Orbital-free bond breaking via machine learning

John C. Snyder, Matthias Rupp, Katja Hansen, Leo Blooston, Klaus-Robert Müller, and Kieron Burke

Citation: The Journal of Chemical Physics **139**, 224104 (2013); doi: 10.1063/1.4834075
View online: http://dx.doi.org/10.1063/1.4834075
View Table of Contents: http://scitation.aip.org/content/aip/journal/jcp/139/22?ver=pdfcov
Published by the AIP Publishing

THE JOURNAL OF CHEMICAL PHYSICS **139**, 224104 (2013)

Orbital-free bond breaking via machine learning

John C. Snyder,[1] Matthias Rupp,[2,a)] Katja Hansen,[3] Leo Blooston,[4]
Klaus-Robert Müller,[5,6] and Kieron Burke[1]
[1]*Departments of Chemistry and of Physics, University of California, Irvine, California 92697, USA*
[2]*Institute of Pharmaceutical Sciences, ETH Zurich, 8093 Zürich, Switzerland*
[3]*Fritz-Haber-Institut der Max-Planck-Gesellschaft, 14195 Berlin, Germany*
[4]*Department of Chemistry, University of California, Irvine, California 92697, USA*
[5]*Machine Learning Group, Technical University of Berlin, 10587 Berlin, Germany*
[6]*Department of Brain and Cognitive Engineering, Korea University, Anam-dong, Seongbuk-gu, Seoul 136-713, South Korea*

(Received 7 June 2013; accepted 13 November 2013; published online 10 December 2013)

Using a one-dimensional model, we explore the ability of machine learning to approximate the non-interacting kinetic energy density functional of diatomics. This nonlinear interpolation between Kohn-Sham reference calculations can (i) accurately dissociate a diatomic, (ii) be systematically improved with increased reference data and (iii) generate accurate self-consistent densities via a projection method that avoids directions with no data. With relatively few densities, the error due to the interpolation is smaller than typical errors in standard exchange-correlation functionals.
 [http://dx.doi.org/10.1063/1.4834075]

I. INTRODUCTION

Kohn-Sham density functional theory (KS DFT)[1,2] is a widely used electronic structure method, striking a balance between accuracy and computational efficiency.[3] KS DFT requires solving a self-consistent set of orbital equations, yielding the exact non-interacting kinetic energy.[4] In return, only a small fraction of the total energy, the exchange-correlation (XC) energy, needs to be approximated as a functional of the electronic spin densities. This produces far greater accuracy than obtained with the original, Thomas-Fermi-based, orbital-free DFT,[5] thus motivating the search for better kinetic energy approximations.[6,7]

The computational bottleneck in KS DFT calculations is the need to solve the KS equations, which formally scale as N^3, where N is the number of electrons. Thus there is strong interest in constructing an orbital-free DFT which would avoid this step.[8] A sufficiently accurate approximation to $T_s[n]$, the kinetic energy (KE) of KS electrons, would produce an orbital-free scheme, greatly reducing the computational cost of DFT without sacrificing accuracy. Sufficient accuracy would be that the error made in approximating T_s be less than that made by current approximations for XC; accuracy beyond this is practically irrelevant.

Several research efforts have recently focused in this direction.[6,7] Unfortunately, the relative accuracy requirements of a KE functional are much stricter than those of an XC functional because T_s is typically comparable to the total energy of the system.[5,9,10] Worse, one also needs accurate functional derivatives, since ultimately the density must be determined by solving an Euler equation instead of the KS equations. The standard XC approximations, which use local and semi-local forms, do not have accurate derivatives,[11] but this is unimportant for obtaining accurate energies in most present calculations,[12] again because only a small part of the Kohn-Sham potential comes from XC. The solution of an effective one-body potential, albeit mildly inaccurate, usually produces extremely accurate densities. As we discuss within, local and semi-local approximations to T_s also have poor functional derivatives, which, in orbital-free DFT, can lead to highly inaccurate densities.

Approximating T_s has proven to be a difficult task for both extended and finite systems.[8] Some approaches build on Thomas-Fermi theory with gradient expansions or the von Weizsäcker KE.[6,7,13] Others use generalized gradient approximations (GGAs),[9,10] some with enhancement factors based on "conjointness."[14] An alternative approach is to use non-local KE functionals[15–19] based on linear response theory. These can produce accuracy comparable to KS DFT for main group metals, but mostly cannot describe covalently bonded materials (although some have not been broadly tested on these materials). For multiple bonds, virtually no KE functional can accurately describe stretched molecules or dissociate diatomics via optimized orbital-free DFT calculations. In general, achieving self-consistent results with an all-electron calculation proves difficult because standard orbital-free KE functionals exhibit poor functional derivatives near the nuclei.[8] Such difficulties can be circumvented through the use of pseudopotentials, effectively softening the singularity at the nuclei.[8,20] In the present work, we use soft-Coulomb interactions (the Coulomb interaction gives a divergent KE in 1d) in analogy to this softening. Thus, we do not test the ability of our method to handle cusps in the density (this will be explored in future work). Only local pseudopotentials can be used in OF DFT, which are less accurate than their non-local counterparts available in KS DFT.

a)Present address: Institute of Physical Chemistry, Department of Chemistry, University of Basel, Klingelbergstr. 80, 4056 Basel, Switzerland.

Introducing KS-derived basis functions and a density matrix in the atomic core region in angular-momentum-dependent OF DFT can recover some of the error introduced by local pseudopotentials.[21] For covalently bonded semiconductor materials, original work of Huang and Carter[22] yielded good results, and a recent extension[20] for covalently bonded molecules achieves moderately accurate optimized dissociation energies and bond lengths for various dimers. Still, self-consistent densities are not very accurate within the bonding region, and it is not clear how the tunable parameters in such approximations would accommodate different, larger molecules.

Dissociating a chemical bond correctly is a particularly difficult problem. Any locally-based approximation has difficulties when the bond length is stretched to large distances. The fragments often contain fractional electron numbers, for which local approximations to XC yield very wrong answers.[23] Similar difficulties beset local and semi-local approximations to T_s. In Sec. II C, we demonstrate this for the KE of a single bond in one dimension, where a local approximation yields fragment energies that are much worse in the stretched limit.

A different approach to this problem is to use machine learning (ML), a powerful tool for learning high-dimensional patterns via induction that has been successful in many applications,[24] including some in quantum chemistry.[25–30] Some of us recently suggested using ML to approximate density functionals.[31] Arbitrary accuracy was achieved by adding enough training data, difficulties with functional derivatives were circumvented, and the predictive variance was used to indicate where the ML approximation to T_s would be accurate. But in that work, all particles were confined to a box, so the difficulties of stretching and even breaking bonds did not arise.

We are interested in applying ML to construct approximations to T_s for general use in electronic structure calculations. This is a totally different way to think about approximating density functionals than the traditional paths described above. ML is a sophisticated interpolation scheme that shines in very high-dimensional spaces. Instead of using exact conditions or starting from local approximations, ML takes results from known examples to predict previously unknown (but similar) cases. Thus, ML cannot produce a transferable approximation satisfying many exact conditions, and is not applicable in totally novel circumstances. Instead, an ML approximation requires examples where the answer is known, and then can only be applied to similar examples where the answer is desired. Since every KS calculation ever run (and each iteration in the self-consistent cycle of every KS calculation) produces a density and its exact T_s, this seems like an ideal situation for attempting to apply ML to density functional construction.

ML approximations are conceptually very different from the usual approach. Superficially, they may appear similar to the popular practice (for XC functionals) of fitting approximate forms to multiple data sets.[32,33] However, the difference in our approach is that ML fits the shape of the functional *itself*, not any human-constructed approximate form, such as a local or semi-local form, or a two-point response function. The resulting approximate ML functional may contain 10^6 parameters, which will change depending on the training data. It may be impossible to prove that even the most basic exact condition for T_s, namely, that it should always be positive, is always satisfied by such a functional. In fact, there is no reason that it must, as the functional need only be accurate on the space of densities that it is designed to be accurate for. Understanding and evaluating ML approximations to density functionals will require developing an entirely new set of tools.

