By-passing the Kohn-Sham equations with machine learning
Supplemental Information

Felix Brockherde,[1,2] Leslie Vogt,[3] Li Li,[4] Mark E. Tuckerman,[3,5,6] Kieron Burke,[7,4,*] and Klaus-Robert Müller[1,8,*]

[1] *Machine Learning Group, Technische Universität Berlin, Marchstr. 23, 10587 Berlin, Germany*
[2] *Max-Planck-Institut für Mikrostrukturphysik, Weinberg 2, 06120 Halle, Germany*
[3] *Department of Chemistry, New York University, New York, NY 10003, USA*
[4] *Departments of Physics and Astronomy, University of California, Irvine, CA 92697, USA*
[5] *Courant Institute of Mathematical Science, New York University, New York, NY 10003, USA*
[6] *NYU-ECNU Center for Computational Chemistry at NYU Shanghai, 3663 Zhongshan Road North, Shanghai 200062, China*
[7] *Departments of Chemistry, University of California, Irvine, CA 92697, USA*
[8] *Department of Brain and Cognitive Engineering, Korea University, Anam-dong, Seongbuk-gu, Seoul 136–713, Republic of Korea*

(Dated: March 1, 2017)

KERNEL RIDGE REGRESSION

Kernel Ridge Regression[1, 2] (KRR) is a machine learning method for regression. We introduce the method for abstract training points (x_i, y_i), i.e. features $x_1, \dots x_M \in \mathbb{R}^d$ and associated labels $\mathbf{Y} = (y_1, \dots, y_M)^T \in \mathbb{R}^M$ and describe the actual models used in the main text afterwards. We want to model a function $f : \mathbb{R}^d \to \mathbb{R}$ that maps from features to labels. This model should not be 'learned by heart' but perform well on unseen data (i.e. *generalize*). We first restrict the set of possible functions to the reproducing kernel Hilbert space (RKHS) $\mathcal{H}$ on the space of discretized densities that is induced by the Gaussian kernel function

$$k(x, x') = \exp\left(-\frac{||x - x'||^2}{2\sigma^2}\right). \tag{1}$$

The restriction is very mild and rather technical; more interesting is the choice of the kernel function which determines the scalar product (and thus the norm) of the RKHS. Leaving rigor aside, the Gaussian kernel induces an RKHS norm $||f||_{\mathcal{H}}$ that is smaller for simpler, smoother functions and higher for more complicated, oscillating functions. We minimize the empirical risk functional

$$\mathcal{C}(f) = \sum_{i=1}^{M} |y_i - f(x_i)|^2 + \lambda \|f\|^2_{\mathcal{H}} \tag{2}$$

that defines a trade-off between error on the training points and smoothness of the function controlled by the hyper-parameter λ.

The representer theorem[3] allows us to assume that the solution to Eq. 2 is given by a linear combination of kernel functions $f = \sum_{i=1}^{M} \boldsymbol{\alpha}_i k(x_i, \cdot)$. It now suffices to solve

$$\mathcal{C}(\boldsymbol{\alpha}) = \sum_{i=1}^{M} |y_i - f(x_i)|^2 + \lambda \|f\|^2_{\mathcal{H}} \tag{3}$$

$$= \sum_{i=1}^{M} |y_i - f(x_i)|^2 + \lambda \boldsymbol{\alpha}^\intercal \mathbf{K} \boldsymbol{\alpha}, \tag{4}$$

where $\mathbf{K}_{ij} = k(x_i, x_j)$ is the kernel matrix. The solution is given by

$$\boldsymbol{\alpha} = (\mathbf{K} - \lambda \mathbf{I})^{-1} \mathbf{Y}. \tag{5}$$

Note that all model parameters and hyper-parameters are estimated on the training set; the hyper-parameter choice makes use of standard cross-validation procedures (see Hansen *et al.* [4]). Once the model is fixed after training, it is applied unchanged out-of-sample.

We use this method for various maps:

Non-interacting kinetic energy functional ($T_s^{ML}[n]$, 1-D). The training points are given by pairs of densities and associated kinetic energies. We discretize the densities and use them in vectorial form, i.e. $n \in \mathbb{R}^G$. Thus, the *functional* $\mathcal{L}^2 \to \mathbb{R}$ is modeled as a *function* $\mathbb{R}^{\mathbb{G}} \to \mathbb{R}$

ML-OF map (1-D). The training points are given by pairs of discretized 1-D box potentials and associated total energies.

ML-KS map (3-D). The training points are given by pairs of discretized Gaussians potentials (as described in the main text) and total energies.

