Supporting Information for *A Data-Driven Construction of the Periodic Table of the Elements*

Michael J. Willatt, Felix Musil and Michele Ceriotti
(Dated: June 2018)

I. OPTIMIZATION OF THE ALCHEMICAL KERNEL IN LINEAR REGRESSION

The generic regularized loss function is defined by

$$L(\mathbf{u}, \mathbf{w}, \sigma_w; \{A\}) = \frac{1}{2} \sum_{N \in A} [y(N) - \langle w|N\rangle]^2 + \frac{1}{2}\sigma_w^2 \langle w|w\rangle \,, \tag{1}$$

where $\mathbf{w} = \{\langle nn'lJJ'|w\rangle\}$ and $\mathbf{u} = \{u_{\alpha J}\}$, on which $|N\rangle$ is implicitly dependent. For k-fold cross validation, there are k optimal linear regression weights $\mathbf{w}_k$, which satisfy the vector equations

$$L_2(\mathbf{u}, \mathbf{w}_k, \sigma_w; \{A_k\}) = 0, \tag{2}$$

where the subscript denotes differentiation with respect to the second argument. Solving these equations, which are linear in $\mathbf{w}_k$, provides relations for $\mathbf{w}_k(\mathbf{u})$. Furthermore,

$$\mathbf{w}_k'(\mathbf{u}) = -L_{22}^{-1}(\mathbf{u}, \mathbf{w}_k(\mathbf{u}), \sigma_w; \{A_k\})\, L_{12}(\mathbf{u}, \mathbf{w}_k(\mathbf{u}), \sigma_w; \{A_k\}). \tag{3}$$

Having calculated these quantities, the total k-fold cross-validation error (the square of the total Root Mean Square Error),

$$L(\mathbf{u}) = \sum_k L(\mathbf{u}, \mathbf{w}_k(\mathbf{u}), \sigma_w = 0; \{A_k\}^c), \tag{4}$$

where the c superscript denotes the set complement, can be minimized (at least locally) by findings the roots of

$$L'(\mathbf{u}) = \sum_k L_1(\mathbf{u}, \mathbf{w}_k(\mathbf{u}), \sigma_w = 0; \{A_k\}^c) + L_2(\mathbf{u}, \mathbf{w}_k(\mathbf{u}), \sigma_w = 0; \{A_k\}^c)\mathbf{w}_k'(\mathbf{u}). \tag{5}$$

To find the roots, we used an implementation of the L-BFGS algorithm in the scipy library for python. As mentioned in the main text, we initialised the matrix u for this optimization by solving $uu^T = \kappa$ with κ given by an exponential kernel involving the Pauling electronegativities and van der Waals radii of each element in the data set.

For the elpasolite crystals and QM9 data sets, we used two-fold cross validation and found that 500 iterations of the L-BFGS algorithm were sufficient to converge to a local minimum of the cross-validation error (Eq. (5)). The optimized u matrices for the elpasolite crystals data set ($d_J = 1, 2, 4$) are included as python .npy files in the SI (see the README.md file).

II. COMPUTATION OF THE SOAP DESCRIPTOR

We used the QUIP package (`http://libatoms.github.io/QUIP/index.html`) with the GAP code (`http://www.libatoms.org/gap/gap_download.html`) to compute the SOAP descriptors, using the quippy python interface with the following input string:

```
'average=F normalise=T soap cutoff_dexp=\$cutoff_dexp cutoff_rate=\$cutoff_rate \
cutoff_scale=\$cutoff_scale central_reference_all_species=F \
central_weight=\$centerweight covariance_sigma0=0.0 atom_sigma=\$awidth \
cutoff=\$rc cutoff_transition_width=0.5 n_max=\$nmax l_max=\$lmax \
n_species=\$nspecies species_Z=\$species n_Z=\$ncentres Z=\$centres'
```

The parameters cutoff_dexp, cutoff_scale, cutoff_rate, central_weight, atom_sigma, cutoff, n_max, l_max are respectively referred to in this paper as m, r_0, c, u_0, g_w, r_c, n_{max} and l_{max}.