The present paper is again a proof of principle, but one aimed at chemical rather than physical problems. Perhaps the most outstanding failure of explicit density functionals has always been the stretching and breaking of chemical bonds. This failure has been highlighted for XC approximations,[34] but is even more of a problem for KE approximations. In fact, it is worse in our 1d examples than in 3d. Thus, here we are not testing our ML approximation for general use in real molecules. We are testing to see if ML approximations show any particular difficulty in breaking bonds, the bane of the usual approaches.

In this work, we construct an orbital-free KE functional based on ML that is capable of accurately describing 1d diatomics from the united atom limit to complete nuclear separation. Moreover, we obtain accurate ground-state densities and molecular forces, and accuracy is systematically improvable with more training densities. The culmination of this work is shown in Fig. 9, where we apply ML algorithms self-consistently to fermions in two wells as they are separated: ML appears to have no difficulty in reproducing the KS DFT result for a situation that is very challenging with traditional orbital-free KE functional approximations.

II. THEORY AND BACKGROUND

Throughout this paper, we use atomic units ($e^2 = \hbar = \mu = 1$). Unless labeled otherwise, energies are given in Hartrees and lengths are given in Bohr radii. All equations are written for 1d systems, and $n(x)$ is assumed to be a spin-unpolarized electron density.

A. KS DFT

The Hohenberg-Kohn theorems[1] allow writing the total ground-state energy of the system as a minimum of a functional of the electron density n,

$$E_v[n] = T_s[n] + V[n] + U[n] + E_{xc}[n], \tag{1}$$

where $T_s[n]$ is the non-interacting KE, $V[n]$ is the energy associated with the one-body potential $v(x)$ (e.g., the interaction of the electron with external fields such as nuclei), $U[n]$ is the Hartree electron-electron repulsion energy, and $E_{xc}[n]$ is the XC energy. In KS DFT, T_s is computed by solving an auxiliary non-interacting system (the KS system).[5] The KS equations for a 1d system are

$$\left\{-\frac{1}{2}\frac{\partial^2}{\partial x^2} + v_s[n](x)\right\}\phi_j(x) = \epsilon_j\phi_j(x), \tag{2}$$

where ϵ_j and ϕ_j are the KS energy levels and orbitals. The KS potential is

$$v_{\rm S}[n](x) = v(x) + v_{\rm H}[n](x) + v_{\rm XC}[n](x), \qquad (3)$$

where $v_{\rm H}[n](x)$ is the Hartree potential, and

$$v_{\rm XC}[n](x) = \frac{\delta E_{\rm XC}[n]}{\delta n(x)}. \qquad (4)$$

For ease, we consider only doubly occupied orbitals, so the electron density is given by

$$n(x) = 2\sum_{j=1}^{N/2} |\phi_j(x)|^2. \qquad (5)$$

Given some approximation for $E_{\rm XC}[n]$, this is a closed set of equations that must be solved self-consistently to give the ground-state density which, inserted in (1), yields an approximation to the ground-state energy.

B. Model system

In the present work, we consider a one-dimensional model of a diatomic molecule, where we replace the true Coulomb potential by a soft form,[35] so that the electron repulsion is

$$v_{\rm ee}(|x - x'|) = \frac{1}{\sqrt{1 + |x - x'|^2}}, \qquad (6)$$

while the attraction to the nuclei of charges Z_α and Z_β is

$$v(x) = -\frac{Z_\alpha}{\sqrt{1 + |x + R/2|^2}} - \frac{Z_\beta}{\sqrt{1 + |x - R/2|^2}} \qquad (7)$$

and the internuclear repulsion is $V_{\rm NN}(R) = Z_\alpha Z_\beta/\sqrt{1 + R^2}$, where R is the distance between the two nuclei. For example, H_2 is modeled by $Z_\alpha = Z_\beta = 1$ and $N = 2$, and LiH is given by $Z_\alpha = 3$, $Z_\beta = 1$, and $N = 4$.

We perform KS DFT calculations using the local density approximation (LDA) for XC. For the 1d gas of soft-Coulomb interacting electrons, the LDA exchange energy is given in Ref. 35 and an accurate parametrization of the correlation energy is given in Ref. 36. As mentioned above, the local approximation causes the well-known failure of the molecule to dissociate into the correct fragments, as the symmetry is restricted to a singlet. This is irrelevant to our purposes here, except to make our 1d calculation more realistic. Our ML KE functional is designed to reproduce the KS LDA results, not correct the failure of LDA XC. A perfect model for $T_{\rm s}$ would reproduce KS LDA exactly (if the exact XC functional were readily available,[35] then a perfect model for $T_{\rm s}$ would reproduce the exact binding curve).

Numerically, we solve the KS equations in a discrete variable representation (DVR)[37] generated by a unitary transformation of the eigenfunctions ξ_j of the 1d (non-interacting) harmonic oscillator (with $\omega = 1$), chosen for its simplicity. The transformation is found by diagonalizing the position operator,

$$X^{\rm DVR} = UXU^\dagger, \qquad (8)$$

where $X_{ij} = \langle \xi_i|\hat{x}|\xi_j\rangle$ for $i, j = 1, \ldots, N_G$, where N_G is the number of basis functions used. The DVR basis is

$$\theta_\alpha(x) = \sum_{j=1}^{N_G} U_{\alpha j}\xi_j(x) \qquad (9)$$

for $\alpha = 1, \ldots, N_G$. This method is closely related to a Hermite-Gaussian quadrature, as the quadrature points x_α are given by the diagonal of $X^{\rm DVR}$. The DVR basis has the special property $\theta_\alpha(x_\beta) = \delta_{\alpha\beta}/\sqrt{w_\beta}$, so the quadrature weights are given by $w_\alpha = 1/\theta_\alpha(x_\alpha)^2$. The expansion of any function $f(x)$ in the DVR basis is given by the values of f at the quadrature points,

$$f(x) \approx \sum_\alpha \sqrt{w_\alpha} f(x_\alpha)\theta_\alpha(x), \qquad (10)$$

and the integral of $f(x)$ on the interval $(-\infty, \infty)$ can be approximated by

$$\int_{-\infty}^{\infty} f(x)\, dx \approx \sum_\alpha w_\alpha f(x_\alpha). \qquad (11)$$

For a more thorough discussion of DVR, see Ref. 37. To find the KS energies and orbitals, we diagonalize the Hamiltonian of the KS system in the DVR basis,

$$H^{\rm DVR} = UTU^\dagger + V_{\rm s}^{\rm DVR}, \qquad (12)$$

where $V_{{\rm s},\alpha\beta}^{\rm DVR} = v_{\rm s}[n](x_\alpha)\delta_{\alpha\beta}$, and the matrix elements of the kinetic energy operator, $T_{ij} = \langle \xi_i| - \frac{1}{2}\frac{\partial^2}{\partial x^2}|\xi_j\rangle$, are computed analytically. The electron density is given by

$$n(x) = \sum_\alpha \sqrt{w_\alpha} n(x_\alpha)\theta_\alpha(x). \qquad (13)$$

Note that only the value of the densities at the quadrature points x_α need be computed and stored,

$$n(x_\alpha) = 2\sum_{j=1}^{N/2} |\phi_j(x_\alpha)|^2. \qquad (14)$$

We compute reference KS LDA energies and densities for 1d H_2, He_2, Li_2, Be_2, and LiH, for nuclear separation R between 0 and 10. The range of kinetic energies for all systems are shown in Fig. 1. With $N_G = 100$, the errors in all

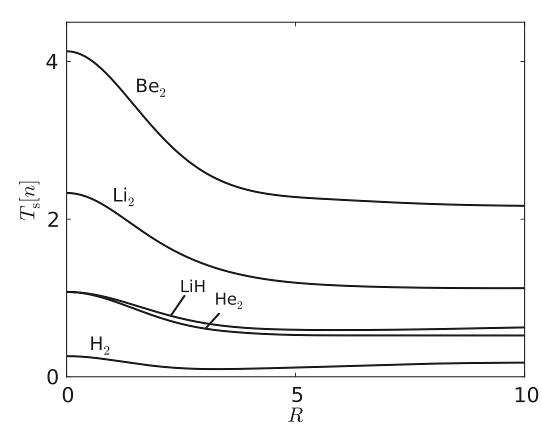

FIG. 1. KS kinetic energies (in Hartrees) for 1d soft-Coulomb models of H_2, He_2, Li_2, Be_2, and LiH, for nuclear separations between 0 and 10.

