doi:10.2533/chimia.2015.182 Chimia 69 (2015) 182–186 © Schweizerische Chemische Gesellschaft

Many Molecular Properties from One Kernel in Chemical Space

Raghunathan Ramakrishnan[§a] and O. Anatole von Lilienfeld*[ab]

[§]SCS-Metrohm Award for best oral presentation

Abstract: We introduce property-independent kernels for machine learning models of arbitrarily many molecular properties. The kernels encode molecular structures for training sets of varying size, as well as similarity measures sufficiently diffuse in chemical space to sample over all training molecules. When provided with the corresponding molecular reference properties, they enable the instantaneous generation of machine learning models which can be systematically improved through the addition of more data. This idea is exemplified for single kernel based modeling of internal energy, enthalpy, free energy, heat capacity, polarizability, electronic spread, zero-point vibrational energy, energies of frontier orbitals, HOMO-LUMO gap, and the highest fundamental vibrational wavenumber. Models of these properties are trained and tested using 112,000 organic molecules of similar size. The resulting models are discussed as well as the kernels' use for generating and using other property models.

Keywords: Chemical Space · Kernel Ridge Regression · Machine learning · Molecular properties · Quantum chemistry

1. Introduction

Strategies for solving computational chemistry problems have evolved in parallel with the capacity and abundance of computer hardware.[1] Access to ever-increasing compute power available at centralized computer facilities, such as the sciCORE at the University of Basel, the Swiss National Supercomputing Centre or the Argonne Leadership Computing Facility, enable 'Big Data' driven computational chemistry which no longer relies on experimental data for training and validation, but rather on virtual data, obtained through predictive modeling and massive simulation efforts. Statistical inference from Big Data holds great promise in many scientific domains, including biology,[2] climate research,[3] high-energy physics,[4] or photonics.[5] Due to their unrivaled computational efficiency (typical execution time is milliseconds), data driven models of molecular property predictions become relevant as soon as they reach the accuracy of well established deductive quantum chemistry methods for solving approximations to the electronic Schrödinger equation, such as Hartree-Fock, Density Functional Theory (DFT), or Coupled-Cluster methods.

For the 'supervised learning' task[6] of inferring a molecular property from structure–property dyads provided *a priori*, machine learning (ML) algorithms have very recently been shown to reach desirable quantum chemical accuracy, even when predicting properties for new (out-of-sample) molecules which had no part in training.[7–11] These developments have also triggered studies on transition state dividing surfaces,[12] orbital-free kinetic density functionals,[13] electronic properties of crystals,[14] transmission coefficients in nano-ribbon models,[15] or densities of states in Anderson impurity models.[16] In order to establish a consistent dataset for which ML models can be improved systematically through the addition of more data, we recently published computed DFT structures and multiple properties of 134,000 organic molecules.[17] Within a previous study,[18] discussed at the Swiss Chemical Society Fall Meeting in 2014, we used some properties of this dataset to investigate the Δ-ML approach that augments less expensive deductive baseline theories by inductive ML models, trained on the baseline's deficiencies. We demonstrated that chemical accuracy can be reached within this Δ-ML approach, as well as strong transferability when applied to all the 134,000 molecules. Here, we investigate the use of property-independent kernels for the simultaneous modeling of multiple properties taken from the same database.[17] To this end, we rely on kernel functions sufficiently diffuse to account for significant similarity measures among all training molecules. This enables us to reapply inverted kernel matrices to any arbitrary set of molecular properties and to generate the corresponding ML models on the same footing. We have validated our approach by simultaneous training and prediction of 13 energetic and electronic molecular properties.

This article is organized as follows: Section 2, Methods, briefly summarizes the ML notations and definitions, along with a discussion of kernel function shape and spread. In Section 3, Computational Details, we describe our molecular data selection strategy, and discuss the selection of properties in the dataset. We present and analyze our results for the performance of property-independent kernels in Section 4. In Section 5 we draw our conclusions. In the appendix we explain how to access and reuse the kernel data.

