Electronic Supplementary Material (ESI) for Chemical Science.


# Multi-scale approach for the prediction of atomic scale properties
## *Supplementary Material*

Andrea Grisafi,[1,*] Jigyasa Nigam,[1,2,3,*] and Michele Ceriotti[1,3,†]

[1] *Laboratory of Computational Science and Modeling, IMX, École Polytechnique Fédérale de Lausanne, 1015 Lausanne, Switzerland*

[2] *Indian Institute of Space Science and Technology, Thiruvananthapuram 695547, India*

[3] *National Centre for Computational Design and Discovery of Novel Materials (MARVEL), École Polytechnique Fédérale de Lausanne, 1015 Lausanne, Switzerland*

* These authors contributed equally to this work

† michele.ceriotti@epfl.ch

## I. DERIVATION OF THE LONG-RANGE INTERACTION ENERGY AS A MULTIPOLE EXPANSION

We report here the compact derivation of the multipole-like expansion that enters the long-range part of the electronic energy $U^{>}$ as predicted by a linear regression model that is based on LODE(1,1) multi-scale representations. For a given atomic center $i$, this reads as follows:

$$
\begin{aligned}
U_i^{>} &= \sum_{l=0}^{l_{\max}} \sum_{a_1 a_2} \int_0^{r_c} \mathrm{d}r_1 \int_0^{r_c} \mathrm{d}r_2 \, \langle U|a_1 r_1; a_2 r_2; l\rangle \sum_{|m|\le l} \langle a_1 r_1 lm|\rho_i^{<}\rangle \, \langle a_2 r_2 lm|V_i^{>}\rangle^{\star} \\
&= \sum_{a_1 a_2} \sum_{l=0}^{l_{\max}} \int_0^{r_c} \mathrm{d}r_2 \int_0^{r_c} \mathrm{d}r_1 \, \langle U|a_1 r_1; a_2 r_2; l\rangle \sum_{|m|\le l} \langle a_1 r_1 lm|\rho_i^{<}\rangle \int_{r_c^+}^{\infty} \mathrm{d}r \, \frac{r_2^l}{r^{l+1}} \, \langle \rho_i^{>}|a_2 r lm\rangle \\
&= \sum_{a_1 a_2} \sum_{l=0}^{l_{\max}} \sum_{|m|\le l} \int_{r_c^+}^{\infty} \mathrm{d}r \, \frac{1}{r^{l+1}} \, \langle \rho_i^{>}|a_2 r lm\rangle \left( \int_0^{r_c} \mathrm{d}r_2 \, r_2^l \int_0^{r_c} \mathrm{d}r_1 \, \langle U|a_1 r_1; a_2 r_2; l\rangle \, \langle a_1 r_1 lm|\rho_i^{<}\rangle \right) \\
&= \sum_{a_1 a_2} \sum_{l=0}^{l_{\max}} \sum_{|m|\le l} \int_{r_c^+}^{\infty} \mathrm{d}r \, \frac{1}{r^{l+1}} \, \langle \rho_i^{>}|a_2 r lm\rangle \, \langle a_1 a_2; lm|M_i^{<}(U)\rangle
\end{aligned}
\tag{S.1}
$$

## II. COMPUTATIONAL DETAILS AND BENCHMARK RESULTS

For all the examples considered, both the $|\overline{\rho_i^{\otimes 2}}\rangle$ and $|\overline{\rho_i \otimes V_i}\rangle$ representations are constructed using Gaussian widths of $\sigma = 0.3$ Å and environment cutoffs of $r_c = 3.0$ Å or $r_c = 4.0$ Å, as indicated in the main text. For finite systems, both the density and potential spherical harmonics projections are computed in real space, following the implementations reported in Refs. [1] and [2] respectively. The radial and angular expansion cutoffs have been selected as $n_{\max} = 8$ and $l_{\max} = 6$. In what follows, we report the computational details for the specific examples considered.