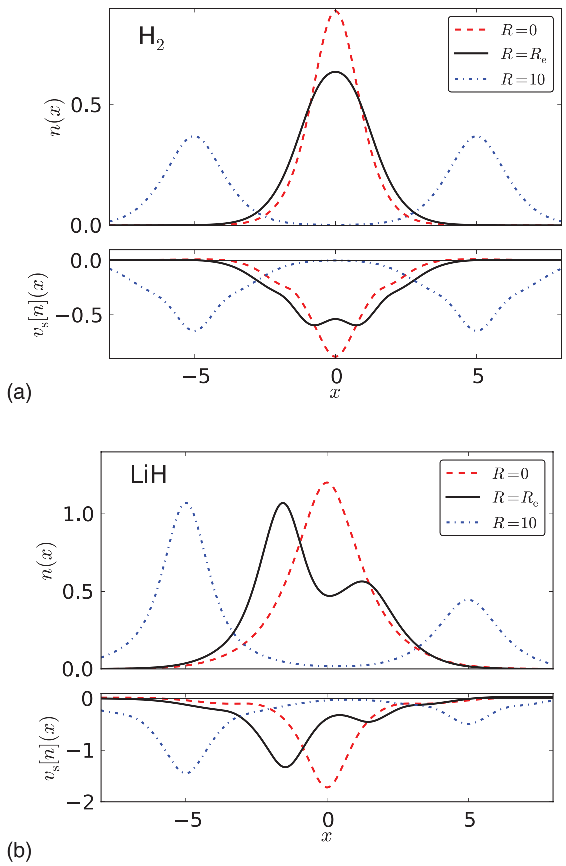

FIG. 2. The 1d soft-Coulomb model for (a) H_2 ($Z_\alpha = Z_\beta = 1$, $N = 2$) and (b) LiH ($Z_\alpha = 3$, $Z_\beta = 1$, $N = 4$). The KS electronic density $n(x)$ and the corresponding KS potential $v_s[n](x)$ are shown at $R = 0$ (dashed), equilibrium bond length R_e (solid), and nearly dissociated $R = 10$ (dotted dashed). R_e values are given in Table I. Values given in atomic units.

reference energies are less than 10^{-7}. Fig. 2 shows the densities and potentials for the united atom, equilibrium bond length, and stretched diatomic for H_2 and LiH. Fig. 9 shows the LDA binding curve of H_2 and LiH. Additionally, we extract equilibrium bond lengths R_e, vibrational frequencies ω_e, and dissociation energies D_0 (which we calculate as the difference in molecular energies between $R = 10$ and $R = R_e$, minus the zero-point vibrational energy), listed in Table I.

C. Orbital-free DFT

In orbital-free DFT, T_s is approximated directly as a functional of n. The ground-state density is found by the constrained minimization

$$\delta\left\{E_v[n] - \mu\left(\int n(x)\,dx - N\right)\right\} = 0, \tag{15}$$

where the chemical potential μ is adjusted to produce the required particle number N. For the KS system, this becomes simply

$$\frac{\delta T_s[n]}{\delta n(x)} = \mu - v_s[n](x). \tag{16}$$

At self-consistency, the functional derivative of the KE is negative the KS potential (up to a constant). This equation can be solved directly for the ground state density $\tilde{n}$—no orbitals are required. Depending on the approximation to T_s, the functional derivative may be ill behaved at the nuclei for real molecular systems, making it difficult or impossible to solve self-consistently. This problem can be avoided by using pseudopotentials, ameliorating the $1/r$ divergence at the nuclei.

Let $\tilde{T}_s$ be an approximate T_s, yielding an approximate $\tilde{E}_v[n]$. There are two tests of an approximate $\tilde{T}_s$. The weaker test is to evaluate $\tilde{T}_s$ on the KS density n and compute the error $\Delta E_F \equiv \tilde{E}_v[n] - E_v[n] = \tilde{T}_s[n] - T_s[n] = \Delta T_s[n]$, where $\tilde{E}_v$ is given by inserting $\tilde{T}_s$ into Eq. (1). This is called the functional-driven error, and is the error due to the approximate functional. Note that, for $N = 2$, the exact non-interacting KE is simply the von Weizsäcker,

$$T^{\mathrm{VW}}[n] = \int dx\,\frac{n'(x)^2}{8n(x)}. \tag{17}$$

The more stringent test is to insert $\tilde{T}_s$ into Eq. (16), find the minimizing density $\tilde{n}$ for the given potential, and compute the error $\Delta E = \tilde{E}_v[\tilde{n}] - E_v[n]$. The density-driven error, defined by $\Delta E_D = \Delta E - \Delta E_F = \tilde{E}_v[\tilde{n}] - \tilde{E}_v[n]$, measures how much additional error is introduced by self-consistency. In general, we seek an approximation to T_s that passes both tests, producing accurate densities and small ΔE_D and ΔE_F. For example, the simplest density functional approximation to T_s is the local approximation, which for spin-unpolarized densities in 1d is

$$T_s^{\mathrm{loc}}[n] = \frac{\pi^2}{24}\int dx\, n^3(x), \tag{18}$$

Fig. 3 shows the functional- and density-driven errors of T_s^{loc} for H_2, highlighting the difficulty the local approximation has as the diatomic is stretched. T_s^{loc} overestimates the KE by 31% at $R = 0$, but in the stretched limit gives a 67% underestimate. This is because the density of H_2 is spin-unpolarized, while each isolated fragment should give the same result as a spin-polarized calculation. Additionally, solving self-consistently roughly doubles the error in the total energy. More sophisticated approximations to T_s, such as a GGA, are built on T_s^{loc} and thus suffer from the same problem.

TABLE I. Reference dissociation energies (in kcal/mol), equilibrium bond lengths (in Bohr radii), and vibrational frequencies (in cm^{-1}) for our 1d diatomic models, computed with LDA. Note that He_2 and Be_2 are unbound.

System	D_0 (kcal/mol)	R_e	ω_e (cm^{-1})
H_2	131	1.63	2960
Li_2	16.6	5.06	272
LiH	67.3	2.95	1460

D. Data topology and representation

In ML, the data typically have a finite representation. For example, in predicting atomization energies from molecular structures,[27] each molecule was represented by its Coulomb

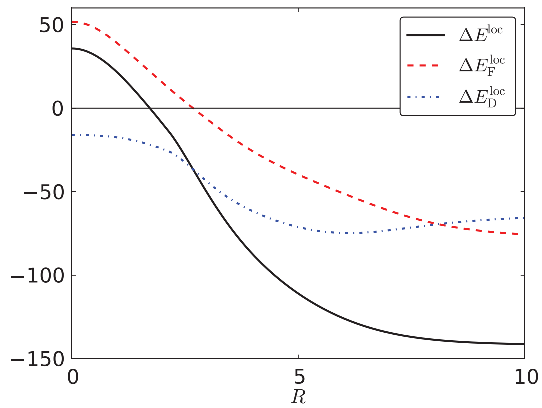

FIG. 3. Errors in total energy (in kcal/mol) of H_2 as a function of nuclear separation R, using the local approximation to the KE $T_{\rm s}^{\rm loc}[n]$, split into the functional- and density-driven components $\Delta E_{\rm F}^{\rm loc}$ and $\Delta E_{\rm D}^{\rm loc}$.

matrix, derived from the charges and positions of the nuclei. In contrast, the electronic density n is a continuous function restricted to the domain

$$\mathcal{J}_N \equiv \left\{ n \,\middle|\, n(x) \geq 0,\, n^{1/2}(x) \in H^1(\mathbb{R}),\ \ \int_{-\infty}^{\infty} n(x)\,dx = N \right\}, \tag{19}$$

where $H^1(\mathbb{R})$ is a Sobolev space.[38] Although $\mathcal{J}_N$ is infinite dimensional, in practice n is expanded in a finite basis (with N_G basis functions). In this work, the discrete variable representation is used, where in previous work[31] we used a real space grid. We use the L^2 inner product and norm between densities n, m,[39]

$$\langle n, m\rangle = \int_{-\infty}^{\infty} dx\, n(x)m(x), \qquad \|n\| = \sqrt{\langle n, n\rangle}. \tag{20}$$

Since the ML method used is expressed in terms of this inner product, the results are independent of the specific representation used as long as the basis is converged.

