorrelation functions motivated by Fig. 5, would improve our ability to allow ML techniques to improve our understanding of computational techniques and the physics they address. Based on the results of this work, we propose a dataset minimally of the form $T = \{1, 5, 10, 20, 50\}$ eV, $Z = \{1, 4, 6, 13, 26\}$, $\rho/\rho_0 = \{0.1, 0.5, 1, 2, 10\}$, where ρ_0 is the standard density of the material. Most important are density variations, which are less commonly explored in the current literature; moreover, building databases with more challenging quantities, such as the velocity autocorrelation function, would further strengthen the quality of future ML studies. With a concerted effort, using a wide range of interactions beyond Yukawa to produce high-quality data, the workflow in Fig. 3 can be adapted to a wider range of problems [48].

M. S. Murillo acknowledges support from the Air Force Office of Scientific Research through Grant No. FA9550-17-1-0394.

*Corresponding author.
murillom@msu.edu
†mathieu.marciante@cea.fr
‡liam.stanton@sjsu.edu

[1] M. Razi, A. Narayan, R. M. Kirby, and D. Bedrov, Fast predictive models based on multi-fidelity sampling of properties in molecular dynamics simulations, Comput. Mater. Sci. **152**, 125 (2018).
[2] G. Pilania, J. E. Gubernatis, and T. Lookman, Multi-fidelity machine learning models for accurate bandgap predictions of solids, Comput. Mater. Sci. **129**, 156 (2017).
[3] M. Fernández-Godino, C. Park, N.-H. Kim, and R. T. Haftka, Review of multi-fidelity models, arXiv:1609.07196.
[4] L. K. Wagner and D. M. Ceperley, Discovering correlated fermions using quantum Monte Carlo, Rep. Prog. Phys. **79**, 094501 (2016).
[5] K. P. Driver, F. Soubiran, and B. Militzer, Path integral Monte Carlo simulations of warm dense aluminum, Phys. Rev. E **97**, 063207 (2018).
[6] J. R. Perilla, B. C. Goh, C. K. Cassidy, B. Liu, R. C. Bernardi, T. Rudack, H. Yu, Z. Wu, and K. Schulten, Molecular dynamics simulations of large macromolecular complexes, Curr. Opin. Struct. Biol. **31**, 64 (2015).
[7] T. C. Germann and K. Kadau, Trillion-atom molecular dynamics becomes a reality, Int. J. Mod. Phys. C **19**, 1315 (2008).
[8] L. G. Stanton, J. N. Glosli, and M. S. Murillo, Multiscale Molecular Dynamics Model for Heterogeneous Charged Systems, Phys. Rev. X **8**, 021044 (2018).
[9] J. Kates-Harbeck, A. Svyatkovskiy, and W. Tang, Predicting disruptive instabilities in controlled fusion plasmas through deep learning, Nature (London) **568**, 526 (2019).
[10] M. Rupp, A. Tkatchenko, K.-R. Müller, and O. A. Von Lilienfeld, Fast and Accurate Modeling of Molecular Atomization Energies with Machine Learning, Phys. Rev. Lett. **108**, 058301 (2012).
[11] V. Botu and R. Ramprasad, Adaptive machine learning framework to accelerate *ab initio* molecular dynamics, Int. J. Quantum Chem. **115**, 1074 (2015).
[12] F. Noé, S. Olsson, J. Köhler, and H. Wu, Boltzmann generators: Sampling equilibrium states of many-body systems with deep learning, Science **365**, eaaw1147 (2019).
[13] J. Wang, S. Olsson, C. Wehmeyer, A. Perez, N. E. Charron, G. De Fabritiis, F. Noe, and C. Clementi, Machine learning of coarse-grained molecular dynamics force fields, ACS Central Sci. **5**, 755 (2019).