### A. $H_2O/CO_2$ interaction

To generate the dataset, we first consider 11 non-degenerate orientations of the $CO_2$ molecule around a fixed $H_2O$ molecule centered at the origin. These orientations are selected from a Lebedev grid comprising 26 points on the unit sphere, as shown in the following figure. Then, 3 different orientations of the $CO_2$ molecule about its center of mass are considered, that are aligned along the 3 Cartesian axis. This makes a total of 33 non-degenerate reciprocal orientations between $H_2O$ and $CO_2$. For each of these, 20 rigid displacements are performed along the direction that joins the centers of the two molecules, spanning a range of distances $R$ ranging from 6.5Å to 11.5Å. For each of the 660 configurations so generated, unrelaxed interaction energies are computed at the DFT/B3LYP level using the FHI-aims software. This functional is chosen to lack of any dispersion correction, implying that the interaction at large distance is dominated by electrostatic effects. A cutoff of $r_c = 3$ Å is used to construct the LODE(1,1) representation.

Figs. 6, 7, 8, at the end of this document, list the asymptotic interaction profiles for all the 33 non-degenerate orientation considered and for all the three-kind of LODE(1,1) representations, namely, $H_2O$-centered, $CO_2$-centered and all-atoms centered.

### B. Lithium/water interaction

For this example, the calculation of the potential-field spherical harmonic projections $\langle arlm|V_i\rangle$ has been carried out extending the plane-waves implementation for periodic systems reported in Ref. [2]. Given that the box size along the $z$-axis is about 100 Å, the number of plane waves needed to reach the spatial resolution of $\sigma = 0.3$Å would be too large to achieve an efficient calculation of the potential-field. For this reason, we made use of a basic implementation of the Ewald method, by representing the potential as the sum of a slow-varying long-range contribution, that can be efficiently computed in reciprocal space, and a screened short-range contribution that can be readily computed in direct space. The Gaussian width used for the Ewald splitting has been optimized to be $\sigma_s = 1.2$ Å.

Fig. 3 shows the binding curves of all the slab-water trajectories which comprised the test set. The learning curves for the energy of interaction using both the local SOAP and multiscale LODE(1,1) representations are shown in Fig. II B.

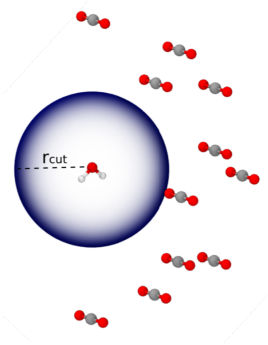


FIG. 1. 11 non-degenerate orientations of $CO_2$ around $H_2O$.

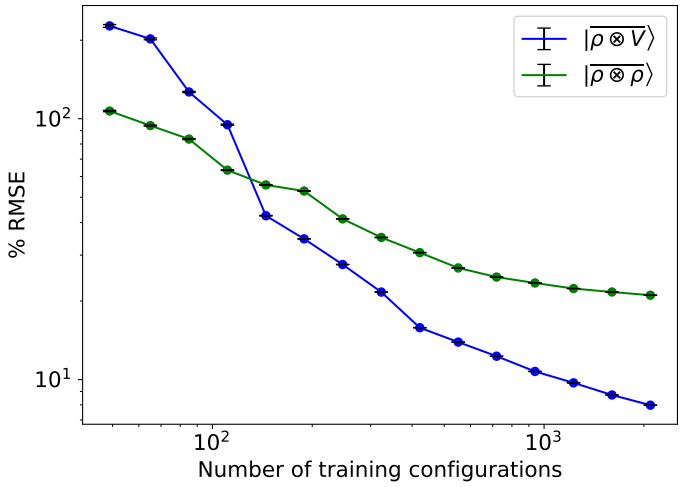


FIG. 2. Learning curves for the binding energy of water-lithium slab interaction. The energies of 75 slab-water molecule trajectories were learnt with SOAP (green) and $|\overline{\rho_i \otimes V_i}\rangle$(blue). Error in predictions on the 6 test trajectories is shown here.

### C. Polarizability of polyaminoacids

As discussed in the main text, the performance of the multiscale LODE model was compared against the quadratic SOAP kernel for learning the $\lambda$=0 polarizability tensor of polyamino acids. The non-linear SOAP kernel was constructed using a $\zeta = 2$ polynomial exponent (i.e. $k^{(\zeta=2)}(A_i, A'_{i'}) \equiv \langle A; \overline{\rho_i^{\otimes 2}} | A; \overline{\rho_{i'}^{\otimes 2}} \rangle^2$, which is equivalent to a tensor product feature space $|[\overline{\rho_i^{\otimes 2}}]^{\otimes 2}\rangle \equiv |\overline{\rho_i^{\otimes 2}}\rangle \otimes |\overline{\rho_i^{\otimes 2}}\rangle$). The underlying SOAP feature vector was sparsified with farthest point sampling, keeping only the 600 most significant features. Linear kernels (both for $|\overline{\rho_i^{\otimes 2}}\rangle$ $|\overline{\rho_i \otimes V_i}\rangle$) were built without truncating the feature vectors.