Even with a truncated basis, $\mathcal{J}_N$ is still high-dimensional, and applying ML to learn the KE of all densities in $\mathcal{J}_N$ would not be feasible. Fortunately, we are only interested in a specific class of potentials (e.g., soft-Coulomb), which greatly reduces the variety of possible densities. In general, let the potential v be parametrized by the variables $\{p_1, \ldots, p_d\}$. We define the density manifold $\mathcal{M}_N \subset \mathcal{J}_N$ as the set of all densities that come from these potentials with a given particle number N. In general, $\mathcal{M}_N$ is a d-dimensional manifold. The training densities n_j, for $j = 1, \ldots, N_T$, are sampled from $\mathcal{M}_N$. For our diatomics, the nuclear separation R is the only variable and thus $d = 1$ (where in previous work[31] our potential had nine variables).

E. Kernel ridge regression (KRR)

KRR is a nonlinear form of regression (i.e., least squares fitting) with regularization to prevent overfitting.[40–42] KRR is a data driven method for interpolating nonlinear functions that is capable of handling high-dimensional functions. As in previous work,[31] our machine learning approximation (MLA) for the KE is

$$T_{\rm s}^{\rm ML}[n] = \sum_{j=1}^{N_T} \alpha_j k[n, n_j], \tag{21}$$

where α_j are weights to be determined, n_j are training densities, N_T is the number of training densities, and k is the kernel, which measures similarity between densities. We choose the Gaussian kernel

$$k[n_i, n_j] = \exp\left(-\frac{\|n_i - n_j\|^2}{2\sigma^2}\right), \tag{22}$$

where σ is called the length scale. The weights are found by minimizing the cost of quadratic error plus regularization

$$\mathcal{C}(\boldsymbol{\alpha}) = \sum_{j=1}^{N_T} \left(T_{\rm s}^{\rm ML}[n_j] - T_{\rm s}[n_j]\right)^2 + \lambda \boldsymbol{\alpha}^T \boldsymbol{K} \boldsymbol{\alpha}, \tag{23}$$

where $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_{N_T})$ and $\boldsymbol{K}$ is the kernel matrix, $\boldsymbol{K}_{ij} = k[n_i, n_j]$. The second term penalizes large weights to prevent overfitting. Minimizing $\mathcal{C}(\boldsymbol{\alpha})$ gives

$$\boldsymbol{\alpha} = (\boldsymbol{K} + \lambda \boldsymbol{I})^{-1} \boldsymbol{T}, \tag{24}$$

where $\boldsymbol{I}$ is the identity matrix and $\boldsymbol{T} = (T_{\rm s}[n_1], \ldots, T_{\rm s}[n_{N_T}])$.

The global parameters σ and λ are determined through leave-one-out cross validation. For each density n_i in the training set, the functional is trained on all other densities n_j ($j \neq i$) and σ and λ are optimized by minimizing the absolute error on n_i. Final values are chosen as the median over all optimum values. To find each optimum, we search over a coarse logarithmic grid in σ and λ. In general, there may be local minima near the global minimum, but typically this area is shallow and so the model is not particularly sensitive to variation in the parameters. Thus we expect some slight variation in the performance of the model depending on the details of the cross validation. To test performance, the functional is evaluated on new densities not in the training set.

III. RESULTS AND DISCUSSION

A. Errors on KS densities

To construct an MLA for this problem, we choose N_T training densities at evenly spaced R between 0 and 10 (inclusive). Table II reports the functional driven error of the MLA, $\Delta E_F = T_{\rm s}^{\rm ML}[n] - T_{\rm s}[n]$, evaluated on KS LDA densities. The global parameters optimized via cross validation are also listed. The performance of the MLA increases rapidly and uniformly with training set size, for both H_2 and LiH. For $N_T = 20$, all errors are much smaller than those due to the XC approximation. In Table III, we fix N_T to either 10 or 20 and compare the performance of the MLA across a variety of systems (H_2, He_2, Li_2, Be_2, and LiH). Errors are roughly similar across the board, independent of molecular size, number of electrons, or bond type or lack thereof (the variation in errors is most likely due to the cross validation not fully optimizing the global parameters). Finally, we combine the training sets for all 5 systems into one MLA, to demonstrate the ability of ML to handle more than one kind of diatomic in the same approximation. Errors are roughly the average of the individual

TABLE II. Parameters and errors (in kcal/mol) for H_2 and LiH with increasing number of training densities N_T. Mean and max values taken over 199 test densities evenly spaced from $R = 0$ to 10 (inclusive) (although the test densities at $R = 0$, 10 are also in the training set, they contribute negligibly to the errors). $\Delta T_{\rm SF}$ is the pure functional-driven error of the MLA, evaluated on KS densities. $\Delta T_{\rm S}$ and ΔE represent errors on constrained optimal densities. The errors in dissociation energy D_0 (corrected for zero-point vibrational energy), equilibrium bond length R_e, and zero-point vibrational frequency ω_e are all calculated on constrained optimal densities (ΔR_e and $\Delta\omega_e$ are given as percentages). The global parameters σ and λ are calculated via cross validation. Note that although $T_s[n]$ is a universal functional, an interpolation of it is not. The length scale and noise level of the model depend on how closely the training densities are spaced.
[a]Due to lack of training densities, we were unable to converge constrained optimal densities near $R = 0$ (see Fig. 9(a)), thus these are ignored for calculating mean and max values.

		$\lvert\Delta T_{\rm SF}\rvert$		$\lvert\Delta T_{\rm S}\rvert$		$\lvert\Delta E\rvert$			(%)	(%)		
Mol.	N_T	Mean	Max	Mean	Max	Mean	Max	ΔD_0	ΔR_e	$\Delta\omega_e$	σ	λ
H_2	6[a]	2.9	12	5.3	64	4.7	29	−2.3	−6.0	−15	2.6	1.3×10^{-3}
	10	0.40	1.8	1.6	16	0.53	2.5	−2.0	1.1	−4.0	6.6	3.2×10^{-6}
	20	0.001	0.005	0.008	0.13	0.002	0.018	−0.002	−0.015	0.003	2.5	1.1×10^{-10}
LiH	6	9.3	35	7.5	33	26	95	−57	−14	31	5.2	4.8×10^{-3}
	10	0.92	5.4	1.5	10	1.2	7.5	−4.3	5.1	−2.5	4.6	1.1×10^{-4}
	20	0.006	0.032	0.026	0.36	0.015	0.18	−0.015	−0.023	0.77	6.1	2.6×10^{-9}

errors for each molecule (although max errors appear to be slightly larger).

B. How much consistency is too much?

Thus far, we have only discussed results evaluating the ML functional on exact densities. In real applications of orbital-free DFT, these would not be available. It is a basic tenet in DFT (and other areas of quantum mechanics) that one should use the variational principle to find the "best" result. In DFT in general, practical calculations employ some approximation to the functional, and then the density is found self-consistently using the variational principle.[5] For the non-interacting KS problem, the minimizing density for a given v is the one that satisfies the Euler equation (Eq. (16)), and we call its solution the self-consistent density.