2. Methods

Inarguably, one of the more appealing ML algorithms is kernel-ridge-regression (KRR)[19] because of its numerical robust-

*Correspondence: Prof. Dr. O. A. von Lilienfeld[ab]
E-mail: anatole.vonlilienfeld@unibas.ch
[a]Institute of Physical Chemistry and National Center for Computational Design and Discovery of Novel Materials (MARVEL)
Department of Chemistry, University of Basel
Klingelbergstrasse 80, CH-4056 Basel
[b]Argonne Leadership Computing Facility
Argonne National Laboratory
9700 S. Cass Avenue, Lemont, IL 60439, USA

ness and conceptual simplicity. Within KRR, the ML *Ansatz* for a given property p of any query molecule, q, is merely a linear combination of similarity measures between q and a finite set of N training molecules t,

$$p_q = \sum_{t=1}^{N} c_t^p K_{qt}. \quad (1)$$

In Eqn. (1) K_{qt} is a kernel matrix element corresponding to molecules q and t. In this study, we wish to investigate if K_{qt} can be made independent of p. We have chosen to use the exponentially decaying (a.k.a. Laplacian) kernel function of the similarity measure between t and q, $K_{qt} = \exp(-D_{tq}/\sigma)$, where $D_{tq} = |\mathbf{d}_t - \mathbf{d}_q|$ is the Manhattan norm of difference between two, typically non-scalar, descriptors of molecules t and q, respectively. The global hyperparameter σ quantifies the kernel width.[19] For the Coulomb matrix (CM) descriptor, the combination of Laplacian kernel with L_1 norm has been shown to yield good ML models of atomization energies.[11]

Prior to predicting molecule q's property p_q according to Eqn. (1), the vector $\mathbf{c}^p$, with optimal regression coefficient c_t^p as weight for every training molecule t, must be obtained through minimization of the penalized Lagrangian function,

$$L = (\mathbf{p}^r - \mathbf{K}\mathbf{c}^p)^{\mathrm{T}}(\mathbf{p}^r - \mathbf{K}\mathbf{c}^p) + \lambda \mathbf{c}^{p\mathrm{T}}\mathbf{K}\mathbf{c}^p \quad (2)$$

where matrices are in upper cases, vectors in lower case, and $()^{\mathrm{T}}$ denotes a transpose. $\mathbf{p}^r$ represents the vector with reference property values for all N training molecules, and $\mathbf{K}$ is the kernel matrix with above defined elements. The λ-term imposes regularization while the first term corresponds to the conventional least square regression. Setting the derivative of L with respect to $\mathbf{c}^p$, to zero, the coefficients which minimize Eqn. (2) can be shown[16] to amount to

$$\mathbf{c}^p = (\mathbf{K} + \lambda\mathbf{I})^{-1}\mathbf{p}^r \quad (3)$$

The dependence of a model's performance on hyperparameters, σ, and λ, can be understood as follows. In the presence of training molecules with extremely outlying properties an optimal value of λ becomes non-zero in order to quench excessively large elements of $\mathbf{c}^p$. In other words, the modeling function becomes more rigid and lessens the danger of overfitting. The meaning of the kernel width, σ, is to control the c_t^p contribution from training molecule t when making a new prediction, see Eqn. (1). Typically, optimal choices of σ, and λ are obtained for every molecular property and training set size through extensive use of cross-validation (CV) within training molecules. CV can become a computational bottleneck for larger training sets. For example, for training sets of $N = 10$ k, a 5-fold CV implies to repeatedly invert 8 k × 8 k matrices, each requiring ~0.5 CPU hours on modern computing hardware.

In this study, we have explored the possibility to always keep all training molecules fixed, to estimate the hyperparameter σ beforehand and to set λ to zero. This allows us to obviate all CVs, and to use a single identical kernel matrix $\mathbf{K}$ for any property. More specifically, once $\mathbf{K}^{-1}$ has been computed and stored, regression coefficients for any number of properties, p_1, p_2, ..., p_n, can be computed simultaneously,

$$\begin{aligned} &[\mathbf{c}^{p_1}\mathbf{c}^{p_2}\ldots\mathbf{c}^{p_n}] = \mathbf{K}^{-1}[\mathbf{p}_1^r\mathbf{p}_2^r\ldots\mathbf{p}_n^r] \\ \Rightarrow \quad &\mathbf{C} = \mathbf{K}^{-1}\mathbf{P}^r, \end{aligned} \quad (4)$$

where $\mathbf{C}$ and $\mathbf{P}^r$ are the $N \times n$ regression coefficient and property matrices, respectively. Consequently, instead of n CVs with computationally demanding multiple kernel matrix inversions, each scaling as O (N^3), the computational cost is now being dominated by one kernel inversion plus n matrix-vector multiplications, each scaling as O (N^2).

