Supporting Information

Hydrogen, oxygen and lead adsorbates on $Al_{13}Co_4(100)$: accurate potential energy surfaces at low computational cost by machine learning and DFT-based data

Nathan Boulangeot,[†,‡] Florian Brix,[†,¶] Frédéric Sur,[‡] and Émilie Gaudry[*,§]

†*Univ. de Lorraine, CNRS UMR7198, Institut Jean Lamour, Campus Artem, 2 allée André Guinier, 54000 Nancy, France*

‡*Univ. de Lorraine, INRIA, CNRS UMR7503, Laboratoire lorrain de recherche en informatique et ses applications, Campus Scientifique, 615 Rue du Jardin-Botanique, 54506 Vandœuvre-lès-Nancy, France*

¶*Center for Interstellar Catalysis, Department of Physics and Astronomy, Aarhus University, DK-8000 Aarhus C, Denmark*

§*Univ. de Lorraine, CNRS UMR7198, Institut Jean Lamour, Campus Artem, 2 allée André Guinier, 54000 Nancy, France*

E-mail: Emilie.Gaudry@univ-lorraine.fr

Contents

S1 Machine learning details

S1.1 SOAP descriptor

The local atomic environment around the adsorbate located at $\mathbf{r}$ is described by the smooth overlap of atomic positions (SOAP) descriptor.[1] Within the DScribe Python package,[2] it is characterized by a vector ($\mathbf{p}$) whose components ($p(\mathbf{r})^{Z_1 Z_2}_{nn'\ell}$) are written as :[3]

$$p(\mathbf{r})^{Z_1 Z_2}_{n_0 n'_0 \ell} = \pi \sqrt{\frac{8}{2\ell+1}} \sum_{-\ell \leq m \leq \ell} [c^{Z_1}_{n_0 \ell m}(\mathbf{r})]^* \, [c^{Z_2}_{n'_0 \ell m}(\mathbf{r})] \tag{1}$$

In the previous equation, $c^{Z_1}_{n_0 \ell m}$ are coefficients based on the atomic density defined around the adsorbate within a sphere of radius R_{cut} (see supporting information). Here, we choose R_{cut} = 7 Å to include most atoms in the periodic system. The parameters n_0 and n'_0 are indices of radial basis functions (n_0 goes from 0 to $N_{max} - 1$), ℓ is the discrete angular degree of spherical harmonics (ℓ goes from 1 to L_{max}) and Z_i is the atomic number of species i. We chose $N_{max} > L_{max}$, in agreement with the literature.[4,5] More precisely, preliminary tests have concluded that $N_{max} = 9$ and $L_{max} = 3$ is a reasonable option (Fig. S1). With this set-up, each descriptor is represented by a vector of dimension 1513.

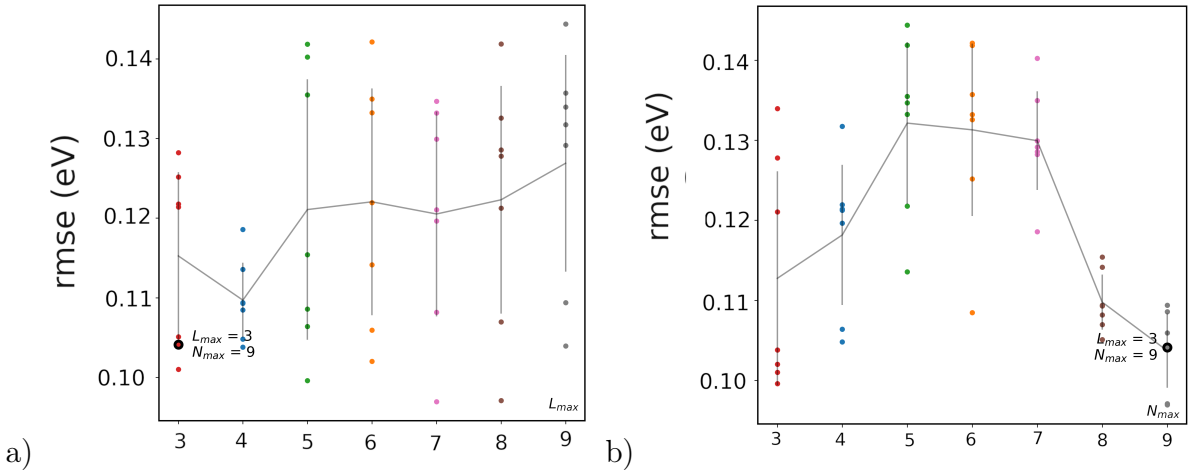

Figure S1: RMSE corresponding to predictions for H/$Al_{13}Co_4$(100) (E^{nr}_{ads}, n =49, FPS ML) for different values of the SOAP parameters L_{max} (left, a) and N_{max} (right, b). Parameters selected in this work ($L_{max} = 3$ and $N_{max} = 9$) are highlighted.

S1.2 Gaussian Kernel

In the paper, the optimization of the Gaussian kernel's hyperparameters is carried out by maximizing a likelihood function.[6] To prevent (i) numerical problems with the inversion of ill-conditioned matrices and (ii) model over-fitting, noise is taken into account and specified as a parameter called α, which adds a constant value to the diagonal of the kernel matrix.[4] We have set α_{Pb}, α_{O},$\alpha_{\mathrm{H}} = 10^{-3}$ when dealing with atomic **Lead**, **Oxygen** and **Hydrogen**, respectfully on a clean surface geometry and $\alpha_{\mathrm{Pb}} = 10^{-1}$, $\alpha_{\mathrm{O}} = 10^{-2}$ and $\alpha_{\mathrm{H}} = 10^{-3}$ when dealing with atomic **Lead**, **Oxygen** and **Hydrogen**, respectfully on an optimized surface geometry. These values result from several tests as shown in Fig. S2.