Of course, the exact $T_{\rm s}$ will produce the exact density (for the given XC approximation), but any approximation to $T_{\rm s}$ yields only an approximate density. There are many computational and conceptual advantages to using this procedure, such as extracting forces via the Hellmann-Feynman theorem. But there is no guarantee that this procedure, applied to an approximation, will yield either an accurate energy or density. One might guess that applying an approximate functional to an exact density would always yield a more accurate result, but this is *not* typically the case for KS calculations with approximate XC.[12]

In fact, inserting $T_{\rm s}^{\rm ML}$ into Eq. (16) and solving for the self-consistent density yields a highly inaccurate energy and density (just as we found in Ref. 31). Fig. 4 illustrates the problem: a gradient descent quickly leaves the "interpolation" region for which the MLA is accurate. Why is this the case? The MLA only knows how the KE changes along the density manifold $\mathcal{M}_N$. In all dimensions orthogonal to $\mathcal{M}_N$ (which is $N_G - d \gg 1$), the MLA has no information, and thus produces an inaccurate derivative. As shown in Fig. 5, the functional derivative of the MLA in this case is very wrong!

In Fig. 4, a minimization with the exact KE functional $T_{\rm s}$ would yield the correct self-consistent density $\tilde{n}$, which is also on the density manifold. Thus, one solution would be to replace the minimization in Eq. (15) (over all densities

TABLE III. Same as Table. II, but for more molecules. "All" indicates that we have conglomerated the training data from all 5 systems into one model, with N_T training densities per molecule. Mean and max values are taken over the test data from all 5 systems as well.

		$\lvert\Delta T_{\rm SF}\rvert$		$\lvert\Delta T_{\rm S}\rvert$		$\lvert\Delta E\rvert$			
N_T	Mol.	Mean	Max	Mean	Max	Mean	Max	σ	λ
10	H_2	0.40	1.8	1.6	16	0.53	2.5	6.6	3.2×10^{-6}
	He_2	0.95	3.9	2.7	26	1.5	5.8	4.9	5.5×10^{-5}
	Li_2	2.0	10	8.6	100	3.4	19	4.5	7.4×10^{-5}
	Be_2	0.61	3.3	3.6	53	1.2	9.6	10	1.4×10^{-6}
	LiH	0.92	5.4	1.5	10	1.2	7.5	4.6	1.1×10^{-4}
	All	1.2	5.5	1.8	41	1.5	17	13	1.4×10^{-8}
20	H_2	0.001	0.005	0.008	0.13	0.002	0.018	2.5	1.1×10^{-10}
	He_2	0.012	0.042	0.037	0.45	0.015	0.12	6.4	1.1×10^{-10}
	Li_2	0.005	0.031	0.036	0.61	0.016	0.34	9.3	1.1×10^{-9}
	Be_2	0.019	0.054	0.085	1.3	0.023	0.16	11	5.0×10^{-10}
	LiH	0.006	0.032	0.026	0.36	0.015	0.18	6.1	2.6×10^{-9}
	All	0.003	0.072	0.039	8.3	0.008	0.43	7.1	1.2×10^{-13}

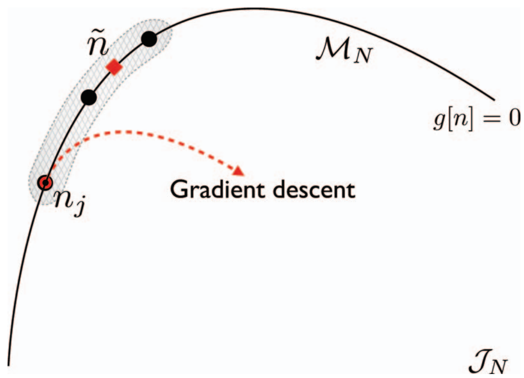

FIG. 4. A cartoon illustrating the difficulty in solving for the self-consistent density with our MLA. Pictured are the density manifold $\mathcal{M}_N$ (curved solid line), the training densities $n_j \in \mathcal{M}_N$ (black circles), and the KS self-consistent density $\tilde{n}$ (red square). Here g is any function that is zero on $\mathcal{M}_N$ and positive elsewhere, so that $\mathcal{M}_N$ is defined implicitly by $g[n] = 0$. The shaded area shows the interpolation region for which the MLA is accurate. The red dashed line shows the solution of Eq. (16) via gradient descent. Because the functional derivative of the MLA is inaccurate orthogonal to $\mathcal{M}_N$, the minimization soon leaves the interpolation region and becomes unstable.

normalized to N particles) by one over $\mathcal{M}_N$,

$$\delta\left\{E_v[n] + \zeta g[n]\right\} = 0, \tag{25}$$

where ζ is a Lagrange multiplier, and g is any function that is zero on $\mathcal{M}_N$ and positive elsewhere (thus $\mathcal{M}_N$ is defined implicitly by $g[n] = 0$). Note that the original constraint of normalization to N particles is additionally satisfied by densities on $\mathcal{M}_N$. Assuming the self-consistent density to be on the density manifold is valid, since we intend to apply our MLA to systems similar to those trained on.[43] We call the solution of this equation a constrained optimal density, which is analogous to the self-consistent solution of Eq. (15), and satisfies $g[\tilde{n}] = 0$. To see that this constraint can yield an accurate minimizing density, we project the functional derivative onto the tangent space of $\mathcal{M}_N$ at n, which is given as all densities n' satisfying $\langle \delta g[n]/\delta n, n' - n\rangle = 0$. From this we can construct an orthogonal basis, $u_j[n](x)$ for $j = 1, \ldots, d$, for the tangent

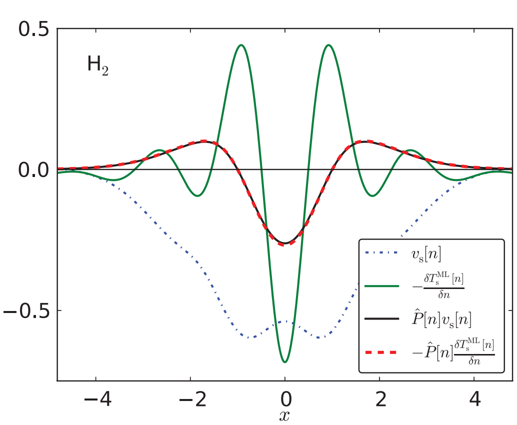

FIG. 5. The functional derivative of our MLA (green) cannot reproduce the exact derivative $v_\mathrm{S}[n]$ (blue dotted dashed) evaluated at the ground-state density, because this information is not contained in the data. However, both agree when projected onto the tangent of the data manifold $\mathcal{M}_N$ at n (black and red dashed). Shown for H_2 at equilibrium bond length $R_e = 1.63$, in atomic units.

space at n, and form the projection operator

$$\hat{P}[n] = \sum_{j=1}^{d} u_j[n] \otimes u_j[n], \tag{26}$$

where $\otimes$ is the tensor product defined by $(a \otimes b)c = \langle a, c\rangle b$. For our diatomic model system, $d = 1$ and the exact projection onto the tangent space is given by $P[n] = u_1[n] \otimes u_1[n]$, where $u_1[n] = (\partial n_R(x)/\partial R)/\|\partial n_R(x)/\partial R\|$ and $n_R(x)$ is the density that comes from the diatomic with nuclear separation R. Fig. 5 shows excellent agreement between the projected functional derivatives, evaluated at the KS density. This demonstrates that the MLA captures the correct derivative of KE along $\mathcal{M}_N$. Thus a gradient descent constrained to $\mathcal{M}_N$ should give an accurate constrained optimal density.

However, we are not given $g[n]$. We must attempt to reconstruct $\mathcal{M}_N$ from the training densities, approximating $g[n]$. In previous work,[31] we used a local principal component analysis (PCA) to approximate $g[n]$. In the present work, however, the density manifold exhibits a higher curvature, yielding inaccurate constrained optimal densities. Thus, we use a more sophisticated approximation based on kernel PCA,[44] called nonlinear gradient denoising (NLGD). A brief summary of NLGD is given in the Appendix. The full derivation is given in Ref. 45. In Section III C, we describe how to solve Eq. (25) for constrained optimal densities.

C. Projected gradient descent algorithm

A schematic of the projected gradient descent is shown in Fig. 6. Given an approximate $g[n]$ and an initial guess for the density $\tilde{n}_0 \in \mathcal{M}_N$ (e.g., one of the training densities), the algorithm is as follows:

1. In a projection step, we first compute the projection operator $\hat{P}[n]$ onto the tangent space of $\mathcal{M}_N$ at $\tilde{n}_t$. Next, a step is taken to lower the energy,

$$\tilde{n}'_t(x) = \tilde{n}_t(x) - \epsilon \hat{P}[\tilde{n}_t]\left(\left.\frac{\delta T_\mathrm{s}^\mathrm{ML}[n]}{\delta n(x)}\right|_{n=\tilde{n}_t} + v_\mathrm{s}[\tilde{n}_t](x)\right), \tag{27}$$

where ϵ is a small positive constant.