[14] A. Rupe, N. Kumar, V. Epifanov, K. Kashinath, O. Pavlyk, F. Schlimbach, M. Patwary, S. Maidanov, V. Lee, J. P. Crutchfield*et al.*, Disco: Physics-based unsupervised discovery of coherent structures in spatiotemporal systems, arXiv:1909.11822.
[15] J. L. Peterson, K. D. Humbird, J. E. Field, S. T. Brandon, S. H. Langer, R. C. Nora, B. K. Spears, and P. T. Springer, Zonal flow generation in inertial confinement fusion implosions, Phys. Plasmas **24**, 032702 (2017).
[16] T. Ott and M. Bonitz, First-principle results for the radial pair distribution function in strongly coupled one-component plasmas Contrib. Plasma Phys. **55**, 243 (2015).
[17] https://github.com/MurilloGroupMSU.
[18] D. Hohl, V. Natoli, D. M. Ceperley, and R. M. Martin, Molecular Dynamics in Dense Hydrogen, Phys. Rev. Lett. **71**, 541 (1993).
[19] S. M. Younger, Many-atom screening effects on diffusion in dense helium, Phys. Rev. A **45**, 8657 (1992).
[20] J. Kohanoff and J.-P. Hansen, Statistical properties of the dense hydrogen plasma: An ab initio molecular dynamics investigation, Phys. Rev. E **54**, 768 (1996).
[21] P. L. Silvestrelli, No evidence of a metal-insulator transition in dense hot aluminum: A first-principles study, Phys. Rev. B **60**, 16382 (1999).
[22] W. Lorenzen, B. Holst, and R. Redmer, First-order liquid-liquid phase transition in dense hydrogen, Phys. Rev. B **82**, 195107 (2010).
[23] K. U. Plagemann, P. Sperling, R. Thiele, M. P. Desjarlais, C. Fortmann, T. Döppner, H. J. Lee, S. H. Glenzer, and R. Redmer, Dynamic structure factor in warm dense beryllium, New J. Phys. **14**, 055020 (2012).
[24] H. Sun, D. Kang, J. Dai, W. Ma, L. Zhou, and J. Zeng, First-principles study on equation of states and electronic structures of shock compressed ar up to warm dense regime. J. Chem. Phys. **144**, 124503 (2016).
[25] F. Lambert, J. Clérouin, and G. Zérah, Very-high-temperature molecular dynamics, Phys. Rev. E **73**, 016403 (2006).
[26] J. Clérouin, Cooking strongly coupled plasmas, Mol. Phys. **113**, 2403 (2015).
[27] J. Clérouin, P. Arnault, C. Ticknor, J. D. Kress, and L. A. Collins, Unified Concept of Effective One Component Plasma for Hot Dense Plasmas, Phys. Rev. Lett. **116**, 115003 2016.
[28] M. W. C. Dharma-Wardana, Electron-ion and ion-ion potentials for modeling warm dense matter: Applications to laser-heated or shock-compressed Al and Si, Phys. Rev. E **86**, 036407 (2012).
[29] R. Bredow, T. h. Bornath, W.-D. Kraeft, M. W. C. Dharma-wardana, and R. Redmer, Classical-map hypernetted chain calculations for dense plasmas, Contrib. Plasma Phys. **55**, 222 (2015).
[30] E. Liberatore, C. Pierleoni, and D. M. Ceperley, Liquid-solid transition in fully ionized hydrogen at ultra-high pressures, J. Chem. Phys. **134**, 184505 (2011).
[31] J. Dai, Y. Hou, D. Kang, H. Sun, J. Wu, and J. Yuan, Structure, equation of state, diffusion and viscosity of warm dense Fe under the conditions of a giant planet core, New J. Phys. **15**, 045003 (2013).
[32] K. K. Mon, N. W. Ashcroft, and G. V. Chester, Core polarization and the structure of simple metals, Phys. Rev. B **19**, 5103 (1979).
[33] J. P. Hansen and I. R. McDonald, Microscopic Simulation of a Hydrogen Plasma, Phys. Rev. Lett. **41**, 1379 (1978).
[34] See Supplemental Material at http://link.aps.org/supplemental/10.1103/PhysRevLett.125.085503 for additional explorations of the data set and for an application to transport.
[35] A. Zheng and A. Casari, *Feature Engineering for Machine Learning: Principles and Techniques for Data Scientists* (O'Reilly Media, Inc., California, 2018).
[36] I. Guyon and A. Elisseeff, An introduction to variable and feature selection, J. Mach. Learn. Res. **3**, 1157 (2003).
[37] J. Hua, Z. Xiong, J. Lowey, E. Suh, and E. R. Dougherty, Optimal number of features as a function of sample size for various classification rules, Bioinformatics **21**, 1509 (2004).
[38] D. A. Augusto and H. J. C. Barbosa, Symbolic regression via genetic programming, in *Proceedings of the Sixth Brazilian Symposium on Neural Networks, Rio de Janeiro, RJ, Brazil* (IEEE, 2000), Vol. 1, pp. 173–178.
[39] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, Scikit-learn: Machine learning in PYTHON, J. Mach. Learn. Res. **12**, 2825 (2011).
[40] K.-Y. Siu, J. Bruck, T. Kailath, and T. Hofmeister, Depth efficient neural networks for division and related problems, IEEE Trans. Inf. Theory **39**, 946 (1993).
[41] S. S. Sahoo, C. H. Lampert, and G. Martius, Learning equations for extrapolation and control, arXiv:1806.07259.
[42] M. S. Murillo, Strongly coupled plasma physics and high energy-density matter, Phys. Plasmas **11**, 2964 (2004).
[43] M. S. Murillo, J. Weisheit, S. B. Hansen, and M. W. C. Dharma-Wardana, Partial ionization in dense plasmas: Comparisons among average-atom density functional models, Phys. Rev. E **87**, 063113 (2013).
[44] P. Angelikopoulos, C. Papadimitriou, and P. Koumoutsakos, Bayesian uncertainty quantification and propagation in molecular dynamics simulations: A high performance computing framework, J. Chem. Phys. **137**, 144103 (2012).
[45] P. N. Patrone, A. Dienstfrey, A. R. Browning, S. Tucker, and S. Christensen, Uncertainty quantification in molecular dynamics studies of the glass transition temperature, Polymer **87**, 246 (2016).
[46] P. Angelikopoulos, C. Papadimitriou, and P. Koumoutsakos, Data driven, predictive molecular dynamics for nanoscale flow simulations under uncertainty, J. Phys. Chem. B **117**, 14808 (2013).
[47] L. G. Stanton and M. S. Murillo, Unified description of linear screening in dense plasmas, Phys. Rev. E **91**, 033104 (2015).
[48] W. Guodong, S. Lanxiang, W. Wei, C. Tong, G. Meiting, and Z. Peng, A feature selection method combined with ridge regression and recursive feature elimination in quantitative analysis of laser induced breakdown spectroscopy, Plasma Sci. Technol **22**, 074002 (2020).