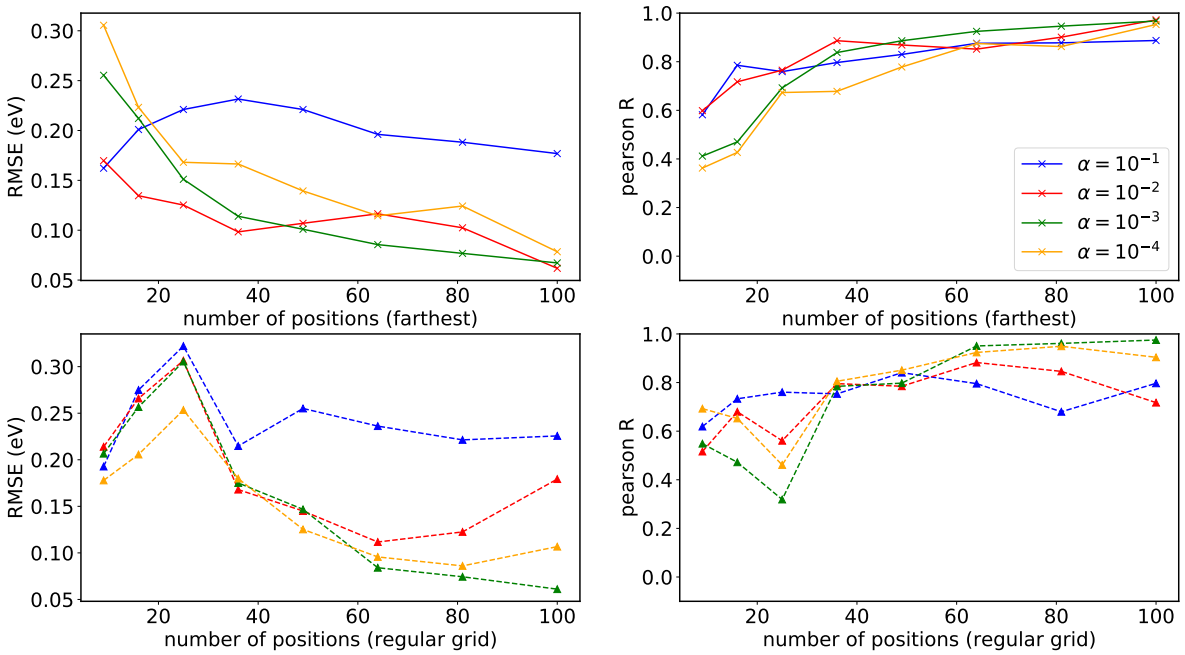

Figure S2: RMSE and Pearson-R values corresponding to predictions for H/$Al_{13}Co_4$(100) (E^{nr}_{ads}, n =49), using different α values. Both FPS ML (solid lines and crosses) and RPS ML (dashed lines and triangles) are considered.

S1.3 Metrics

The R coefficient measures the strength and direction of a linear relationship between two sets of values. It ranges from -1 to 1, where 1 (resp. -1) indicates a perfect positive (resp.

negative) linear relationship, i.e. as one variable increases, the other variable also increases (resp. decreases) proportionally. A zero value indicates no linear relationship between the variables. The average value being subtracted from $\hat{E}_{ads}$ and E^i_{ads}, R is not impacted by systematic additive errors. The $RMSE$ is a positive quantity, and a perfect prediction would correspond to $RMSE = 0$. The $ubRMSE$ is similar to $RMSE$, but it is bias-insensitive.[7] Both quantities are related through $ubRMSE^2 = RMSE^2 - b^2$, where $b = \langle \hat{E}_{ads} \rangle - \langle E_{ads} \rangle$ is the estimation bias. The unbiased metrics is used to evaluate the quality of our predictions without being impacted by systematic errors. It is expressed as a function of the standard deviations of $\hat{E}^i_{ads}$ and E^i_{ads} ($\sigma_{\hat{E}}$ and σ_E, respectively), as

$$ubRMSE = \sqrt{\sigma_E^2 + \sigma_{\hat{E}}^2 - 2R\sigma_E\sigma_{\hat{E}}} \tag{2}$$

Thus, a perfect correlation (i.e., $R = 1$) between predicted and calculated values gives $ubRMSE = |\sigma_E - \sigma_{\hat{E}}| \simeq 0$, since, in practice, $\sigma_E \simeq \sigma_{\hat{E}}$.

S1.4 MACE setup

Table S1: MACE parameters (1/2)

Parameter	**Value**	**Parameter**	**Value**
name	model	radial MLP	[64, 64, 64]
seed	1	hidden irreps	128x0e + 128x1o
log dir	logs	num channels	128
model dir	.	max L	1
checkpoints dir	checkpoints	gate	silu
results dir	results	scaling	rms forces scaling
device	cuda	avg num neighbors	1
default dtype	float32	compute avg num neighbors	True
distributed	True	compute stress	True
log level	INFO	compute forces	True
error table	PerAtomRMSE	train file	data.traj
model	ScaleShiftMACE	valid file	None
r max	6.0	valid fraction	0.2
radial type	bessel	test file	None
num radial basis	10	test dir	None
num cutoff basis	5	multi processed test	False
interaction	RealAgnosticResidual InteractionBlock	num workers	16
interaction first	RealAgnosticResidual InteractionBlock	pin memory	True
max ell	3	atomic numbers	None
correlation	3	mean	None
num interactions	2	std	None
MLP irreps	16x0e	E0s	E0s

Table S2: MACE parameters (2/2)

Parameter	**Value**	**Parameter**	**Value**
energy key	energy	lr factor	0.8
forces key	forces	scheduler patience	5
virials key	virials	lr scheduler gamma	0.9993
stress key	stress	swa	False
dipole key	dipole	start swa	None
charges key	charges	ema	True
loss	weighted	ema decay	0.995
forces weight	10.0	max num epochs	100
swa forces weight	100.0	patience	100
energy weight	1.0	eval interval	1
swa energy weight	1000.0	keep checkpoints	True
virials weight	1.0	restart latest	True
swa virials weight	10.0	save cpu	False
stress weight	100.0	clip grad	100.0
swa stress weight	10.0	wandb	False
dipole weight	1.0	wandb project	mace universal
swa dipole weight	1.0	wandb entity	astagroup
config type weights	{"Default": 1.0}	wandb name	03 faster 02
huber delta	0.01	wandb log hypers	{num channels, max L, correlation, lr, swa lr, weight decay, batch size, max num epochs, start swa, energy weight, forces weight}
optimizer	adam		
batch size	4		
valid batch size	4		
lr	0.001		
swa lr	0.001		
weight decay	1e-08		
amsgrad	True		
scheduler	ReduceLROnPlateau		

S2 Results

S2.1 Prediction of atomic Hydrogen adsorption energies

We first compare the adsorption energies calculated by DFT with values predicted by ML (Tab. S1). We focus here on the five sites with the lowest adsorption energy. Labels are those of Ref.[8]

Table S3: Adsorption energies (eV) of the five most stable sites for atomic **Hydrogen** (green squares in Fig. S3), according to ML, based on a training set built by the FPS method (ML_{FPS}) or by using the values of a regular grid (ML_{regu}), both with $n = 64$. ML energies ($\hat{E}^{nr}_{ads}$) are compared with DFT energies (E^{nr}_{ads}) at the closest site.

site	B8	B2	B16	B10	B20
ML_{FPS}	0.00	0.01	0.04	-0.06	-0.01
ML_{regu}	-0.09	-0.08	0.08	0.08	0.06
DFT_{fix}	-0.05	-0.03	0.03	0.16	0.025
DFT_{relax}	-0.20	-0.18	-0.12	-0.07	-0.12
DFT[8]	-0.16	-0.15	-0.15	-0.12	-0.09

S2.1.1 Adsorption energy maps

A summary of all adsorption energy maps (AEMs) predicted for H adsorption on $Al_13Co_4(100)$ is shown in Fig. S4. At least 100 positions are required to built AEMs that qualitatively mimic the ones calculated with 400 positions, when interpolation methods are used. In contrast, maps are quite well predicted with a low number of positions (in the range [9:36]) when machine learning methods are considered.