Total energy functional ($E^{ML}[n]$, 3-D). The training points are given by pairs of densities in basis function representation (see below) and associated total energies. Just as for T_s^{ML}, this *functional* is modeled as a *function*.

ML HOHENBERG-KOHN MAP

The basis representation for the densities is given by

$$n(x) = \sum_{l=1}^{L} u^{(l)} \phi_l(x), \tag{6}$$

where ϕ_l are the L basis functions. We introduce some notation and continue to write the density in grid representation as n, and its basis coefficients as u. We can then write the HK map model as

$$n^{\mathrm{ML}}[v](x) = \sum_{l=1}^{L} u^{(l)}[v] \phi_l(x), \tag{7}$$

where the L basis function coefficients are regular KRR models,

$$u^{(l)}[v] = \sum_{i=1}^{M} \beta_i^{(l)} k(v, v_i), \tag{8}$$

of external potentials v with a Gaussian kernel function. The contribution of the error to the cost function can be formulated as

$$e(\boldsymbol{\beta}) = \sum_{i=1}^{M} \| n_i - n^{\mathrm{ML}}[v_i] \|_{\mathcal{L}_2}^2 \tag{9}$$

$$= \sum_{i=1}^{M} \left\| n_i - \sum_{l=1}^{L} \sum_{j=1}^{M} \beta_j^{(l)} k(v_i, v_j) \phi_l \right\|_{\mathcal{L}_2}, \tag{10}$$

with the $\mathcal{L}_2$ norm. We write this cost function in terms of basis function coefficients. This can be viewed as projecting the inside of the norm on each basis function. Assuming orthogonality of the basis functions yields

$$e(\boldsymbol{\beta}) = \sum_{i=1}^{M} \sum_{l=1}^{L} \left| u_i^{(l)} - \sum_{j=1}^{M} \beta_j^{(l)} k(v_i, v_j) \right|^2. \tag{11}$$

where $u_i^{(l)} = \langle n_i, \phi_l \rangle$ is the l-th basis function coefficient of the i-th training density, as defined in Eq. 6 if orthogonality is satisfied. After reordering the sums over i and l, we view each l independently and solve analogously to regular KRR

$$\boldsymbol{\beta}^{(l)} = \left(\mathbf{K}_{\sigma^{(l)}} + \lambda^{(l)} \mathbf{I} \right)^{-1} \mathbf{u}^{(l)}, \quad l = 1, \ldots, L \tag{12}$$

where, for each basis function l, $\lambda^{(l)}$ is a regularization parameter, $\mathbf{K}_{\sigma^{(l)}}$ is a Gaussian kernel with kernel width $\sigma^{(l)}$. The $\lambda^{(l)}$ and $\sigma^{(l)}$ can be chosen individually for each basis function via independent cross-validation (see [4, 5]).

BASIS FUNCTIONS

Fourier basis. We define the basis as

$$\phi_l(x) = \begin{cases} \cos\{2\pi x(l-1)/2\}, & l \text{ odd} \\ \sin\{2\pi x l/2\}, & l \text{ even} \end{cases} \quad l = 1, \ldots, L. \tag{13}$$

We transform the density efficiently via the discrete Fourier transform

$$u_i^{(l)} = \sum_{m=1}^{G} n_i(x_m) \phi_l(x_m). \tag{14}$$

The back-projection is written as

$$n^{\mathrm{ML}}[v](x) = \sum_{l=1}^{L} u^{(l)}[v] \phi_l(x). \tag{15}$$

KPCA basis. We define the basis as:

$$\phi_l^{\mathrm{KPCA}} = \sum_{j=1}^{M} p_j^{(l)} \Phi(n_j). \tag{16}$$

The parameters $p_j^{(l)}$ are found by eigen-decomposition of the Kernel matrix. The KCPA basis coefficients are given by

$$u_i^{(l)} = \langle \Phi(n_i), \phi_l^{\mathrm{KPCA}} \rangle = \sum_{j=1}^{M} p_j^{(l)} k(n_j, n_i) \tag{17}$$

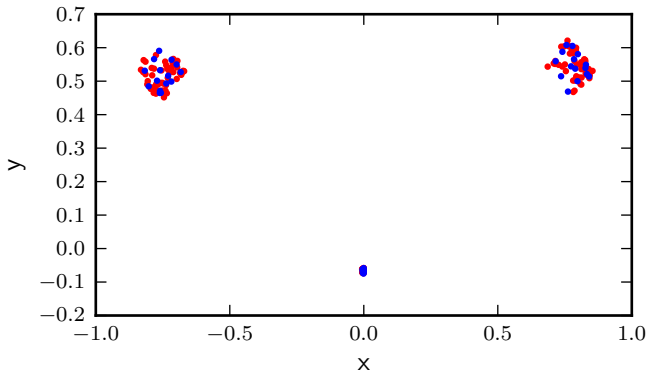
with kernel map Φ. The back-projection for KPCA is not trivial but several solutions exist. We follow Bakir *et al.* [6] and learn the back-projection map.