The learning curves for the trace ($\lambda = 0$) of the polarizability tensor using the linear $|\overline{\rho_i^{\otimes 2}}\rangle$ and its optimal combination with $|\overline{\rho_i \otimes V_i}\rangle$ are compared with the quadratic SOAP and the LODE(1,1) representations in Fig. 4, while the prediction error for oligopeptides of different lengths is shown in Fig. 5. It is interesting to see that while the nonlinear $|[\overline{\rho_i^{\otimes 2}}]^{\otimes 2}\rangle$ model performs more poorly than its linear counterpart $|\overline{\rho_i^{\otimes 2}}\rangle$, with worse extrapolative behavior and clear signs of overfitting, when used in combination with the multi-scale LODE it leads to significantly better performance. This indicates that the use of a flexible, non-linear short-range term does not prevent learning the most effective partitioning between local and long-range contributions, provided that the overall representation contains information on far-field correlations.

[1] A. Grisafi, D. M. Wilkins, M. J. Willatt, and M. Ceriotti, "Atomic-scale representation and statistical learning of tensorial properties," in *Machine Learning in Chemistry: Data-Driven Algorithms, Learning Systems, and Predictions*, Chap. 1, pp. 1–21.

[2] A. Grisafi and M. Ceriotti, J. Chem. Phys. **151**, 204105 (2019).

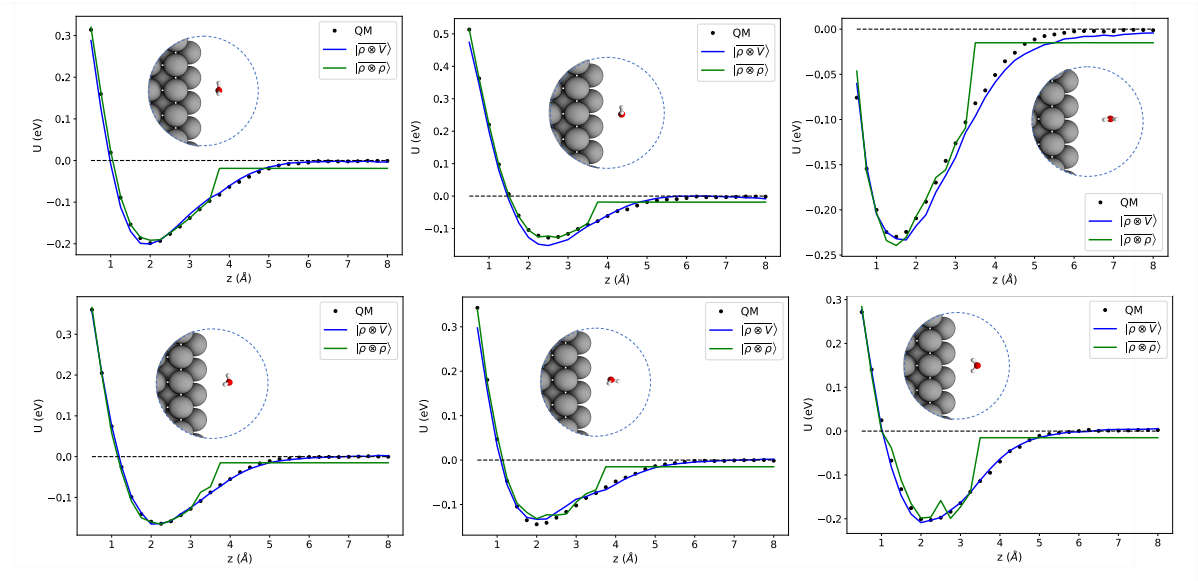


FIG. 3. Binding curves for all of the test trajectories. The inset shows the orientation of the water molecule at a distance of 4 Å from the lithium slab. While black points show the reference quantum mechanical (QM) values) of the energy of interaction, predictions using the $|\overline{\rho_i^{\otimes 2}}\rangle$ and $|\overline{\rho_i \otimes V_i}\rangle$ are shown by the green and blue curves respectively.