3. Computational Details

For typical molecular datasets, with N > 5 k, we find that optimal σ converges towards a large value to ensure non-vanishing contributions from all training compounds while λ converges towards zero. The choice of setting λ to zero is easily justified: As also seen below, property values in the molecular dataset show distributions centered at an average value with relatively few outliers, if any. Over-fitting, leading to large coefficients for these outliers, is therefore unlikely to influence the performance of ML models based on thousands of well-behaved training molecules. For the Laplacian kernel used here, extreme (too large/small) values of σ lead to loss of information regarding training descriptor distances, D_{ij}. For $\sigma \approx 0$, off-diagonal elements of the kernel matrix, $K_{qt} = \exp(-D_{tq}/\sigma)$, vanish, resulting in a unit-kernel-matrix, $\mathbf{K} = \mathbf{I}$. On the other hand, for $\sigma \gg 1$, a kernel-matrix of ones is obtained which would be singular (and hence non-invertible) for $N > 1$ and which does not resolve D_{ij}. Here, we define the optimal σ value to be defined such that all kernel elements are in between 0.5 and 1. This can be accomplished through the constraint for the smallest kernel matrix element, *i.e.* the kernel element which corresponds to the two most distant training molecules,

$$\exp\left(-D_{ij}^{\max}/\sigma_{\mathrm{opt}}\right) = 1/2, \quad (5)$$

resulting in $\sigma_{\mathrm{opt}} = D_{ij}^{max}/\log(2)$. As such, through use of σ_{opt} and $\lambda = 0$, a property-independent global kernel matrix is obtained which only needs to be inverted once before it is used to generate property-dependent regression coefficients for all molecular properties of interest. For randomly sampled 1 k molecules, from the dataset considered in this study (*vide infra*), D_{ij}^{max} = 677 a.u., suggesting $\sigma_{\mathrm{opt}} \approx 977$ a.u. This number is consistent with a numerical grid search for σ_{opt} which identifies $\sigma = 1000$ a.u. to be optimal for a 1 k ML model of atomization enthalpies. In the remainder of this study we discuss the performance of ML models based on inverted global kernels for up to 13 molecular properties, always with $\sigma_{\mathrm{opt}} = 1000$ a.u. and $\lambda = 0$, irrespective of the training set size N.

We have also investigated if our observations depend on the choice of molecular descriptor, $\mathbf{d}$. To this end, we have considered two different descriptors, namely the Coulomb-matrix (CM) with rows and columns uniquely permuted,[9] as well as the bag-of-bonds (BOB)[20] descriptor, amounting to an ordered set of weighted interatomic distances.

We have used the published[17] quantum chemistry results for the smallest 133885 (134 k) organic molecules subset of the GDB-17 published by Reymond *et al.*,[21] which contains over 166 giga molecules. This 134 k dataset contains relaxed geometries and chemical properties computed using the DFT (B3LYP with basis set 6-31G(2df,p)). Here, we have pruned this dataset by eliminating all molecules with up to eight 'heavy' atoms (not counting hydrogens) which, trivially, would be outliers since the dataset is dominated by molecules with nine heavy atoms. For the resulting 111594 (112 k) molecules, we have considered the following 13 computed properties: Zero-Kelvin internal energy, U_0, thermochemistry energetics at 298.15 K (internal energy, U, enthalpy, H, free energy, G, all three properties for the process of atomization, and heat capacity C_v); isotropic molecular polarizability α; electronic radial expectation value $\langle R^2 \rangle$; harmonic zero-point vibrational energy, ZPVE; energy of highest occupied molecular orbital (HOMO), $\varepsilon_{\mathrm{HOMO}}$; energy of lowest unoccupied molecular orbital (LUMO), $\varepsilon_{\mathrm{LUMO}}$; HOMO-LUMO gap, $\Delta\varepsilon$; and the highest fundamental vibrational wavenumber, ω_1, in the 3000–3900 cm^{-1} range.

Fig. 1 features the density distributions of the relative values of these properties,

scaled with respect to their maximal values in the dataset. We note the properties with large count densities to be polarizability, dipole moment, energy of HOMO, and radial expectation value. Compared to these, all thermochemical properties are less densely distributed. Properties even more sparsely distributed include LUMO energy, gap and ZPVE; interestingly their densities also exhibit multimodal distributions (see Fig. 1), possibly arising from characteristic functional group moieties present in the dataset. The distribution of ω_1 shows three narrow peaks, which can readily be interpreted as arising from C–H, N–H, and O–H (symmetric and asymmetric) stretching modes. Corresponding values from literature[22] for similar wavenumbers at the same level of theory read for CH_4 (A_1, 3038, and T_2, 3152 cm^{-1}), NH_3 (A_1, 3459, and E, 3576 cm^{-1}), and H_2O (A_1, 3802, and B_{12}, 3906 cm^{-1}). Further details regarding the genesis of this dataset can be found in ref. [17].