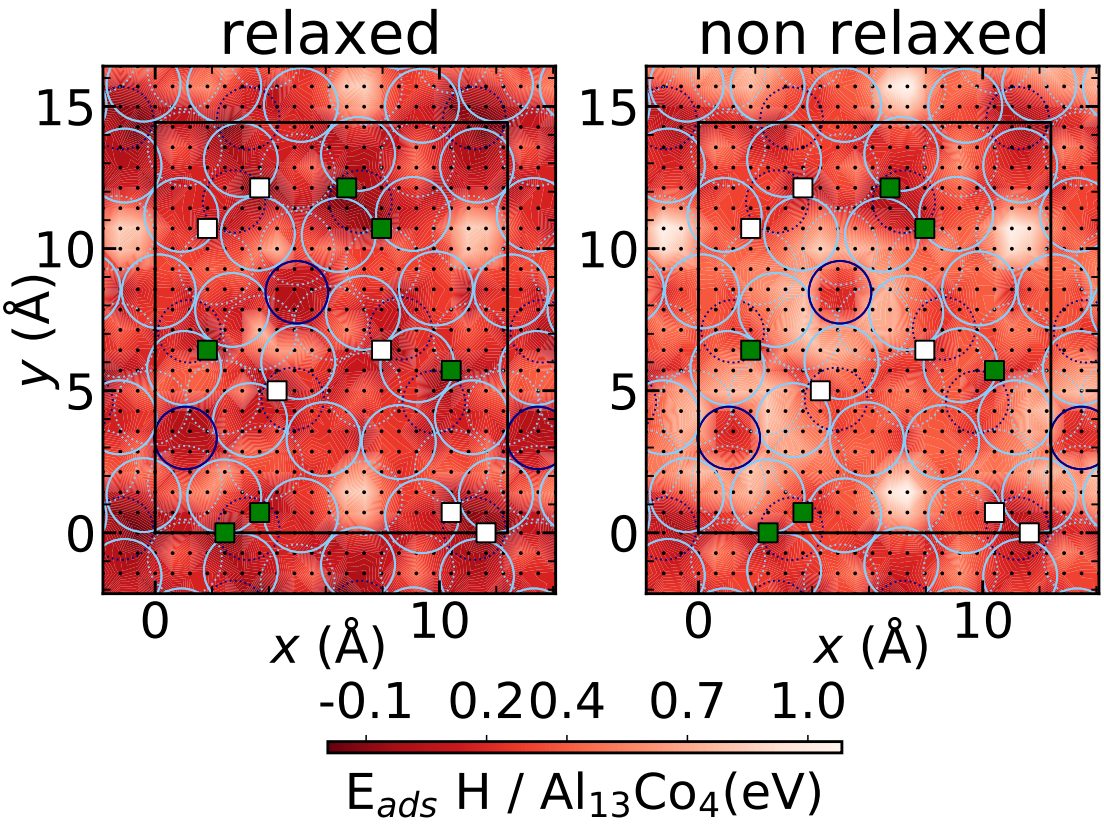

Figure S3: Adsorption energy maps, plotted for E_{ads}^{nr} (top) and E_{ads}^{r} (bottom) by interpolation between 400 DFT optimized values (regular 20×20 grid) for atomic **Hydrogen**. The atomic arrangements at the $Al_{13}Co_4(100)$ surface are superimposed. Topmost and subsurface atoms are shown in full and dotted lines, respectively. Color code : Al = light blue, Co = dark blue.