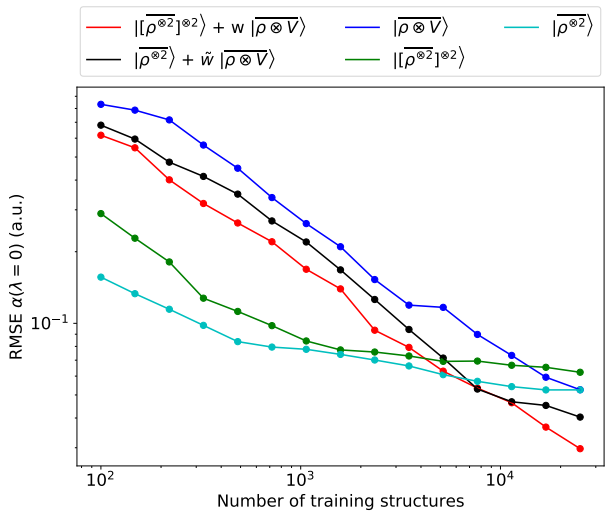


FIG. 4. Learning curves for the $\lambda = 0$ component of the polarizability tensor of a database of polypeptide conformers. The green curve corresponds to the non-linear kernel which is equivalent to $|[\overline{\rho_i^{\otimes 2}}]^{\otimes 2}\rangle$, the blue curve to a linear kernel based on $|\overline{\rho_i \otimes V_i}\rangle$, and the black one to an optimal linear combination of the two. The cyan curve shows the performance of the linear $|\overline{\rho_i^{\otimes 2}}\rangle$ kernel and the black curve is obtained using an optimal combination of the two linear kernels.

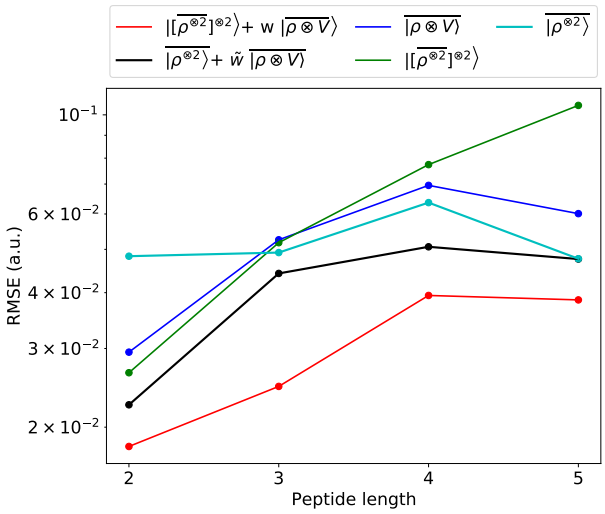


FIG. 5. Absolute RMSE in learning the l=0 spherical tensor of polarizability of polypeptides as a function of the peptide length. The model was trained on 27428 single-amino acids and 370 dipeptides. The error was computed on 30 dipeptides, 20 tripeptides, 16 tetrapeptides and 10 pentapeptides respectively.