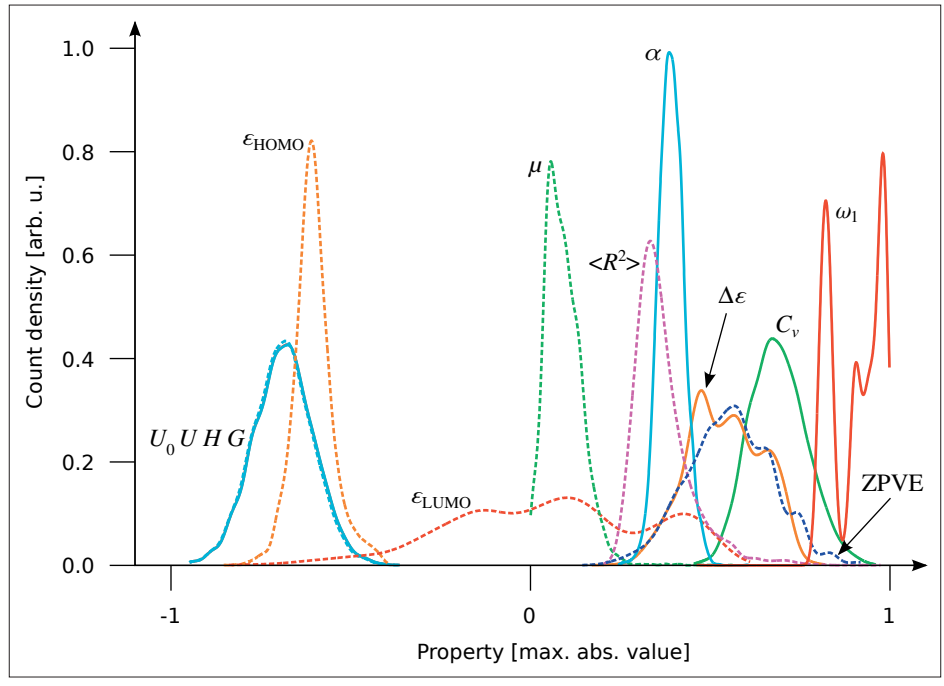

Fig. 1. Density distributions of thirteen properties for the set of 112 k organic molecules made up of CHONF. The abscissa corresponds to values of the properties relative to corresponding maximal absolute values in the dataset: $|U_0|_{max}$ = 2608.5 kcal/mol, $|U|_{max}$ = 2626.4 kcal/mol, $|H|_{max}$ = 2643.0 kcal/mol, $|G|_{max}$ = 2417.1 kcal/mol, $|\alpha|_{max}$ = 196.6 a_0^3, $|C_v|_{max}$ = 47.0 cal/mol/K, $|\mu|_{max}$ = 30.0 D, $|\langle R^2\rangle|_{max}$ = 3375 a_0^2, $|ZPVE|_{max}$ = 171.9 kcal/mol, $|\varepsilon_{HOMO}|_{max}$ = 10.78 eV, $|\varepsilon_{LUMO}|_{max}$ = 4.68 eV, $|\Delta\varepsilon|_{max}$ = 12.57 eV, and $|\omega_1|_{max}$ = 3876.7 cm^{-1}.

4. Results and Discussion

Using single kernels of varying size, the systematic decay of ML prediction errors is summarized in Fig. 2 for all the aforementioned properties. To compare the error across different properties, irrespective of units and scale, we have used the mean absolute error (MAE) relative (*i.e.* RMAE) to desired quantum chemistry accuracy norms as a suitable error measure. Note that all reported error measures refer to out-of-sample predictions, *i.e.* for a given training set of size *N*, errors are presented as measured on the remaining 112 k - *N* molecules. The target accuracy for the thermochemical quantities, and orbital energies is the highly coveted 'chemical accuracy' for energetics, *i.e.* 1 kcal/mol. For ω_1, and ZPVE, both within the harmonic approximation, we have selected a target accuracy of 10 cm^{-1}. This value is slightly larger than the average accuracy of coupled cluster method, CCSD(T) + with converged basis sets,[23] for predicting harmonic wavenumbers of small molecules, as measured by comparison to their experimentally determined counterparts. For dipole moment and isotropic polarizability, the target accuracies employed are 0.1 D, and 0.1 a_0^3 respectively. These thresholds are within the uncertainty of predicted values of the same properties at the CCSD level of theory.[24]

The most compelling feature in Fig. 2 is the systematic decay in RMAEs for *all* molecular properties. This amounts to numerical evidence that predictive ML-models for multiple properties can be built using a single kernel matrix with no property-specific parametrization.