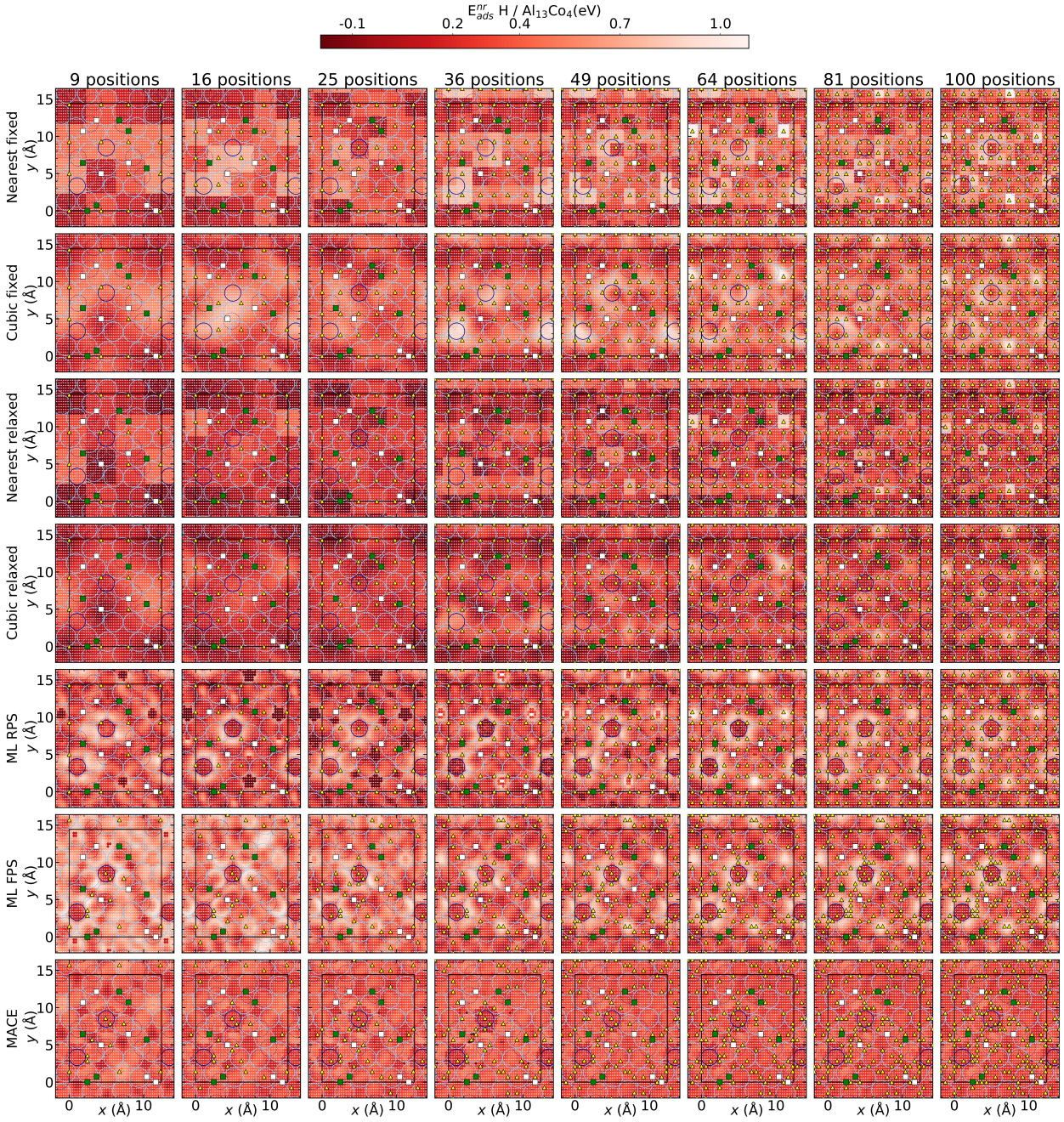

Figure S4: **Hydrogen** adsorption energy maps built by considering 9, 25, 36, 49, 64, 91 and 100 (x, y) adsorbate's positions. Machine learning (rows 5 and 6 for RPS and FPS with E^{r}_{ads}(**relaxed**) training, row 7 and 8 for RPS and FPS with E^{nr}_{ads}(**fixed**) training), as well as nearest neighbors and cubic interpolations, are used on fixed (row 1 and 2) and relaxed surface (rows 3 and 4).

S2.1.2 Error maps and histograms

Figures S5, S6, S7, S8 show maps of normalized residuals (Eq. 5 in the main paper), as well as histograms of errors and scatter plots related to the prediction of H adsorption energies on Al13Co4(100). Training is performed with E_{ads}^{nr} (Figs. S5,S6) or E_{ads}^{r} (Figs. S7,S8). Histograms for the RPS approach (in blue-green) are more symmetrically distributed. Overall, this illustrates well that there is a risk of inaccurate prediction and improper determination of adsorption sites with the RPS approach. As the training data set increases, the distribution of errors becomes narrower, and the scatter plot aligns more closely with the diagonal. Predictions on fixed surfaces give slightly overestimated adsorption energies, while predictions on relaxed surfaces give slightly underestimated adsorption energies (absolute values).

Figures S9, S10, S11 and S12 show the metrics measured when training is performed on with E_{ads}^{nr} (Figs. S5,S6) or E_{ads}^{r} (Figs. S7,S8). Reference energies are either the ones on the non relaxed surface (Figs. S5,S7) or on the relaxed surface (Figs. S6,S8). The influence of the surface relaxation (following adsorption) on the ML results is illustrated in Fig. S13.

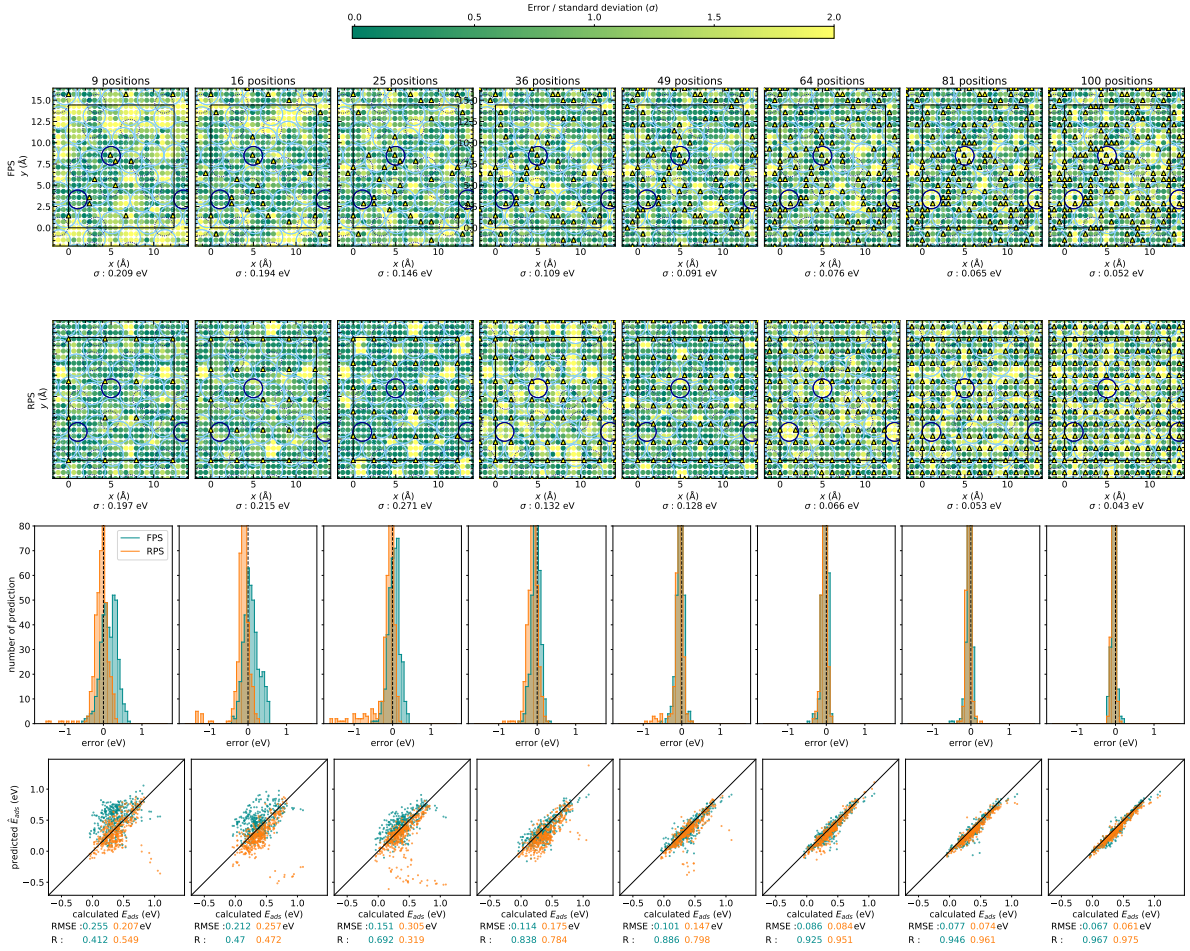

Figure S5: Error maps and histograms for atomic **Hydrogen** adsorption energies. Training is done on E_{ads}^{nr}(**fixed**). Results are compared with E_{ads}^{nr}(**fixed**). The positions are selected on regular grids and FPS method.