III. DETAILS OF THE GRID-SEARCH OPTIMIZATIONS

Note that the effective regularization parameter of the KRR model we use is given by $reg_{eff} = \frac{\sigma^2 \,\mathrm{Tr}\{K\}}{var(y)N}$, where σ is the reported regularization parameter, K is the kernel matrix for the training set, $var(y)$ is the variance of the training properties and N is the number of training samples.

σ	10^{-7}	10^{-6}	10^{-5}	10^{-4}	10^{-3}	0.01	0.1		
m	0	1	2	4	5	6	7	8	9
r_0	1	2	3	4					

TABLE I. Parameters used for the grid search of the optimal radial scaling on QM9. All the possible combinations of the three parameters were evaluated.

A. Radial Scaling on QM9

The optimization of the parameters of the radially-scaled kernel involved r_0 and m (Eqs. 21 from the main text) and the regularization parameter (see table I for the choice of values). The other parameters of the kernel were fixed to $r_c = 5$, $c = 1$, $u_0 = 1$, $\zeta = 2$, $\nu = 2$, $n_{max} = 12$, $l_{max} = 9$ and Gaussian width $g_w = 0.3$. The performance of each set of parameters was scored with the Mean Absolute Error (MAE) averaged over 10-fold cross-validation and the results are available in the SI (see README.md file).

B. Chemical Correlations on QM9

We optimized the parameters σ_ϵ and σ_r (see Eq. 20 from the main text) by evaluating the MAE averaged over 10-fold cross validation for each combination on the range from 0.1 to 1 for σ_ϵ and 1 to 2 for σ_r with an increment of 0.1 and from 10^{-7} to 10^{-4} for σ with an increment of factors of 10. The other kernel parameters were fixed as follow: $r_c = 5$, $\zeta = 2$, $\nu = 2$, $n_{max} = 12$, $l_{max} = 9$, $g_w = 0.3$, $u_0 = 2$, $c = 1$, $r_0 = 2$, $m = 7$. The results of this procedure are available in the SI (see README.md file).

Note that the Pauling's atomic electronegativities and the van der Waals radii were standardized (mean is removed and scaled with the standard deviation) to simplify the determination of the search range for σ_ϵ and σ_r.

IV. RMSE LEARNING CURVES

Figures 1-4 show some additional diagnostics on the machine-learning models discussed in the main text.