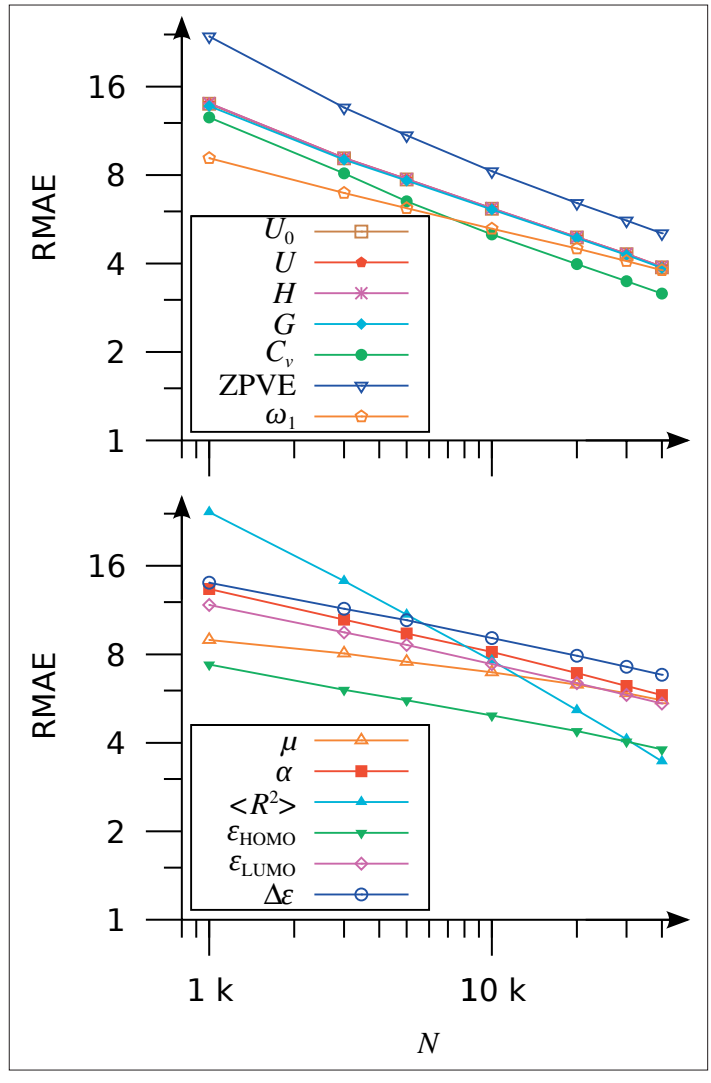

Fig. 2. Relative mean absolute errors (RMAE) in all ML predicted properties of out-of-sample molecules, in the 112 k set using the same kernel function of size up to N = 40 k. See text for the definition of RMAE for respective properties. Top: Thermochemistry and vibrational properties. Bottom: Electronic properties.

Furthermore, the accuracy of these single kernel models can systematically be improved through the addition of more training data. Among all properties, we note maximal learning rates for $\langle R^2 \rangle$ – a measure of diffuseness of the electron density, and quite possibly more directly linked to molecular geometry and composition than the other observables. ZPVE exhibits the second best learning rate, implying a potential scope to invest on building larger ML models for the data-driven, rapid and accurate estimations of ZPVE corrections for a multitude of energetics such as barrier heights, reaction/dissociation energies, and thermochemistry. At the limit of 40 k training molecules, none of the 13 properties were predicted within an RMAE of 1, indicating the need to employ larger training set sizes. μ and ω_1 are the most difficult properties to learn with the smallest learning rates. We note that despite their strongly differing distribution (shown in Fig. 1) HOMO and LUMO values have the same learning rate and only differ in their offset. For all properties, the near-linear learning curves suggest that target values of RMAE = 1 could be reached if only sufficiently large kernels were constructed. Due to the logarithmic scaling, however, the necessary computational investment for training set generation will grow increasingly prohibitive. We have also found such behavior to be independent of descriptor choice: When repeating the ML calculations using the new descriptor BOB,[20] instead of the Coulomb matrix, we noted the same trends in error decays with slightly better overall performance.

In previous studies we have already analyzed the effect on regression coefficients of a specific property due to tuning an external parameter,[16] here we can exploit the single kernel *Ansatz* to directly compare regression coefficients of different properties for the same training molecules. To demonstrate this point we have considered the 40 k coefficients used to predict H, α, and $\langle R^2 \rangle$. Fig. 3 illustrates the pairwise relationships between properties H vs. α, and H vs. $\langle R^2 \rangle$, as well as between the corresponding regression coefficients, c_t, [Eqn. (1)]. On the one hand, α exhibits the familiar lower linear bound in H, as also mentioned in ref. [10], and reminiscent of the minimal polarizability principle,[25] or the related maximum hardness principle.[26] By contrast, the c_t^H vs. c_t^α scatterplot exhibits a complete absence of correlation merely implying a disk shaped bivariate normal distribution which sets an upper bound of $(c_t^H)^2 + (c_t^\alpha)^2$, for any training molecule, t. On the other hand, and in contrast to α, the electron spread, $\langle R^2 \rangle$, correlates poorly with H (see bottom panel in Fig. 3). The scatterplot for the corresponding c_t however, shows a characteristic cross shape, suggesting a mutually complementary mode of action among training molecules. More specifically, training molecules very relevant for modeling one property (*i.e.* large c_t) are insignificant for modeling the other property, and vice versa. Such a pattern could possibly arise from a bivariate normal distribution bound of c_t^H and $c_t^{\langle R^2 \rangle}$ with L_P -norm for $0 < P < 1$. We note in passing that the properties and coefficients displayed in Fig. 3 are dimensionless because of the use of relative unit scales. Hence our results are comparable to those from other methods as long as standards units are employed, and systematic errors are accounted for.