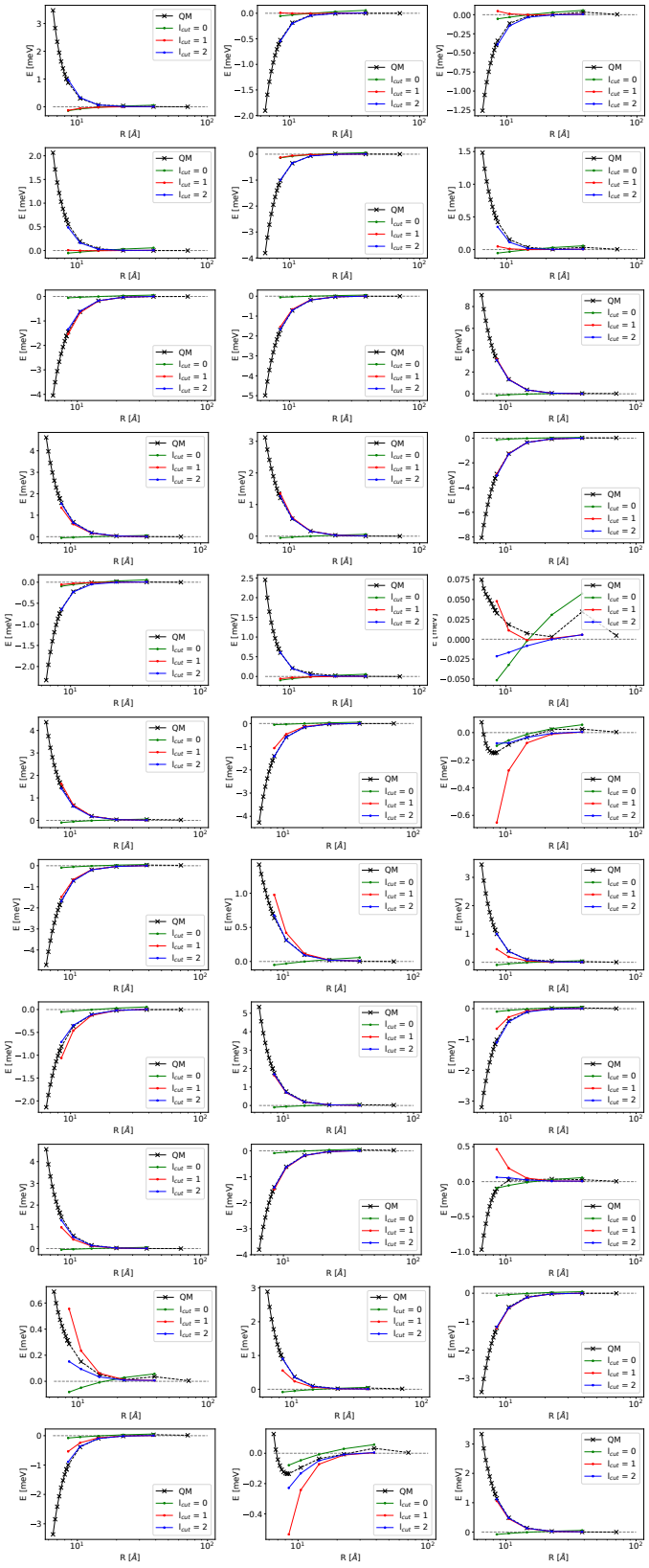


FIG. 6. $H_2O$-centered predictions for the 33 $H_2O/CO_2$ configurations.