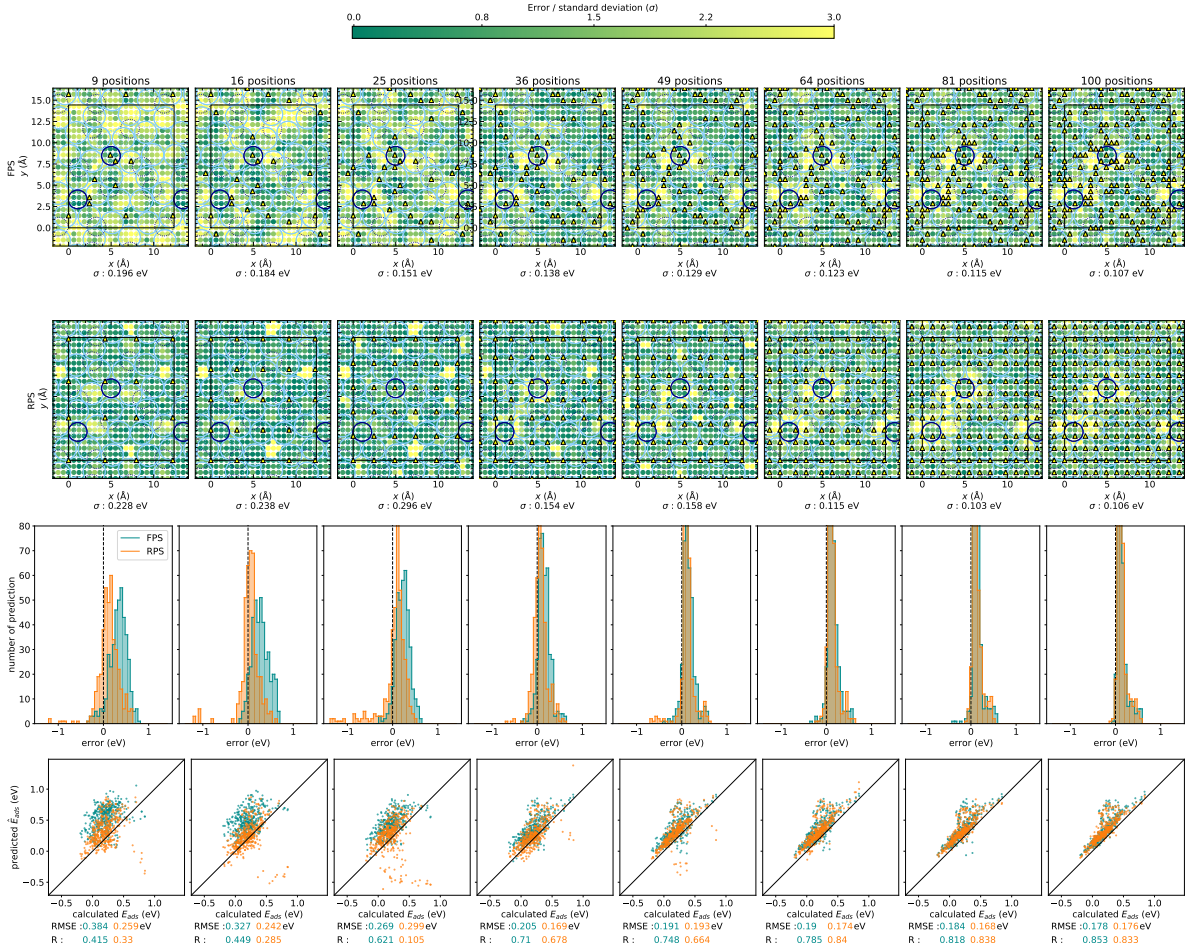

Figure S6: Error maps and histograms for atomic **Hydrogen** adsorption energies. Training is done on E^{nr}_{ads}(**fixed**). Results are compared with E^{r}_{ads}(**relaxed**). The positions are selected on regular grids and FPS method.