Overall, however, we note that the distribution of c_t^p is governed by the limits imposed through σ_{opt} . For the above mentioned trivial case of ultra-tight kernels, *i.e.* in the limit that $\sigma \to 0$, we have $\mathbf{K} = \mathbf{K}^{-1} = \mathbf{I}$, and hence $\mathbf{c}^p = \mathbf{p}^r$. This implies that for the diffuse kernel functions used in this study through choice of $\sigma_{opt} = D_{ij}^{max} / \log(2)$, one should not expect $\mathbf{c}$ to reflect the same trends as properties. Finally we note the coefficients of all properties to show regularized distributions that are peaked at zero (*i.e.* no over-fitting of outliers), giving further justification to our choice of $\lambda = 0$.

5. Conclusion

We have validated and exploited the fact that ML models based on property-invariant kernels in chemical space provide a consistent framework to learn any arbitrary set of global molecular properties on the exact same footing. Using quantum chemistry data for over 100 k organic molecules, we have presented evidence how this facilitates the simultaneous modeling of several properties. Numerical results have been discussed for the single kernel based ML model of a wide variety of electronic and energetic molecular properties, including vibrational wavenumbers. Overall, we have generated 182 kernels of varying sizes for this study. To enable the reuse of the kernels for new properties by the community, we have made them publicly accessible (see Appendix). Due to the fact that the computationally most demanding step in the development of ML models is matrix inversion, requiring up to 2 CPU days when using 40 k training molecules, we include the inverse of the kernel matrices along with the data. This should enable the accelerated development of models for new properties that can be applied to any molecules which fall within the D_{ij}^{max} of the employed kernel.