GRADIENT DESCENT ISSUES

There are two ways to remedy problems of the gradient descent procedure: First, the gradient descent step can be "de-noised" by projecting the gradient onto the data manifold and thus removing the noisy directions. Secondly, the directions outside of the data manifold can be removed in a preprocessing step to get rid of the influence of the noisy directions on the gradient completely. Both methods yield similar results.

Several approaches exist for describing and projecting onto the data manifold. Common to each approach is the idea to find principle components and to project on

Figure 1. The extent of the H_2O dataset. The figure shows the atom coordinates in angstrom. Blue are atoms from 15 training points, red from 50 test points.



those in which direction the densities have largest variance. Best results are reported [7] by using Kernel Principle Component Analysis[8] (KPCA), a non-linear generalization of PCA.

There are three issues with the assumed gradient-based approaches: First, the correct choice of the number of (K)PCA components K has to be made. It is generally possible to view it as a hyper-parameter and find the optimal K via cross-validation. However, we can not choose fractional Ks. One K might be not enough and $K+1$ too much information. Second, the data points only lie in a bounded region of a manifold that can be described via PCA components. It is still possible for the gradient descent to walk outside this bounded region toward a point where the model has no information and thus the gradients become inaccurate. A (K)PCA method that only accesses the scalar products between points in the data set can not solve this[9]. Third, it might not be possible to find a suitable pre-image for a ground-state density given by (K)PCA coefficients[10].

MOLECULAR DATASETS

The extent of the dataset for H_2O is visualized in Fig. 1. In this case, conformers were generated from random displacements from the optimized geometry.

For benzene and ethane, conformers were generated from isothermal molecular dynamics (MD) trajectories. The range of atomic positions from combined 1 ns 300 K and 350 K trajectories is shown in Fig. 2 for benzene and Fig. 3 for ethane after snapshots are aligned to a reference molecule. For malonaldehyde, the classical MD trajectories include 0.5 ns for each tautomer at each temperature. Resulting conformers that are used to create the K-means sampled training set are shown in Fig. 4. The test set is taken from an ab initio MD trajectory at 300 K.

Figure 2. The extent of the benzene conformers generated by MD (red points). K-means sampling is used to select 2,000 representative points. Test points from an independent trajectory are in blue and are offset for clarity.

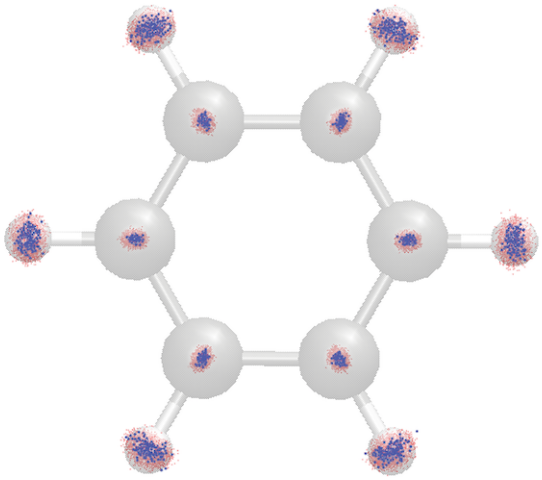
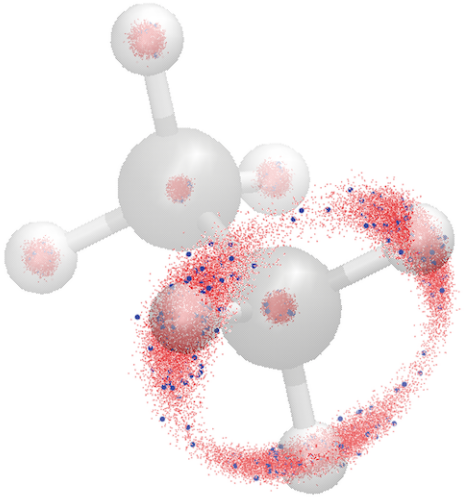