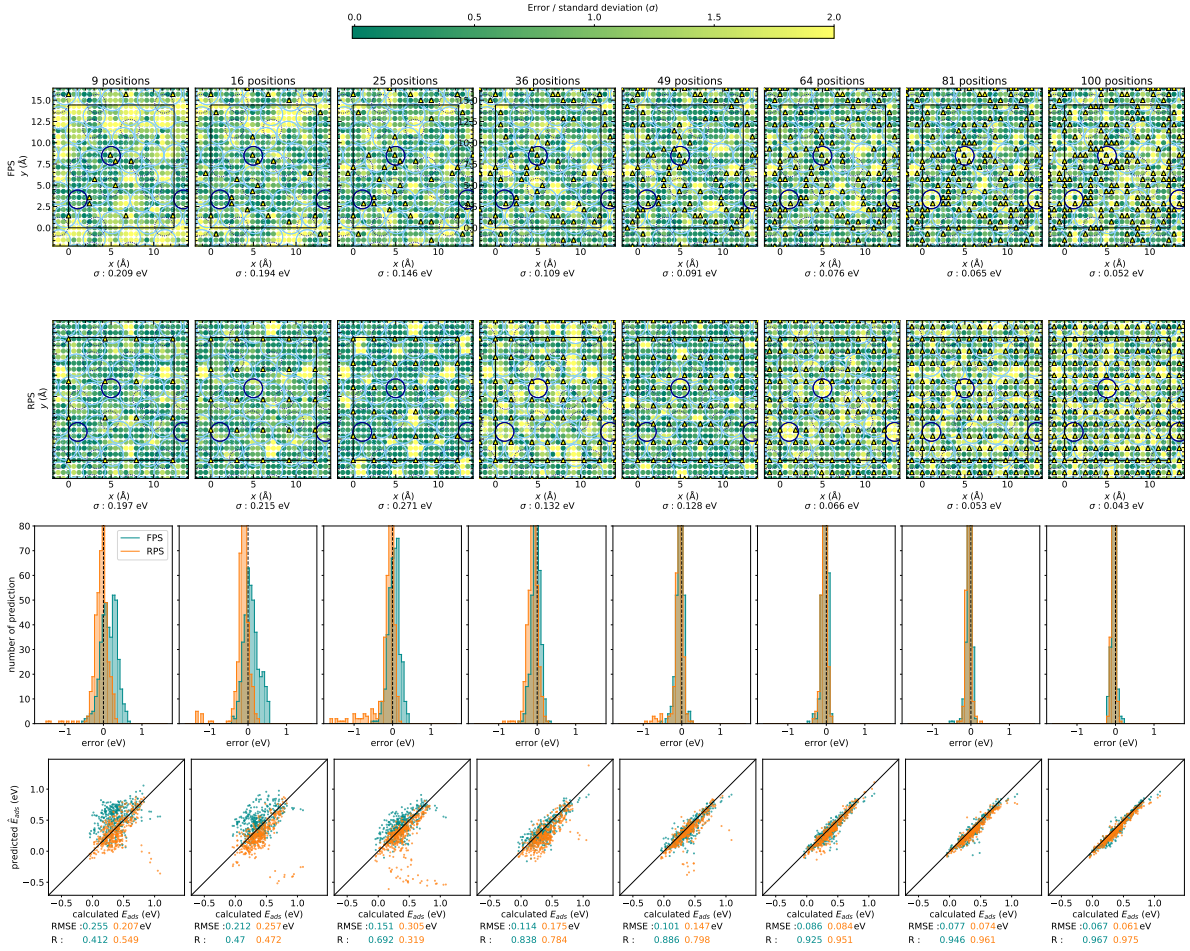

Figure S7: Error maps and histograms for atomic **Hydrogen** adsorption energies. Training is done on E^{r}_{ads}(**relaxed**). Results are compared with E^{nr}_{ads}(**fixed**). The positions are selected on regular grids and FPS method.

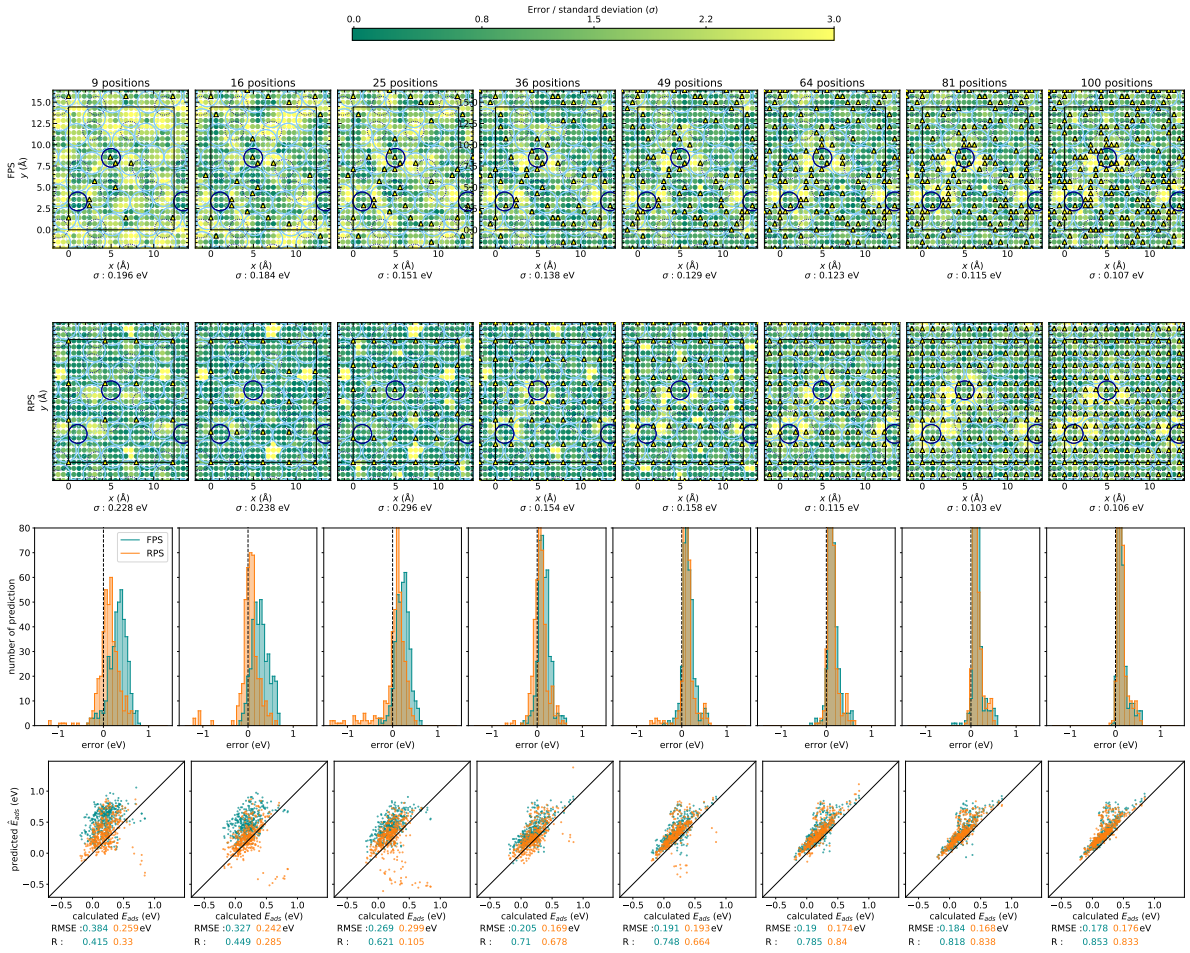

Figure S8: Error maps and histograms for atomic **Hydrogen** adsorption energies. Training is done on E^r_{ads}(**relaxed**). Results are compared with E^r_{ads}(**relaxed**). The positions are selected on regular grids and FPS method.

S2.1.3 Metrics

Metrics related to predictions are shown in Figs. S9,S10,S11,S12. Training is performed with E^{nr}_{ads} (Figs. S9,S10) or E^r_{ads} (Figs. S11,S12). Reference energies are either the ones on the non relaxed surface (Figs. S9,S11) or the relaxed surface (Figs. S10,S8). In all cases, ML approaches outperform interpolation methods.