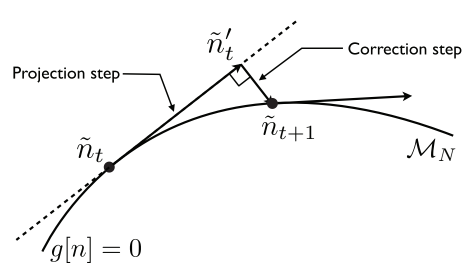

FIG. 6. A schematic of the projected gradient descent. First, we project the functional derivative onto the tangent space of the data manifold $\mathcal{M}_N$ at $\tilde{n}_t$ (dashed line). Next, we take a step along the projected functional derivative to $\tilde{n}'_t$ to lower the energy. Finally, we minimize $g[n]$ orthogonal to the tangent space to ensure the minimization stays on $\mathcal{M}_N$.

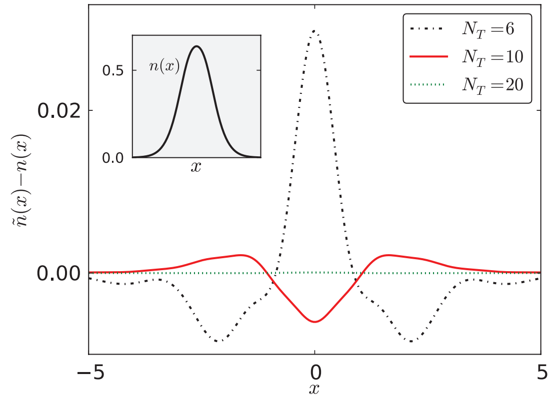

FIG. 7. Difference between the constrained optimal density $\tilde{n}(x)$ and the KS density $n(x)$ for various numbers of training densities N_T. The error decreases uniformly for all x. The system is H_2 at equilibrium bond length. The inset shows the KS density.

2. In a correction step, we minimize $g[n]$ orthogonal to the tangent space of the previous step, starting from $\tilde{n}'_t$. Let $\tilde{n}_{t+1}$ be the minimizing density.

We iterate these two steps until convergence is achieved (the changes in energy and density are smaller than the desired tolerance). Details are given in the Appendix.

D. Errors of constrained optimal densities

To illustrate how well this constrained minimization (Eq. (25)) works, in Fig. 7 we plot the error in the constrained optimal densities for several values of N_T, for H_2 at equilibrium bond length R_e. Clearly, we approach an accurate constrained optimal density rapidly with increasing training data. In Table II, we list mean and max errors (in both KE and total energy), as well as errors in dissociation energy D_0 (corrected for zero-point vibrational energy), equilibrium bond length R_e and zero-point vibrational frequency ω_e, of the MLA evaluated on constrained optimal densities. To see how these errors vary over R, in Fig. 8 we plot the error in total energy for H_2 and LiH, $\Delta E = E^{\mathrm{ML}}[\tilde{n}] - E[n]$, where n is the exact density and $\tilde{n}$ is the constrained optimal density. As we did for T_s^{loc} in Fig. 3, this is split into the functional-driven error, $\Delta E_{\mathrm{F}} = E^{\mathrm{ML}}[n] - E[n]$, and the density-driven error, $\Delta E_{\mathrm{D}} = E^{\mathrm{ML}}[\tilde{n}] - E^{\mathrm{ML}}[n]$.[12] For smaller R, ΔE_{F} is largest because $T_{\mathrm{s}}[n]$ is changing most rapidly with R, and we train on data at fixed separations in R. As R grows, ΔE_{F} becomes much smaller. Once the MLA has achieved high accuracy, density-driven errors become negligible. We do not use the model between the last two training points because the density-driven error becomes much larger (an order of magnitude), presumably due to lack of training data beyond $R = 10$.

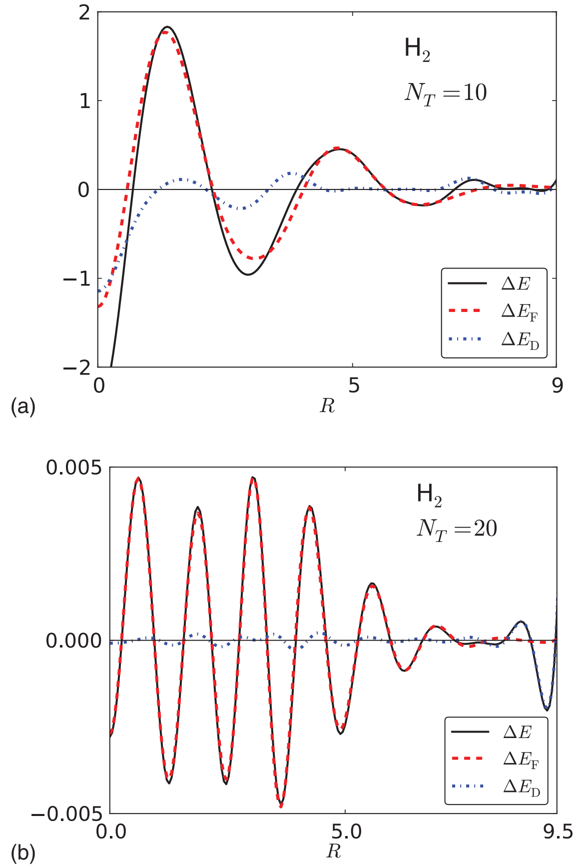

FIG. 8. The total error of the model and the functional- and density-driven errors ΔE_F and ΔE_D for H_2 with (a) 10 and (b) 20 training densities.

Finally, Fig. 9 shows the molecular binding curves for H_2 (for various N_T) and LiH, evaluated on constrained optimal densities. The binding curve for H_2 rapidly converges to the

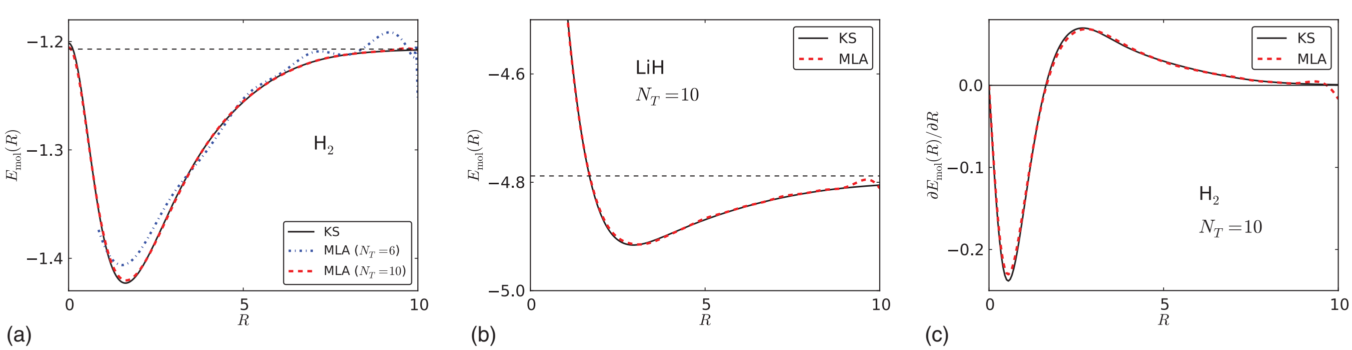

FIG. 9. (a) The molecular binding curve for our 1d model of H_2. The MLA curve is found on constrained optimal densities, for various number of KS densities (spaced evenly from $R = 0$ to 10) for training. The MLA curve converges quickly to the KS curve, and is indistinguishable by $N_T = 15$. The incorrect large R limit is due to deficiencies in LDA XC, not our MLA KE. (b) The molecular binding curve for 1d LiH, for 10 training densities. (c) Molecular forces as a function of R for H_2 with 10 training densities. Derivatives are calculated via finite-difference. All values are in atomic units.

