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*

Follow this and additional works at: [https://scholarworks.sjsu.edu/faculty_rsc](https://scholarworks.sjsu.edu/faculty_rsc)

**Recommended Citation**

This Article is brought to you for free and open access by SJSU ScholarWorks. It has been accepted for inclusion in Faculty Research, Scholarly, and Creative Activity by an authorized administrator of SJSU ScholarWorks. For more information, please contact scholarworks@sjsu.edu.
Machine Learning Discovery of Computational Model Efficacy Boundaries

Michael S. Murillo, Mathieu Marcianiec, and Liam G. Stanton

Department of Computational Mathematics, Science and Engineering, Michigan State University, East Lansing, Michigan 48824, USA
CEA-DAM, DIF F-91297 Arpajon, France
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 hyper-netted 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 $a = (3/4\pi n) \frac{1}{3}$. 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 $(Z/e)$. Conversely, the weakly interacting electrons are treated in a long-wavelength linear response model to obtain the electronic screening cloud, with screening length $\lambda$, around the ionic core. This procedure yields the Yukawa ion-ion pair interaction energy between ions

$$u_Y(r) = \frac{(Z)^2 e^2}{r} \exp(-r/\lambda),$$

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 $(Z)$ and $\lambda$ [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 not 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.

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

![Figure 1](image1.png)

**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).

![Figure 2](image2.png)

**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.
These new features are nonlinear combinations of those in \( F \) that exist beyond the current best combination of the basic features by updating the feature set; thus, the three physical dimensions inherent in \( F_0 \) 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 \( F_0 \) we obtain \( F_{\text{poly}} = \{ T, \rho, Z^2, T^2, \rho Z, T Z, \rho^2, Z^2 \} \). Importantly, note that constants are included. Polynomial terms constructed from the feature vector \( F \) 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: \( F^n \to F^{n+1} \). For example, RFE revealed that the square of \( \log(T) \) was a strong feature, and thus, the feature space \( F \) 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 \( \log(T)/Z = \log(T/Z) \) were identified as strong features, and RFU led to the inequality

\[
\xi = \log^2(T/eV) \left( \frac{\rho + 10}{(g/cm^3)} \right) \frac{1}{Z} > 2.0,
\]

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 \( \xi \) and accurate predictions for larger values of \( \xi \).

The decision boundary implied by \( \xi \) 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.
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 ISVM 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 $\zeta$ and plot accuracy versus $\zeta$ in Fig. 6, which should be compared with Fig. 4. From this figure, we find an accuracy boundary of

$$\zeta = \frac{\langle Z \rangle}{Z} > 0.35.$$  (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, ρ, T} are not predictive as unary features or in pairs. More powerful ML approaches,
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 \( \zeta = (Z)/Z \) 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 \( \rho \) 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 \( \rho \) [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 \( F_\alpha \), 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 \( \rho_0 \) 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.**

murrilom@msu.edu

mathieu.marcianete@cea.fr

liam.stanton@sjtu.edu