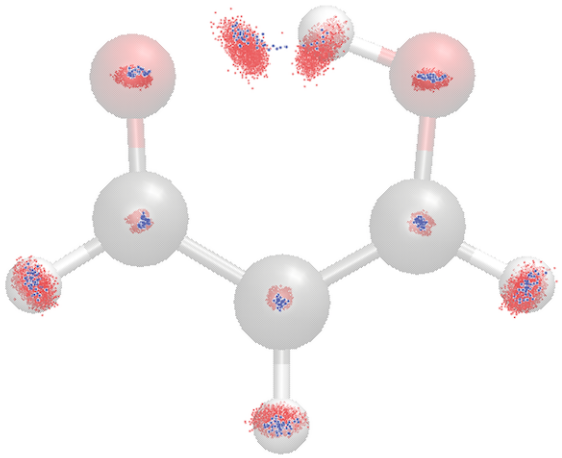
Figure 3. The extent of the ethane conformers generated by MD (red points). K-means sampling is used to select 2,000 representative points. Test points from an independent trajectory are in blue.



DFT CONVERGENCE

For our 3-D DFT calculations in Quantum Espresso[11], we center a water molecule in a cubic cell and converge three variables: the kinetic energy

Figure 4. The extent of the malonaldehyde conformers generated by MD (red points). K-means sampling is used to select 2,000 representative points. Test points from an independent ab initio MD trajectory are in blue and are offset for clarity.



cutoff for wavefunctions `ecutwfc` in steps of 10 Ry, the kinetic energy cutoff for charge density and potential `ecutrho` in steps of 40 Ry, and the cell dimension `celldm` in steps of 1 bohr. We increase parameters until increasing any parameter does not change the equilibrium position total energy by more than 0.01 kcal/mol for H_2O. We end up with `ecutwfc` of 90 Ry, `ecutrho` of 360 Ry, and `celldm` of 20 bohr, which are used for all other molecules in this work.

SAMPLING

For H_2, since there is only one atomic distance to adjust, we take the M equi-distant points in the parameter range and for each of these points select the training point that is closest.

For larger molecules with more parameters (H_2O, Benzene, Ethane, Malonaldehyde) we also want to cover the conformer space in a way that all conformers are relatively close to at least one training point.

Assuming p_i are the parameters of conformer i and $i \in \tilde{P}_j$ if and only if $\tilde{p}_j$ is closest to p_i, we want to find $\tilde{p}_j$, $j = 1 \dots M$ that minimize

$$\sum_{j=1}^{M} \sum_{i \in \tilde{P}_j} ||\tilde{p}_j - p_i||^2. \tag{18}$$

K-means[12] solves this problem for continuous $\tilde{p}_j$. However, since K-means returns only locally optimal solutions, we rerun the algorithm 50 times and select the solution which minimizes Eq. 18. We choose the points p_i closest to each $\tilde{p}_j$ as training points.

LOGIC OF DENSITY FUNCTIONAL THEORY (DFT)

Within the Born-Oppenheimer approximation in non-relativistic quantum mechanics, and using atomic units, the Hohenberg-Kohn paper[13] laid the theoretical framework of all modern DFT. The first statement is that the mapping

$$v(\mathbf{r}) \longleftrightarrow n(\mathbf{r}) \tag{19}$$

is one-to-one, i.e., at most one potential can give rise to a given ground-state density, even in a quantum many-body problem, for given interaction among particles and statistics (i.e., fermions or bosons). A follow-up claim is that the ground-state energy of an electronic system can be found from

$$E[v] = \min_n \left\{ F[n] + \int d^3r\, n(\mathbf{r}) v(\mathbf{r}) \right\} \tag{20}$$

where $F[n]$ is a density functional containing all many-body effects. The minimizing density is the solution to the Euler equation:

$$\frac{\delta F}{\delta n(\mathbf{r})} + v(\mathbf{r}) = \text{const} \tag{21}$$

It is the direct map between densities and potentials that we machine-learn in this paper. We call it the HK density map, $n[v](\mathbf{r})$.