KS curve. By $N_T = 15$, the curves are indistinguishable to the eye. The same is true for He_2, Li_2, and Be_2 (not shown). The figure also shows molecular forces, calculated via finite-difference. The forces are very accurate and should be suitable for, e.g., an *ab initio* molecular dynamics calculation.

IV. CONCLUSIONS

In the present work, we have shown how ML can produce a kinetic energy functional that allows an orbital-free calculation of bond breaking for a simple model system. We have also shown how highly accurate constrained optimal densities can be found with such a model, despite inaccurate functional derivatives. The more training data are used, the more accurate the results become. In this toy model of diatomic molecules, 20 training points yielded chemical accuracy, even on constrained optimal densities. The systematic improvability of this method is sorely lacking in traditional density functional approximations.

We expect results of a similar quality if several simple real (3d) diatomics were stretched. But beyond this, i.e., for large molecules with many bonds of different types and many internal degrees of freedom, one needs to find a compact representation of the density that is invariant with respect to symmetry operations, such as translation and rotation. Two important questions for future work are: how will the amount of training data required to achieve chemical accuracy scale with molecular size, and how can we ensure that the training densities are sampled uniformly over the Born-Oppenheimer surface and restricted to the region of interest. For example, in an *ab initio* molecular dynamics simulation of a molecule, we want to build a model that trains on ground-state densities with energies up to a certain threshold. Such a model would be useful in running long-time molecular dynamics simulations. In scaling the method up to much larger systems (i.e., millions of atoms), we envision taking advantage of near-sightedness, i.e., combining our method with a partitioning scheme such as partition DFT[46] and using ML to learn the interaction energy between fragments. We are currently working on applications to simple atoms and ions, as well as to more complex systems, and pursuing several approaches to answering these questions. Ultimately, only realistic calculations in a quantum chemical code can determine if ML provides a useful solution to the orbital-free DFT problem. The results shown here are both a necessary condition and a promising second step.

ACKNOWLEDGMENTS

The authors thank IPAM at UCLA for hospitality and acknowledge NSF Grant No. CHE-1240252 (J.S., K.B.), EU PASCAL2 and DFG Grant Nos. MU 987/4-2 and MU 987/17-1 (K.H., K.R.M.), Einstein Foundation (K.R.M.), EU Marie Curie Grant No. IEF 273039 (M.R.), and NRF Korea Grant BK21 (K.R.M.).

APPENDIX: NONLINEAR GRADIENT DENOISING

When the training densities are not sampled densely enough on $\mathcal{M}_N$, a local PCA approximation is not sufficient. In this case, kernel PCA can provide a better reconstruction of $\mathcal{M}_N$. We call this method NLGD.[45,47]

In kernel-based methods in ML, the inputs are mapped to a higher or infinite-dimensional vector space called "feature space."[40] Let $\Phi[n]$ be the map to feature space associated with the kernel k. Then, the kernel is equivalent to the inner product in feature space:

$$k[n, n'] = \langle \Phi[n], \Phi[n'] \rangle. \tag{A1}$$

Φ is given implicitly by the choice of kernel. Kernel PCA is simply applying PCA to the data mapped into feature space. Since PCA can be expressed solely in terms of the inner product, $\Phi[n]$ is not needed explicitly in kernel PCA. There, one performs an eigendecomposition of the kernel matrix $\boldsymbol{K}$, with elements $\boldsymbol{K}_{ij} = k[n_i, n_j]$,

$$\tilde{\boldsymbol{K}} \boldsymbol{\alpha}_j = N_T \lambda_j \boldsymbol{\alpha}_j, \tag{A2}$$

where $\tilde{\boldsymbol{K}} = (\boldsymbol{I} - \boldsymbol{1}/N_T)\boldsymbol{K}(\boldsymbol{I} - \boldsymbol{1}/N_T)$ is the centered kernel matrix, $\boldsymbol{\alpha}_j$ are the eigenvectors normalized by $\|\boldsymbol{\alpha}_j\| = 1/\sqrt{N_T \lambda_j}$, λ_j are eigenvalues ordered from largest to smallest magnitude for $j = 1, \ldots, N_T$, $\boldsymbol{I}$ is the $N_T \times N_T$ identity matrix and $\boldsymbol{1}$ is the $N_T \times N_T$ matrix whose entries are all 1. Kernel PCA is performed on all the given training densities. The kernel chosen is the same used throughout the paper, given in Eq. (22), but the length scale used is different: we choose 4 times the median over all nearest neighbor L^2-distances between training densities. The principal components (PCs) in feature space are

$$\boldsymbol{v}_i = \sum_{j=1}^{N_T} \boldsymbol{\alpha}_{i,j} \tilde{\Phi}[n_j], \tag{A3}$$

where $\tilde{\Phi}[n] = \Phi[n] - \sum_{j=1}^{N_T} \Phi[n_j]/N_T$. The projection in feature space onto the first q PCs is

$$\hat{Q} = \sum_{i=1}^{q} \boldsymbol{v}_i \boldsymbol{v}_i^T. \tag{A4}$$

We choose q such that we keep all nonzero eigenvalues λ_j in this projection. The kernel PCA projection error is

$$p[n] = \|(1 - \hat{Q})\tilde{\Phi}[n]\|^2. \tag{A5}$$

Let H be the Hessian of $p[n]$,

$$H[n](x, x') = \frac{\delta p[n]}{\delta n(x) \delta n(x')}. \tag{A6}$$

We find the eigenvalues and eigenfunctions of H,

$$\langle H[n], u_j[n] \rangle = \beta_j u_j[n]. \tag{A7}$$

If a given density n is on $\mathcal{M}_N$, then the eigenfunctions $u_j[n](x)$ with small $|\beta_j|$ form a basis for the tangent space of $\mathcal{M}_N$ at n. There will be a sudden increase in β_j after d eigenvalues, where d is the dimensionality of $\mathcal{M}_N$. The projection operator onto the tangent plane is given by

$$\hat{P}^{\mathrm{NLGD}}[n] = \sum_{j=1}^{d} u_j[n] \otimes u_j[n], \tag{A8}$$

where $a \otimes b$ denotes the tensor product defined by $(a \otimes b)c = \langle a, c \rangle b$. Finally, we define $g^{\mathrm{NLGD}}[n]$ as the squared L^2-norm of the functional derivative of $p[n]$ orthogonal to the tangent space of $\mathcal{M}_N$ at n,

$$g^{\mathrm{NLGD}}[n] = \left\| (1 - \hat{P}^{\mathrm{NLGD}}[n]) \frac{\delta p[n]}{\delta n} \right\|^2. \tag{A9}$$

These approximations to $g[n]$ and $\hat{P}[n]$ are then used in the projection algorithm of Sec. III C. The derivation of this approximation and further details are given in Ref. 45.