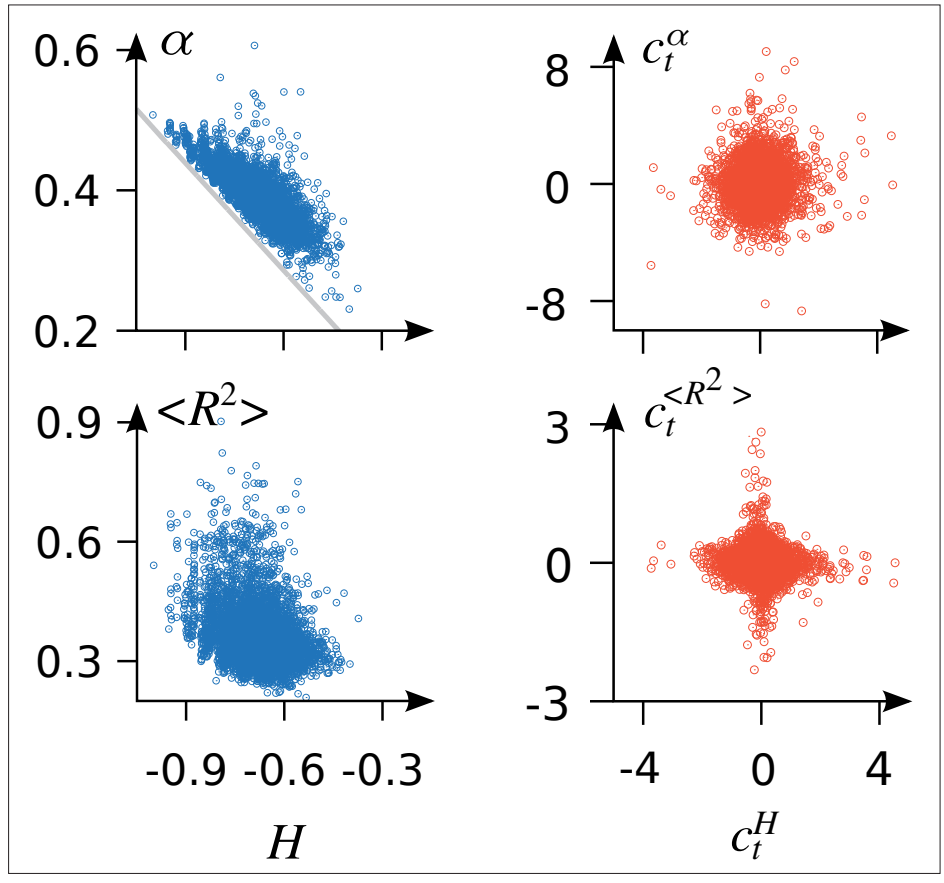

Fig. 3. Scatter plots for properties and corresponding coefficients (using the 40 k kernel). Left: Atomization enthalpy, H, versus isotropic polarizability, α (TOP), and versus radial expectation value, $\langle R^2 \rangle$ (bottom). The gray line ($y = -0.51\,x - 0.02$) indicates the linear fit to the lower bound α for any given H. Right: Scatterplots of the regression coefficients, c_t^H vs. c_t^α (top) and c_t^H vs. $c_t^{\langle R^2 \rangle}$ (bottom). All values are in units of the maximal absolute property as defined in Fig. 1.

6. Appendix: Usage of Kernels

We provide full access to all data generated in this study.[27] Fig. 4 illustrates the organization and expected usage of the data. Using inverse kernel matrices provided in the dataset, **c** vectors (see Eqn. (1)) of new properties can be computed through a matrix-vector operation (see Eqn. (3)). For this purpose, one can use properties reported in ref. [17], or compute them fresh. It is possible to train a model for a property computed at geometries from a theory slightly different than the one employed here. In such cases, **c** will account for both changes in theory, and in geometries (see discussions in ref. [18]). This **c** vector can then be used for the estimation of properties of query molecules with nine of the C, N, O, and F atoms. We caution the user that one should not expect predictive power for molecules that differ substantially from training set molecules in composition or geometry.

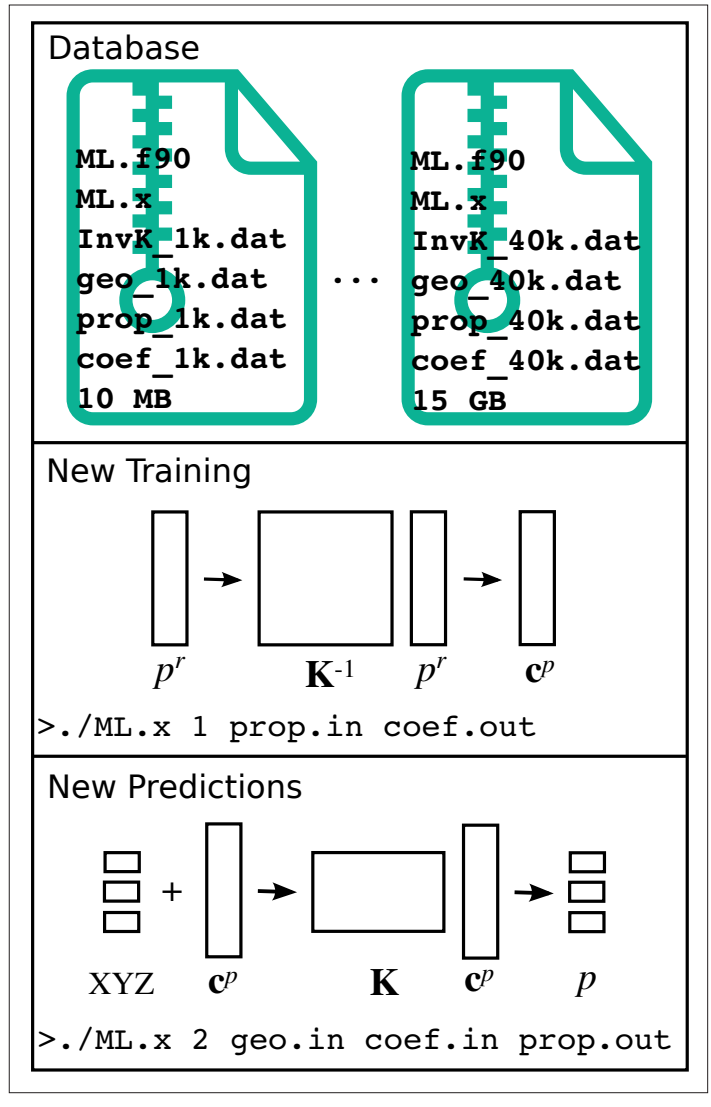

Fig. 4. Schematic description of the database access and usage. Top: For various training set sizes between 1 k, and 40 k, inverse of kernel matrices, descriptors, an example program, and corresponding indices of the training molecules in the 134 k dataset[17] are archived (see Appendix for details). Middle: With the input argument 1, and input property values, the program calculates the corresponding **c** vector. Bottom: With the input argument 2, **c** vector computed *a priori*, and geometries of query molecules, the program estimates the corresponding properties.