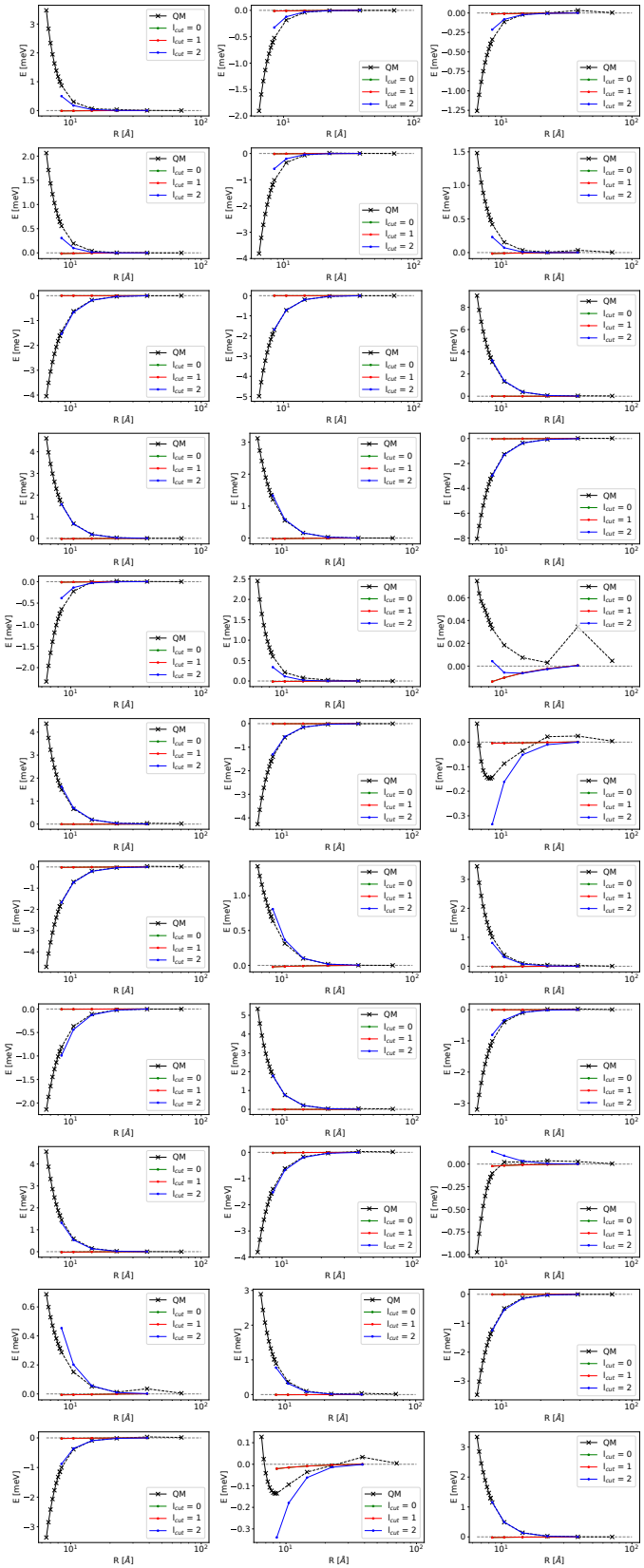


FIG. 7. $CO_2$-centered predictions for the 33 $H_2O/CO_2$ configurations.

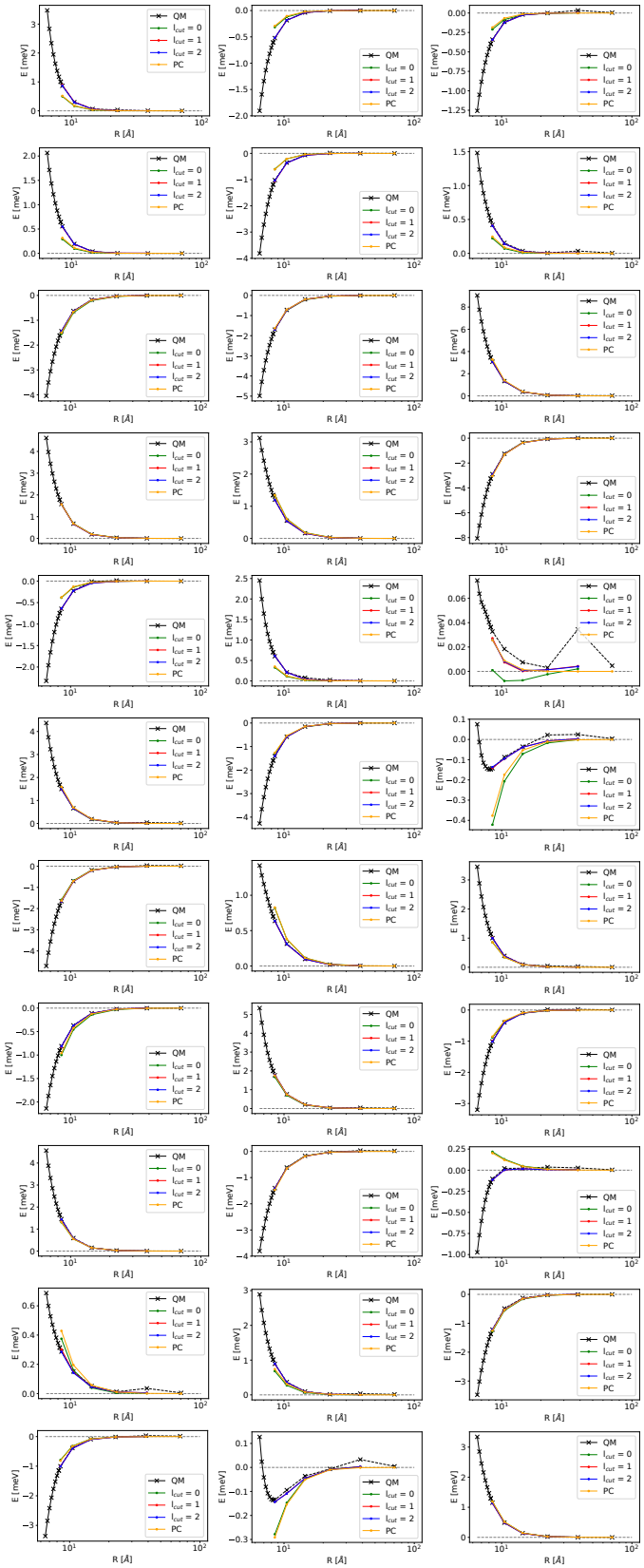


FIG. 8. All-atoms-centered predictions for the 33 $H_2O/CO_2$ configurations.