[1] P. Hohenberg and W. Kohn, "Inhomogeneous electron gas," Phys. Rev. **136**, B864–B871 (1964).
[2] W. Kohn and L. J. Sham, "Self-consistent equations including exchange and correlation effects," Phys. Rev. **140**, A1133–A1138 (1965).
[3] K. Burke, "Perspective on density functional theory," J. Chem. Phys. **136**, 150901 (2012).
[4] K. Burke and L. O. Wagner, "Dft in a nutshell," Int. J. Quantum Chem. **113**, 96–101 (2013).
[5] R. M. Dreizler and E. K. U. Gross, *Density Functional Theory: An Approach to the Quantum Many-Body Problem* (Springer, 1990).
[6] Y. A. Wang and E. A. Carter, "Orbital-free kinetic-energy density functional theory," in *Theoretical Methods in Condensed Phase Chemistry*, edited by S. D. Schwartz (Kluwer, New York, 2000).
[7] V. V. Karasiev, D. Chakraborty, and S. B. Trickey, "Progress on new approaches to old ideas: Orbital-free density functionals," in *Many-Electron Approaches in Physics, Chemistry, and Mathematics*, edited by L. Delle Site and V. Bach (Springer Verlag, Kluwer, NY) (to be published).
[8] V. Karasiev and S. B. Trickey, "Issues and challenges in orbital-free density functional calculations," Comput. Phys. Commun. **183**, 2519–2527 (2012).
[9] V. V. Karasiev, R. S. Jones, S. B. Trickey, and F. E. Harris, "Properties of constraint-based single-point approximate kinetic energy functionals," Phys. Rev. B **80**, 245120 (2009).
[10] V. V. Karasiev, R. S. Jones, S. B. Trickey, and F. E. Harris, "Erratum: Properties of constraint-based single-point approximate kinetic energy functionals [Phys. Rev. B **80**, 245120 (2009)]," Phys. Rev. B **87**, 239903 (2013).
[11] C. J. Umrigar and X. Gonze, "Accurate exchange-correlation potentials and total-energy components for the helium isoelectronic series," Phys. Rev. A **50**, 3827–3837 (1994).
[12] M.-C. Kim, E. Sim, and K. Burke, "Understanding and reducing errors in density functional calculations," Phys. Rev. Lett. **111**, 073003 (2013).
[13] C. F. V. Weizsäcker, "Zur theorie der kernmassen," Z Phys. **96**, 431–458 (1935).
[14] F. Tran and T. A. Wesolowski, "Link between the kinetic- and exchange-energy functionals in the generalized gradient approximation," Int. J. Quantum Chem. **89**, 441–446 (2002).
[15] E. Chacón, J. E. Alvarellos, and P. Tarazona, "Nonlocal kinetic energy functional for nonhomogeneous electron systems," Phys. Rev. B **32**, 7868–7877 (1985).
[16] P. García-González, J. E. Alvarellos, and E. Chacón, "Nonlocal kinetic-energy-density functionals," Phys. Rev. B **53**, 9509–9512 (1996).
[17] P. García-González, J. E. Alvarellos, and E. Chacón, "Nonlocal symmetrized kinetic-energy density functional: Application to simple surfaces," Phys. Rev. B **57**, 4857–4862 (1998).
[18] L.-W. Wang and M. P. Teter, "Kinetic-energy functional of the electron density," Phys. Rev. B **45**, 13196–13220 (1992).
[19] Y. A. Wang, N. Govind, and E. A. Carter, "Orbital-free kinetic-energy density functionals with a density-dependent kernel," Phys. Rev. B **60**, 16350–16358 (1999).
[20] J. Xia, C. Huang, I. Shin, and E. A. Carter, "Can orbital-free density functional theory simulate molecules?" J. Chem. Phys. **136**, 084102 (2012).
[21] Y. Ke, F. Libisch, J. Xia, L.-W. Wang, and E. A. Carter, "Angular-momentum-dependent orbital-free density functional theory," Phys. Rev. Lett. **111**, 066402 (2013).
[22] C. Huang and E. A. Carter, "Nonlocal orbital-free kinetic energy density functional for semiconductors," Phys. Rev. B **81**, 045206 (2010).
[23] A. J. Cohen, P. Mori-Sánchez, and W. Yang, "Insights into current limitations of density functional theory," Science **321**, 792–794 (2008).
[24] K.-R. Müller, S. Mika, G. Rätsch, K. Tsuda, and B. Schölkopf, "An introduction to kernel-based learning algorithms," IEEE Trans. Neural Network **12**, 181–201 (2001).
[25] O. Ivanciuc, "Applications of support vector machines in chemistry," in *Reviews in Computational Chemistry*, edited by Kenny Lipkowitz and Tom Cundari (Wiley, Hoboken, 2007), Vol. 23, Chap. 6, pp. 291–400.
[26] A. P. Bartók, M. C. Payne, R. Kondor, and G. Csányi, "Gaussian approximation potentials: The accuracy of quantum mechanics, without the electrons," Phys. Rev. Lett. **104**, 136403 (2010).
[27] 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).
[28] M. Rupp, M. R. Bauer, R. Wilcken, A. Lange, M. Reutlinger, F. M. Boeckler, and G. Schneider, "Machine learning estimates of natural product conformational energies," PLOS Comput. Biol. (in press).
[29] Z. D. Pozun, K. Hansen, D. Sheppard, M. Rupp, K.-R. Müller, and G. Henkelman, "Optimizing transition states via kernel-based machine learning," J. Chem. Phys. **136**, 174101 (2012).
[30] G. Montavon, M. Rupp, V. Gobre, A. Vazquez-Mayagoitia, K. Hansen, A. Tkatchenko, K.-R. Müller, and O. A. von Lilienfeld, "Machine learning of molecular electronic properties in chemical compound space," New J. Phys. **15**, 095003 (2013).
[31] J. C. Snyder, M. Rupp, K. Hansen, K.-R. Müller, and K. Burke, "Finding density functionals with machine learning," Phys. Rev. Lett. **108**, 253002 (2012).
[32] Y. Zhao and D. Truhlar, "The M06 suite of density functionals for main group thermochemistry, thermochemical kinetics, noncovalent interactions, excited states, and transition elements: Two new functionals and systematic testing of four M06-class functionals and 12 other functionals," Theor. Chem. Acc. **120**, 215–241 (2008).
[33] V. Petzold, T. Bligaard, and K. W. Jacobsen, "Construction of new electronic density functionals with error estimation through fitting," Top. Catal. **55**, 402–417 (2012).
[34] P. Mori-Sanchez, A. J. Cohen, and W. Yang, "Discontinuous nature of the exchange-correlation functional in strongly correlated systems," Phys. Rev. Lett. **102**, 066403 (2009).
[35] L. O. Wagner, E. M. Stoudenmire, K. Burke, and S. R. White, "Reference electronic structure calculations in one dimension," Phys. Chem. Chem. Phys. **14**, 8581–8590 (2012).
[36] N. Helbig, J. I. Fuks, M. Casula, M. J. Verstraete, M. A. L. Marques, I. V. Tokatly, and A. Rubio, "Density functional theory beyond the linear regime: Validating an adiabatic local density approximation," Phys. Rev. A **83**, 032503 (2011).
[37] J. C. Light and T. Carrington, "Discrete-variable representations and their utilization," in *Advances in Chemical Physics* (John Wiley & Sons, Inc., 2007), pp. 263–310.
[38] R. A. Adams and J. J. F. Fournier, *Sobolev Spaces*, Pure and Applied Mathematics (Elsevier Science, 2003).
[39] In real calculations all densities are represented by a finite basis, and thus will have a finite L^2 norm.
[40] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning. Data Mining, Inference, and Prediction*, 2nd ed. (Springer, New York, 2009).
[41] K. Hansen, G. Montavon, F. Biegler, S. Fazli, M. Rupp, M. Scheffler, O. A. von Lilienfeld, A. Tkatchenko, and K.-R. Müller, "Assessment and validation of machine learning methods for predicting molecular atomization energies," J. Chem. Theory Comput. **9**, 3404–3419 (2013).
[42] G. Montavon, M. Braun, T. Krueger, and K.-R. Müller, "Analyzing local structure in kernel-based learning: Explanation, complexity, and reliability assessment," IEEE Signal Process. Mag. **30**, 62–74 (2013).
[43] In situations where this is not the case (e.g., incomplete sampling of $\mathcal{M}_N$), the predictive variance could be used to detect when leaving the interpolation domain and further reference calculations could then be done to augment the training set accordingly.
[44] B. Schölkopf, A. Smola, and K. R. Müller, "Nonlinear component analysis as a kernel eigenvalue problem," Neural Comput. **10**, 1299–1319 (1998).
[45] J. C. Snyder, M. Rupp, K. Hansen, K.-R. Müller, and K. Burke, "Accurate densities from inaccurate functional derivatives" (unpublished).
[46] P. Elliott, K. Burke, M. H. Cohen, and A. Wasserman, "Partition density-functional theory," Phys. Rev. A **82**, 024501 (2010).
[47] J. C. Snyder, S. Mika, K. Burke, and K.-R. Müller, "Kernels, pre-images and optimization," in *Empirical Inference – Festschrift in Honor of Vladimir N. Vapnik*, edited by Bernhard Schoelkopf, Zhiyuan Luo, and Vladimir Vovk (Springer, Heidelberg) (to be published).