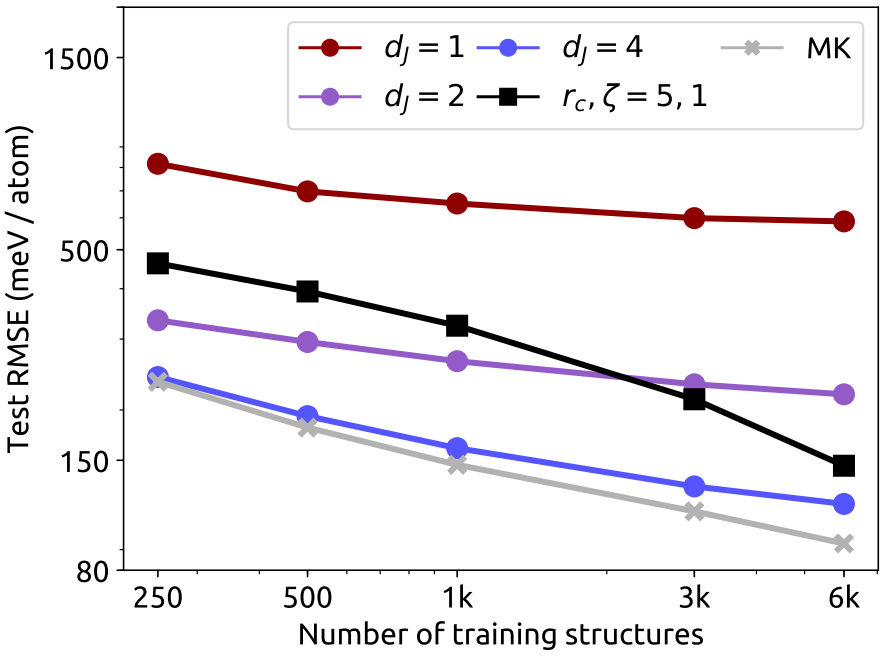

FIG. 1. Learning curves for the elpasolite crystals. The standard SOAP curve is shown in black and the optimized curves are shown in dark red ($d_J = 1$), purple ($d_J = 2$) and blue ($d_J = 4$). For each of these models, the kernels were constructed with $r_c = 5$Å and $\zeta = 1$. The multiple-kernel model (shown in grey) combines three standard SOAP kernels ($\zeta = 1$, $r_c = 4$; $\zeta = 1$, $r_c = 6$; $\zeta = 4$, $r_c = 6$) and one optimized kernel ($d_J = 4$, $\zeta = 1$, $r_c = 5$) in the ratio $4 : 3 : 1 : 220$. All of the kernels were constructed with $\nu = 2$, $n_{\max} = 12$ radial basis functions and $l_{\max} = 9$ non-degenerate spherical harmonics. Error bars are omitted because they are as small as the data point markers.

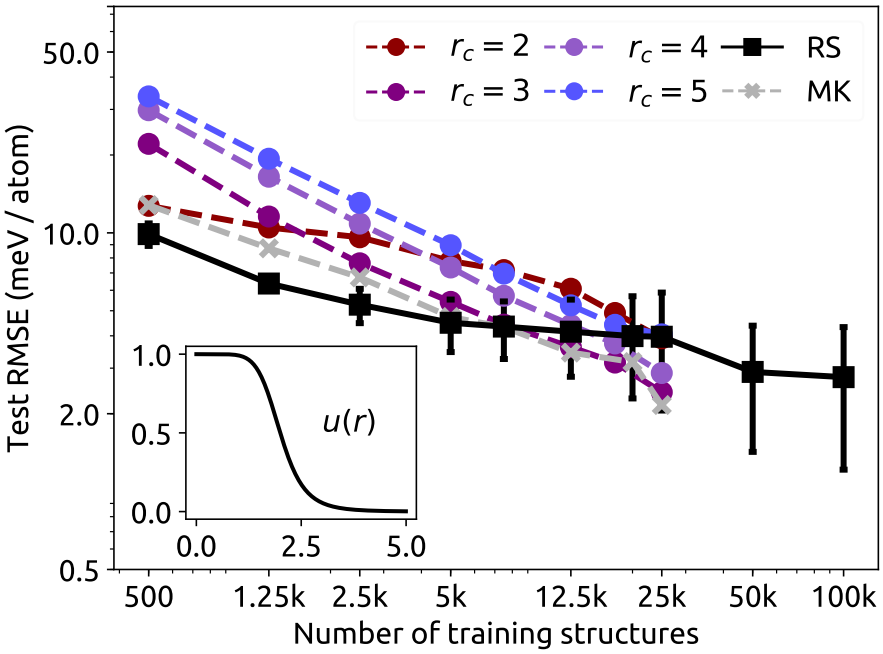

FIG. 2. Learning curves for the QM9 data set. Four of the lines show the MAE on the test set for various standard SOAP kernels ($\zeta = 2$) with different cutoff radii (dashed lines graduating from red to blue). The other lines show the MAE on the test set for the optimal radially-scaled (RS) and multiple-kernel (MK) SOAP models (black and grey lines respectively). In every model, the kernels were constructed with $\nu = 2$, $n_{\max} = 12$ radial basis functions and $l_{\max} = 9$ non-degenerate spherical harmonics. The inset shows the radial-scaling function $u(r)$ from $r = 0$Å to $r = 5$Å with the parameters that were found to minimize the ten-fold cross validation MAE on the optimization set through a grid search, $r_0 = 2$Å and $m = 7$. The multiple-kernel model combines the $r_c = 2, 3, 4$ and RS kernels in the ratio 100,000:1:2:10,000.

