We introduce property-independent kernels for machine learning modeling 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. Corresponding molecular reference properties provided, they enable the instantaneous generation of ML models which can systematically be 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 kilo organic molecules of similar size. 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
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.

II. METHODS

Inarguably, one of the more appealing ML algorithms is kernel-ridge-regression (KRR) because of its numerical robustness 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 \( t \),

\[
p_q = \sum_{t=1}^{N} c^p_t K_{qt}. \tag{1}
\]

In Eq. \( 1 \) \( K_{qt} \) is a kernel matrix element corresponding to molecules \( q \) and \( t \). Here, 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 similarity measure between \( t \) and \( q \), \( K_{qt} = \exp \left( -D_{tq}/\sigma \right) \), where \( D_{tq} = |d_t - d_q| \) is the Manhattan (a.k.a. \( L_1 \)) norm of difference between two, typically non-scalar, descriptors of molecules \( t \) and \( q \), respectively. The global hyperparameter \( \sigma \) 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 Eq. \( 1 \), the vector \( c^p \), with optimal regression coefficient \( c^p_t \) as element for every training molecule \( t \), must be obtained through minimization of the penalized Lagrangian function,

\[
L = (p^r - \mathbf{K}c^p)^T (p^r - \mathbf{K}c^p) + \lambda c^p T \mathbf{K}c^p, \tag{2}
\]

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

\[
c^p = (\mathbf{K} + \lambda \mathbf{I})^{-1} p^r. \tag{3}
\]

The dependence of a model’s performance on hyperparameters, \( \sigma \), and \( \lambda \), can be understood as follows. In the presence of training molecules with extremely outlying properties an optimal value of \( \lambda \) becomes non-zero in order to quench excessively large elements of \( c^p \). In other words, the modeling function becomes more rigid and lessens the danger of overfitting. The meaning of the kernel width, \( \sigma \), is to control the \( c^p_t \) contribution from training molecule \( t \) when making a new prediction, see Eq. \( 1 \). Typically, optimal choices of \( \sigma \), and \( \lambda \) 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 \( N = 10^4 \), a 5-fold CV implies to repeatedly invert \( 8k \times 8k \) matrices, each requiring \( \sim 0.5 \) CPU hours on modern compute hardware.

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

\[
[c^{p_1}c^{p_2} \ldots c^{p_n}] = \mathbf{K}^{-1} [p_1^r p_2^r \ldots p_n^r] \Rightarrow \mathbf{C} = \mathbf{K}^{-1}\mathbf{P}^r, \tag{4}
\]

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 \( O(N^2) \).

III. COMPUTATIONAL DETAILS

For typical molecular datasets, with \( N > 5k \), we find that optimal \( \sigma \) converge towards a large value which ensures non-vanishing contributions from all training compounds, while \( \lambda \) converges towards zero. The choice of setting \( \lambda \) 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. Overfitting, 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 \( \sigma \) lead to
Figure 1. Density distributions of thirteen properties for the set of 112k organic molecules made up of CHONF. The abscissa corresponds to values of the properties relative to corresponding maximal absolute values in the dataset: $|U_0|_{\text{max}} = 2608.5$ kcal/mol, $|U|_{\text{max}} = 2626.4$ kcal/mol, $|H|_{\text{max}} = 2643.0$ kcal/mol, $|G|_{\text{max}} = 2417.1$ kcal/mol, $|\alpha|_{\text{max}} = 196.6$ $a_0^{-1}$, $|C_v|_{\text{max}} = 47.0$ cal/mol/K, $|\mu|_{\text{max}} = 30.0$ D, $|\langle R^2 \rangle|_{\text{max}} = 3375$ $a_0^2$, $|\text{ZPVE}|_{\text{max}} = 171.9$ kcal/mol, $|\varepsilon_{\text{HOMO}}|_{\text{max}} = 10.78$ eV, $|\varepsilon_{\text{LUMO}}|_{\text{max}} = 4.68$ eV, $|\alpha|_{\text{max}} = 12.57$ eV, and $|\omega_1|_{\text{max}} = 3876.7$ cm$^{-1}$.

loss of information regarding training descriptor distances, $D_{ij}$. For $\sigma \approx 0$, off-diagonal elements of the kernel matrix, $K_{ij} = \exp(-D_{ij}/\sigma)$ vanish, resulting in a unit-kernel-matrix, $K = I$. On the other hand, for $\sigma >> 1$, a kernel-matrix of ones is obtained which would be singular (and hence non-invertible) for $N > 1$, and anyways does not resolve $D_{ij}$. Here, we define the optimal $\sigma$ 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,