Acknowledgements

OAvL acknowledges funding from the Swiss National Science foundation (No. PP00P2_138932). Some calculations were performed at sciCORE (http://scicore.unibas.ch/) scientific computing core facility at University of Basel. This research used resources of the Argonne Leadership Computing Facility at Argonne National Laboratory, which is supported by the Office of Science of the U.S. DOE under contract DE-AC02-06CH11357.

Received: February 16, 2015

[1] E. K. Wilson, *Chem. Eng. News* **2000**, *78*, 39.
[2] V. Marx, *Nature* **2013**, *498*, 255.
[3] C. A. Mattmann, *Nature* **2013**, *493*, 473.
[4] C. Doctorow, *Nature News* **2008**, *455*, 16.
[5] A. Wright, *Comm. ACM*, **2014**, *57*, 13.
[6] M. Mohri, A. Rostamizadeh, A. Talwalkar, 'Foundations of machine learning', MIT press, **2012**.
[7] M. Rupp, A. Tkatchenko, K.-R. Müller, O. A. von Lilienfeld, *Phys. Rev. Lett.* **2012**, *108*, 058301.
[8] G. Montavon, K. Hansen, S. Fazli, M. Rupp, F. Biegler, A. Ziehe, A. Tkatchenko, O. A. von Lilienfeld, K.-R. Müller, *Adv. Neural Information Processing Systems* **2012**, *25*, 440.
[9] O. A. von Lilienfeld, *Int. J. Quantum Chem.* **2013**, *113*, 1676.
[10] G. Montavon, M. Rupp, V. Gobre, A. Vazquez-Mayagoitia, K. Hansen, A. Tkatchenko, K.-R. Müller, O. A. von Lilienfeld, *New J. Phys.* **2013**, *15*, 095003.
[11] K. Hansen, G. Montavon, F. Biegler, S. Fazli, M. Rupp, M. Scheffler, O. A. von Lilienfeld, A. Tkatchenko, K.-R. Müller, *J. Chem. Theory Comput.* **2013**, *9*, 3404.
[12] Z. D. Pozun, K. Hansen, D. Sheppard, M. Rupp, K.-R. Müller, G. Henkelman, *J. Comp. Phys.* **2012**, *136*, 174101.
[13] J. C. Snyder, M. Rupp, K. Hansen, K.-R. Müller, K. Burke, *Phys. Rev. Lett.* **2012**, *108*, 253002.
[14] K. T. Schütt, H. Glawe, F. Brockherde, A. Sanna, K. R. Müller, E. K. U. Gross, *Phys. Rev. B* **2014**, *89*, 205118.
[15] A. Lopez-Bezanilla, O. A. von Lilienfeld, *Phys. Rev. B* **2014**, *89*, 235411.
[16] L.-F. Arsenault, A. Lopez-Bezanilla, O. A. von Lilienfeld, A. J. Millis, *Phys. Rev. B* **2014**, *90*, 155136.
[17] R. Ramakrishnan, P. O. Dral, M. Rupp, O. A. von Lilienfeld, *Scientific Data* **2014**, 1.
[18] R. Ramakrishnan, P. O. Dral, M. Rupp, O. A. von Lilienfeld, submitted, **2015**, *http://arxiv.org/abs/1503.04987*.
[19] S. Y. Kung, 'Kernel Methods and Machine Learning', Cambridge University Press, **2014**.
[20] K. Hansen, F. Biegler, R. Ramakrashnan, W. Pronobos, O. A. von Lilienfeld, K.-R. Müller, A. Tkatchenko, submitted, **2015**.
[21] L. Ruddigkeit, R. Van Deursen, L. C. Blum, J.-L. Reymond, *J. Chem. Inf. Model.* **2012**, *52*, 2864.
[22] R. D. Johnson III, *http://cccbdb.nist.gov (2013)*.
[23] D. P. Tew, W. Klopper, M. Heckert, J. Gauss, *J. Phys. Chem. A* **2007**, *111*, 11242.
[24] A. L. Hickey, C. N. Rowley, *J. Phys. Chem. A* **2014**, *18*, 3678.
[25] U. Hohm, *J. Phys. Chem. A* **2000**, *104*, 8418.
[26] R. G. Pearson, *Acc. Chem. Res.* **1993**, *26*, 250.
[27] For data contact authors or visit *http://tinyurl.com/qjkwxky*