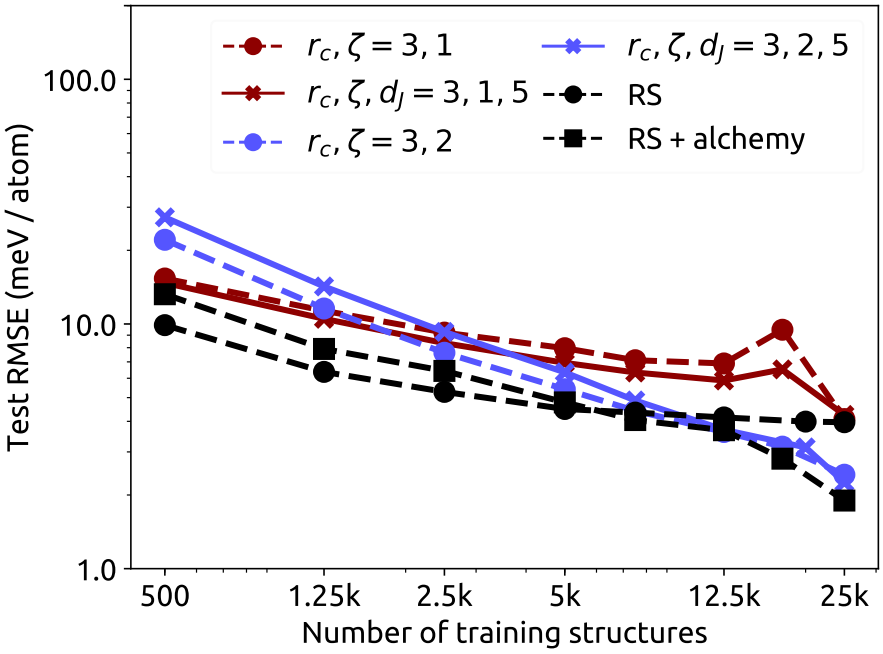

FIG. 3. Learning curves for the QM9 data set after inclusion of radially-scaled and alchemically-optimized SOAP kernels. Standard SOAP kernels with different cutoff radii are compared with the result of optimizing alchemical correlations using the scheme presented previously for the elpasolite crystal data set (blue and red lines). The learning curve of the optimized radially-scaled kernel (dashed black line with circles) is improved through inclusion of a Gaussian alchemical kernel (dashed black line with squares), which was optimized specifically for $\zeta = 2$ using a grid search. In every SOAP-based model, the kernels were constructed with $\nu = 2$, $n_{\max} = 12$ radial basis functions and $l_{\max} = 9$ non-degenerate spherical harmonics.

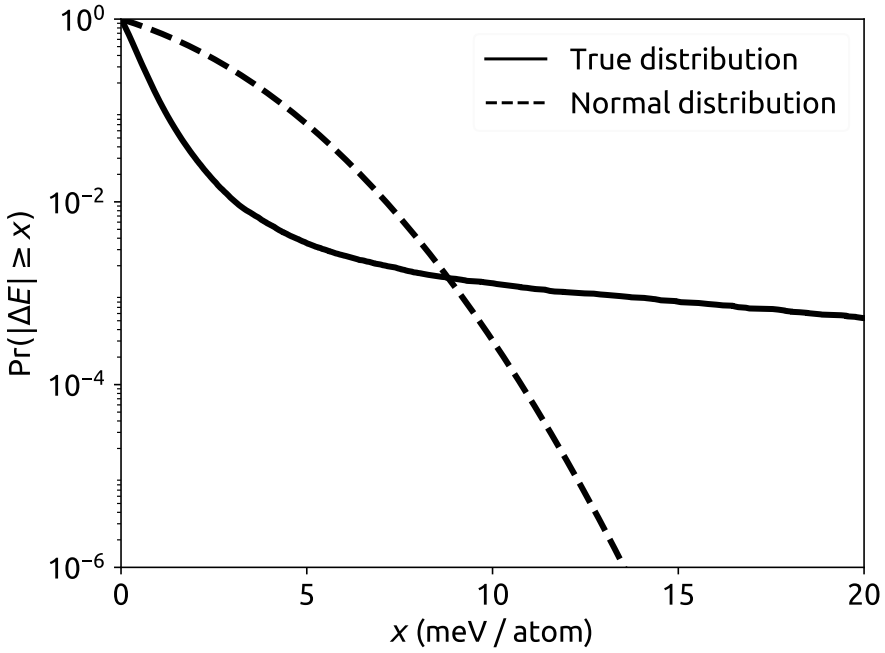

FIG. 4. Cumulative distribution function of absolute prediction errors $|\Delta E|$ for the QM9 data set, the radially-scaled kernel and 100k training structures. For reference, the dashed line shows the cumulative distribution function of a normal distribution with a standard deviation equal to the test RMSE of the model, 2.77meV/atom.