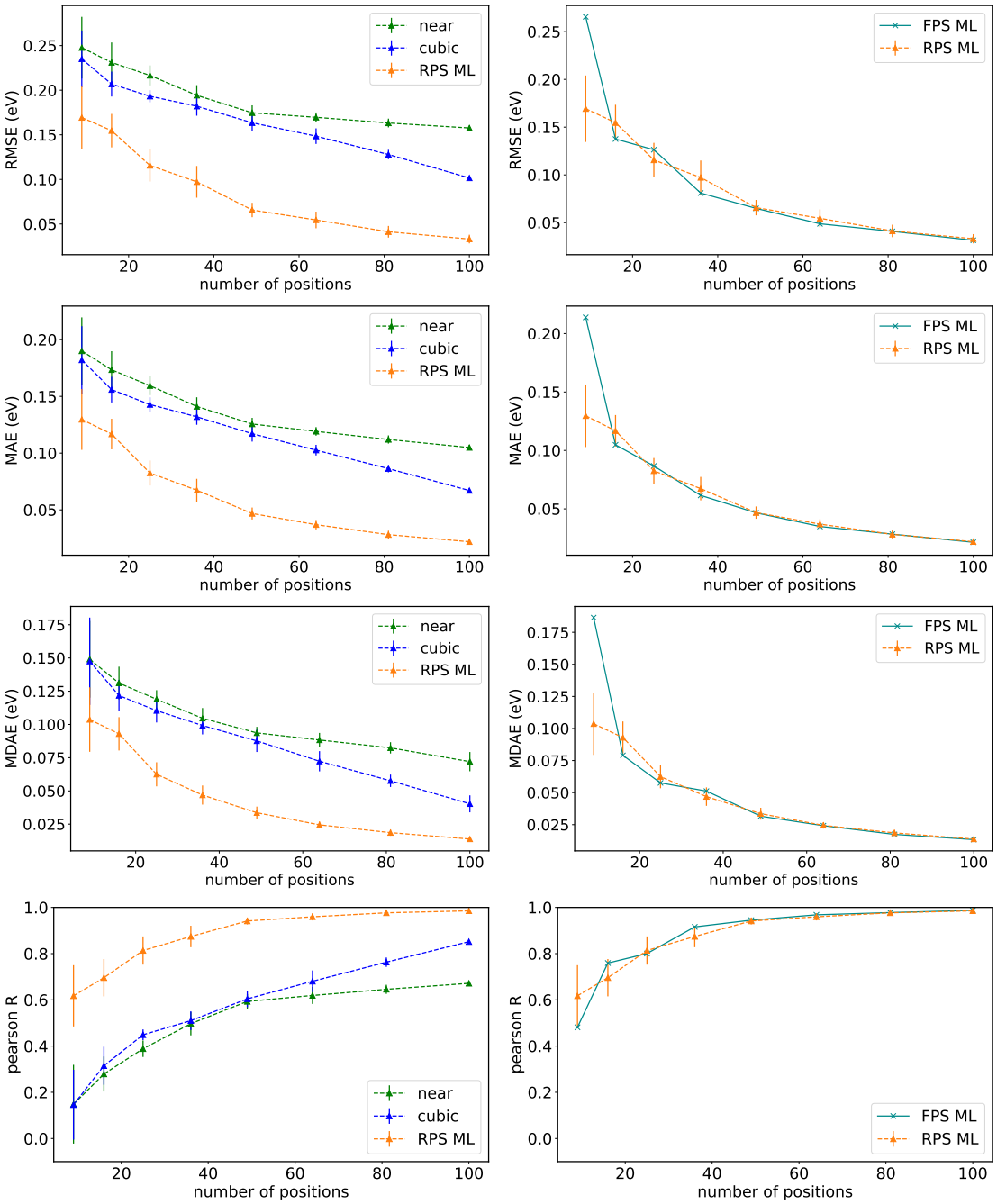

Figure S9: Metrics for **Hydrogen** adsorption. Training is done on E_{ads}^{nr}(**fixed**). Results are compared with E_{ads}^{nr}(**fixed**). RMSE (first line), MAE (second line), MAD (third line) and Pearson coefficient (fourth line) as a function of the number of DFT-calculated values in the training set. The sites considered for the training are located on a regularly spaced grid (dashed lines and triangles, labeled by *regular grid*) or are selected by our method (solid lines and crosses, labeled by *FPS*). Results of interpolation methods are represented in green (nearest neighbor interpolation) and in blue (cubic interpolation). Results of the machine learning approach are shown in orange (*regular grid*) and cyan (*FPS*).