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

resulting in $\sigma_{\text{opt}} = D_{ij}^{\text{max}}/\log(2)$. As such, through use of $\sigma_{\text{opt}}$ and $\lambda = 0$, a property-independent global kernel matrix is obtained which only needs to be inverted once before it is used to generate regression coefficients for all molecular properties of interest. For randomly sampled 1k molecules, from the dataset considered in this study (vide infra), $D_{ij}^{\text{max}} = 677$ a.u., suggesting $\sigma_{\text{opt}} \approx 977$ a.u. This number is consistent with a numerical grid search for $\sigma_{\text{opt}}$ which identifies $\sigma = 1000$ a.u. to be optimal for a 1k 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 $\sigma_{\text{opt}} = 1000$ 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, $d$. 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 (134k) organic molecules subset of the GDB-17 published by Reymond et al. [21], which contains over 166 giga molecules. This 134k 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 8 “heavy” atoms (not counting hydrogens) which, trivially, would be outliers since the dataset is dominated by molecules with 9 heavy atoms. For the resulting 111594 (112k) molecules, we have considered the following 13 computed properties: Zero-Kelvin internal energy $U_0$, thermochem-
chemistry 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, $\alpha$; electronic radial expectation value, $\langle R^2 \rangle$; harmonic zero-point vibrational energy, ZPVE; energy of highest occupied molecular orbital (HOMO), $\varepsilon_{\text{HOMO}}$; energy of lowest unoccupied molecular orbital (LUMO), $\varepsilon_{\text{LUMO}}$; HOMO-LUMO gap, $\Delta \varepsilon$; and the highest fundamental vibrational wavenumber, $\omega_1$, 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 property density, polarizability, dipole moment, energy of HOMO, and radial expectation value. Compared to these properties, all thermochemical properties are less densely distributed. Properties even more sparsely distributed include LUMO energy, gap and ZVE; interestingly their densities also exhibit multimodal distributions (see Fig. 1), possibly arising from characteristic functional group moieties present in the dataset. The distribution of $\omega_1$ 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$_2$O ($A_1$, 3802, and $B_2$, 3906 cm$^{-1}$). Further details regarding the genesis of the dataset can be found in Ref. [17].

IV. 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. 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 112k - $N$ molecules. 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. The target accuracy for the thermochemical quantities, and orbital energies is the highly coveted “chemical accuracy” for energetics, i.e., 1 kcal/mol. For $\omega_1$, 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 these 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. 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 40k training molecules, none of the 13 properties were predicted within an RMAE of 1, indicating the need to employ larger training set sizes. $\mu$ and $\alpha$ are the more difficult properties to learn as evinced in Fig. 2. 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 off-set. 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.

Overall, however, we note that the distribution of $\sigma_{opt}^p$ is governed by the limits imposed through $\sigma_{opt}$. For the above mentioned trivial case of ultra-tight kernels, i.e. in the limit that $\sigma \to 0$, we have $K = K^{-1} = I$, and hence $c^p = p^\top$. This implies that for the diffuse kernel functions used in this study through choice of $\sigma_{opt} = D^{max}_ij / \log(2)$, one should not expect $c$ 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 $\lambda = 0$.

In previous studies we have already analyzed the effect on regression coefficients of a specific property due to tuning an external parameter [10], 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 40k coefficients used to predict $H$, $\alpha$, and $\langle R^2 \rangle$. Fig. 3 illustrates the pairwise relationships between properties $H$ vs. $\alpha$, and $H$ vs. $\langle R^2 \rangle$, as well as between the corresponding regression coefficients, $c_t$. [Eq. (1)]. On the one hand, $\alpha$ exhibits the familiar lower linear bound in $H$, as also mentioned in Ref. [8] and reminiscent of the minimal polarizability principle [25], or the related maximum hardness principle [26]. By contrast, the $c_t^H$ vs. $c_t^\alpha$ scatterplot exhibits 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 $\alpha$, the electron spread, $\langle R^2 \rangle$, correlates rather poorly with $H$ (see bottom panel in Fig. 3). The scatterplot for the corresponding $c_t$, however, now 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 $0 < P < 1$.

V. 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 100k 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 40k 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 $D_{ij}^{\text{max}}$ of the employed kernel.

VII. APPENDIX: USAGE OF KERNELS

We provide full access to all data generated in this study [27]. Fig. illustrates the organization and expected usage of the data. Using inverse kernel matrices provided in the dataset, c vectors (see Eq. (1)) of new properties can be computed through a matrix-vector operation (see Eq. (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 trainingset molecules in composition or geometry.

VI. 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.
[15] A. Lopez-Bezanilla and O. A. von Lilienfeld, Phys. Rev. B 89, 235411 (2014).
[16] L.-F. Arsenault, A. Lopez-Bezanilla, O. A. von Lilienfeld, and A. J. Millis, Phys. Rev. B 90, 155136 (2014).
[17] R. Ramakrishnan, P. O. Dral, M. Rupp, and O. A. von Lilienfeld, Scientific Data 1 (2014).
[18] R. Ramakrishnan, P. O. Dral, M. Rupp, and O. A. von Lilienfeld, (2015), submitted.
[19] S. Y. Kung, *Kernel Methods and Machine Learning* (Cambridge University Press, 2014).
[20] K. Hansen, F. Biegler, R. Ramakrishnan, W. Pronobis, O. A. von Lilienfeld, K.-R. Müller, and A. Tkatchenko, (2015), submitted.
[21] L. Ruddigkeit, R. Van Deursen, L. C. Blum, and J.-L. Reymond, J. Chem. Inf. Model. 52, 2864 (2012).
[22] R. D. Johnson III, http://cccbdb.nist.gov (2013).
[23] D. P. Tew, W. Klopper, M. Heckert, and J. Gauss, J. Phys. Chem. A 111, 11242 (2007).
[24] A. L. Hickey and C. N. Rowley, J. Phys. Chem. A 118, 3678 (2014).
[25] U. Hohm, J. Phys. Chem. A 104, 8418 (2000).
[26] R. G. Pearson, Acc. Chem. Res. 26, 250 (1993).
[27] For data contact authors or visit http://tinyurl.com/qjkwzky.