The KS scheme avoids direct approximation of F by imagining a fictitious system of non-interacting electrons with the same density as the real one[14]. The KS equations are:

$$\left\{ -\frac{1}{2}\nabla^2 + v_\mathrm{s}(\mathbf{r}) \right\} \phi_i(\mathbf{r}) = \epsilon_i \phi_i(\mathbf{r}) \tag{22}$$

where ϵ_i are the KS eigenvalues and ϕ_i the KS orbitals.

$$v_\mathrm{s}(\mathbf{r}) = v(\mathbf{r}) + v_\mathrm{H}(\mathbf{r}) + v_\mathrm{XC}(\mathbf{r}) \tag{23}$$

where $v_\mathrm{H}(\mathbf{r})$ is the Hartree potential and $v_\mathrm{XC}(\mathbf{r})$ is the exchange-correlation potential. The true energy of the system is then reconstructed from the self-consistent density $n(\mathbf{r}) = \sum_i |\phi_i(\mathbf{r})|^2$ via

$$E[n] = T_\mathrm{s}[n] + U[n] + \int d^3r\, n(\mathbf{r}) v(\mathbf{r}) + E_\mathrm{XC}[n] \tag{24}$$

where $T_s[n]$ is the kinetic energy of the non-interacting electrons and $U[n]$ is the Hartree energy. $E_{XC}[n]$ is the exchange-correlation (XC) energy and implicitly defined by Eq. 24. Most calculations[15] use simple approximations that depend only on the density and its gradient to determine E_{XC}, called generalized gradient approximations, or replace a fixed fraction of the approximate exchange with the exact exchange from a Hartree-Fock calculation (called a hybrid). Requiring the XC potential to be the functional derivative of E_{XC} ensures that the self-consistent solution of Eq. 22 minimizes the energy of Eq. 24 for the given $v(\mathbf{r})$ and $E_{XC}[n]$.

* to whom correspondence should be addressed.

[1] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning - Data Mining, Inference, and Prediction*, 2nd ed., Springer Series in Statistics (Springer, 2009).

[2] K. Vu, J. C. Snyder, L. Li, M. Rupp, B. F. Chen, T. Khelif, K.-R. Müller, and K. Burke, International Journal of Quantum Chemistry **115**, 1115 (2015).

[3] B. Schölkopf, R. Herbrich, and A. J. Smola, "A generalized representer theorem," in *Computational Learning Theory: 14th Annual Conference on Computational Learning Theory, COLT 2001 and 5th European Conference on Computational Learning Theory, EuroCOLT 2001 Amsterdam, The Netherlands, July 16–19, 2001 Proceedings*, edited by D. Helmbold and B. Williamson (Springer Berlin Heidelberg, Berlin, Heidelberg, 2001) pp. 416–426.

[4] K. Hansen, G. Montavon, F. Biegler, S. Fazli, M. Rupp, M. Scheffler, O. A. von Lilienfeld, A. Tkatchenko, and K.-R. Müller, J. Chem. Theory Comput. **9**, 3404 (2013).

[5] K.-R. Müller, S. Mika, G. Rätsch, K. Tsuda, and B. Schölkopf, IEEE Trans. Neural Netw. **12**, 181 (2001).

[6] G. H. Bakir, J. Weston, and B. Schölkopf, in *Advances in Neural Information Processing Systems*, Vol. 16, edited by S. Thrun, L. K. Saul, and B. Schölkopf (MIT Press, 2004) pp. 449–456.

[7] J. C. Snyder, M. Rupp, K.-R. Müller, and K. Burke, Int. J. Quantum Chem. **115**, 1102 (2015).

[8] B. Schölkopf, A. Smola, and K.-R. Müller, Neural Computation **10**, 1299 (1998).

[9] F. J. Király, M. Kreuzer, and L. Theran, arXiv:1406.2646 [cs, math, stat] (2014).

[10] B. Schölkopf, S. Mika, C. Burges, P. Knirsch, K.-R. Müller, G. Rätsch, and A. Smola, IEEE Trans. Neural Netw. **10**, 1000 (1999).

[11] P. Giannozzi, S. Baroni, N. Bonini, M. Calandra, R. Car, C. Cavazzoni, D. Ceresoli, G. L. Chiarotti, M. Cococcioni, I. Dabo, A. Dal Corso, S. de Gironcoli, S. Fabris, G. Fratesi, R. Gebauer, U. Gerstmann, C. Gougoussis, A. Kokalj, M. Lazzeri, L. Martin-Samos, N. Marzari, F. Mauri, R. Mazzarello, S. Paolini, A. Pasquarello, L. Paulatto, C. Sbraccia, S. Scandolo, G. Sclauzero, A. P. Seitsonen, A. Smogunov, P. Umari, and R. M. Wentzcovitch, Journal of Physics: Condensed Matter **21**, 395502 (19pp) (2009).

[12] H. Steinhaus, Bull. Acad. Polon. Sci. Cl. III. **4**, 801 (1956).

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

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

[15] A. Pribram-Jones, D. A. Gross, and K. Burke, Annual Review of Physical Chemistry **66**, 283 (2015).