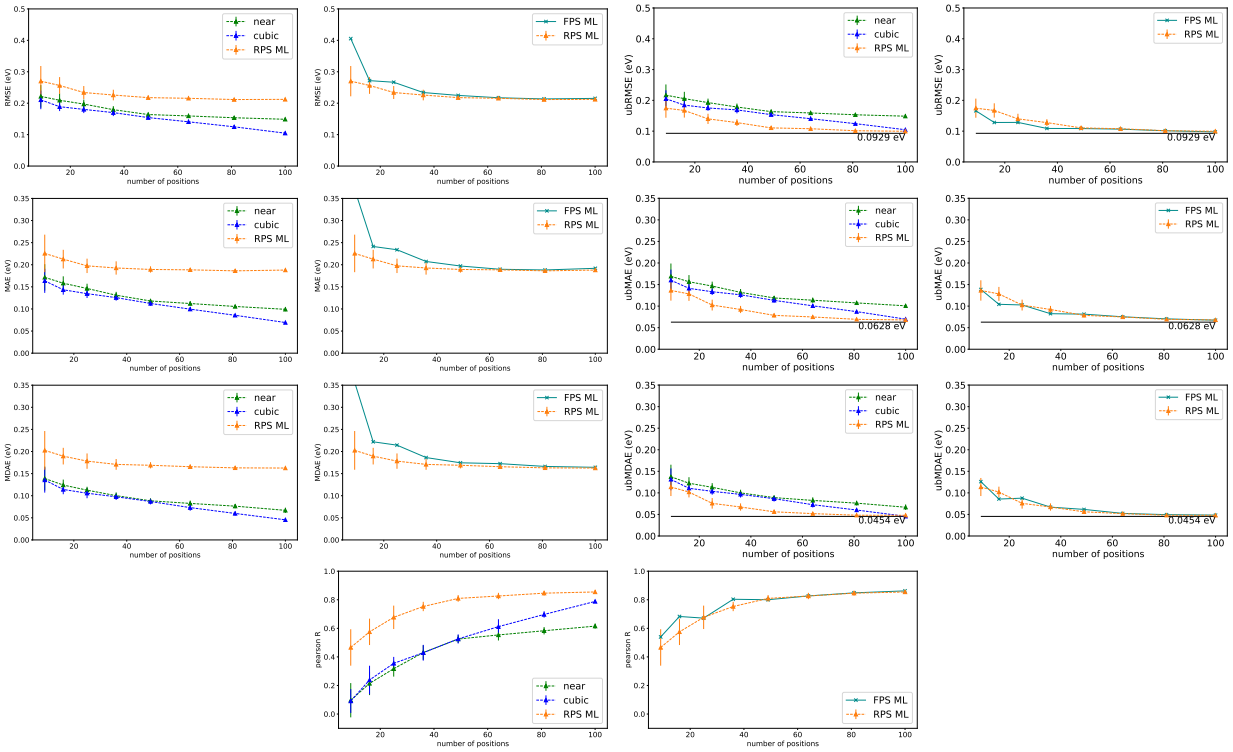

Figure S10: Metrics for **Hydrogen** adsorption. Training is done on E_{ads}^{nr}(**fixed**). Results are compared with E_{ads}^{r}(**relaxed**). RMSE (first line), MAE (second line), MAD (third line) and Pearson coefficient (fourth line) as a function of the number of DFT-calculated values in the training set. The sites considered for the training are located on a regularly spaced grid (dashed lines and triangles, labeled by *regular grid*) or are selected by our method (solid lines and crosses, labeled by *FPS*). Results of interpolation methods are represented in green (nearest neighbor interpolation) and in blue (cubic interpolation). Results of the machine learning approach are shown in orange (*regular grid*) and cyan (*FPS*). for the three first line (RMSE, MAE, MAD), results are shown with biased (first and second column) and unbiased (third and fourth column) errors. Horizontal black line represented the value given by the metric considering the difference between E_{ads}^{nr}(**fixed**) and E_{ads}^{r}(**relaxed**) as the error (see Fig S13

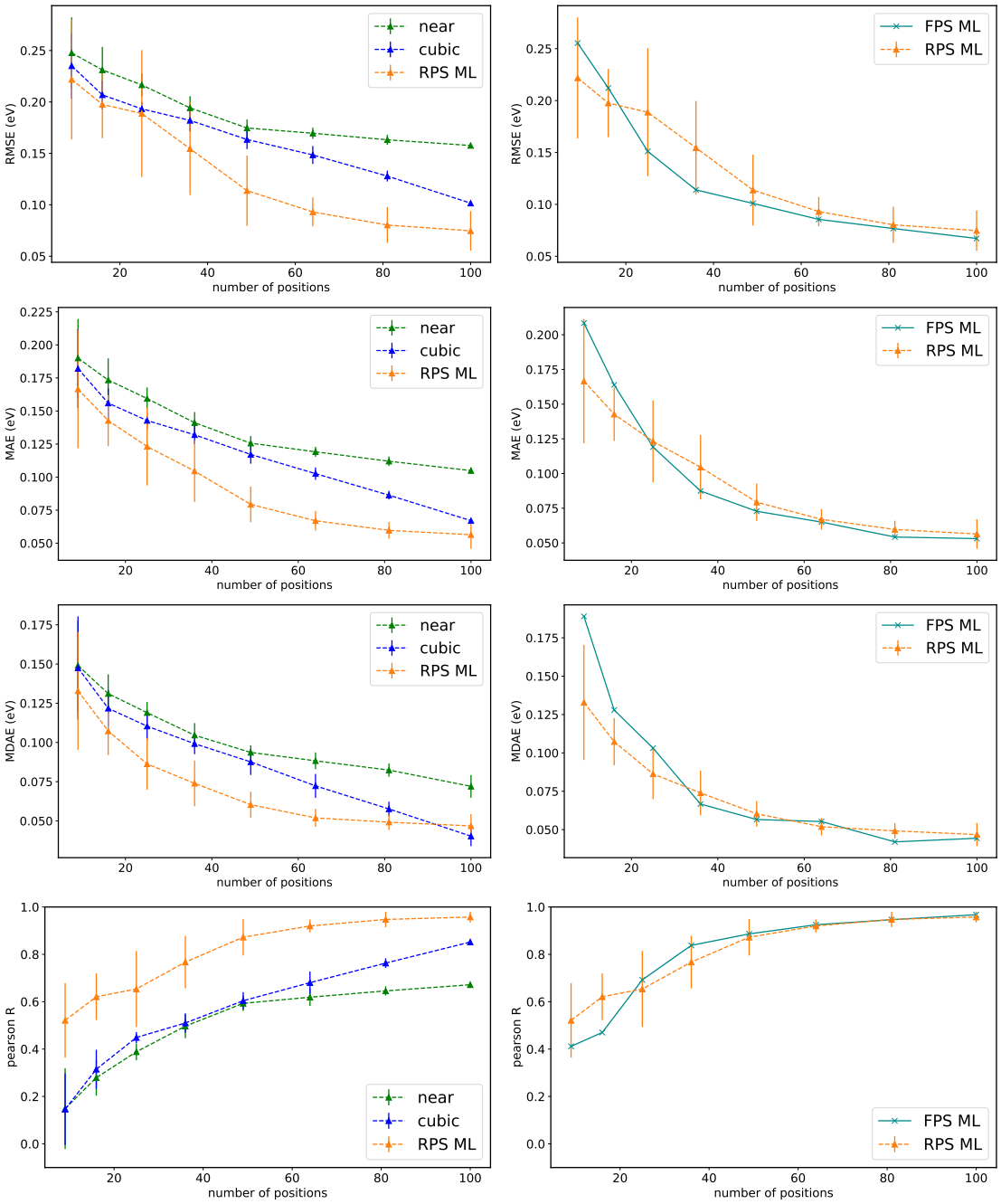

Figure S11: Metrics for **Hydrogen** adsorption. Training is done on E^r_{ads}(**relaxed**). Results are compared with E^{nr}_{ads}(**fixed**). RMSE (first line), MAE (second line), MAD (third line) and Pearson coefficient (fourth line) as a function of the number of DFT-calculated values in the training set. The sites considered for the training are located on a regularly spaced grid (dashed lines and triangles, labeled by *regular grid*) or are selected by our method (solid lines and crosses, labeled by *FPS*). Results of interpolation methods are represented in green (nearest neighbor interpolation) and in blue (cubic interpolation). Results of the machine learning approach are shown in orange (*regular grid*) and cyan (*FPS*).

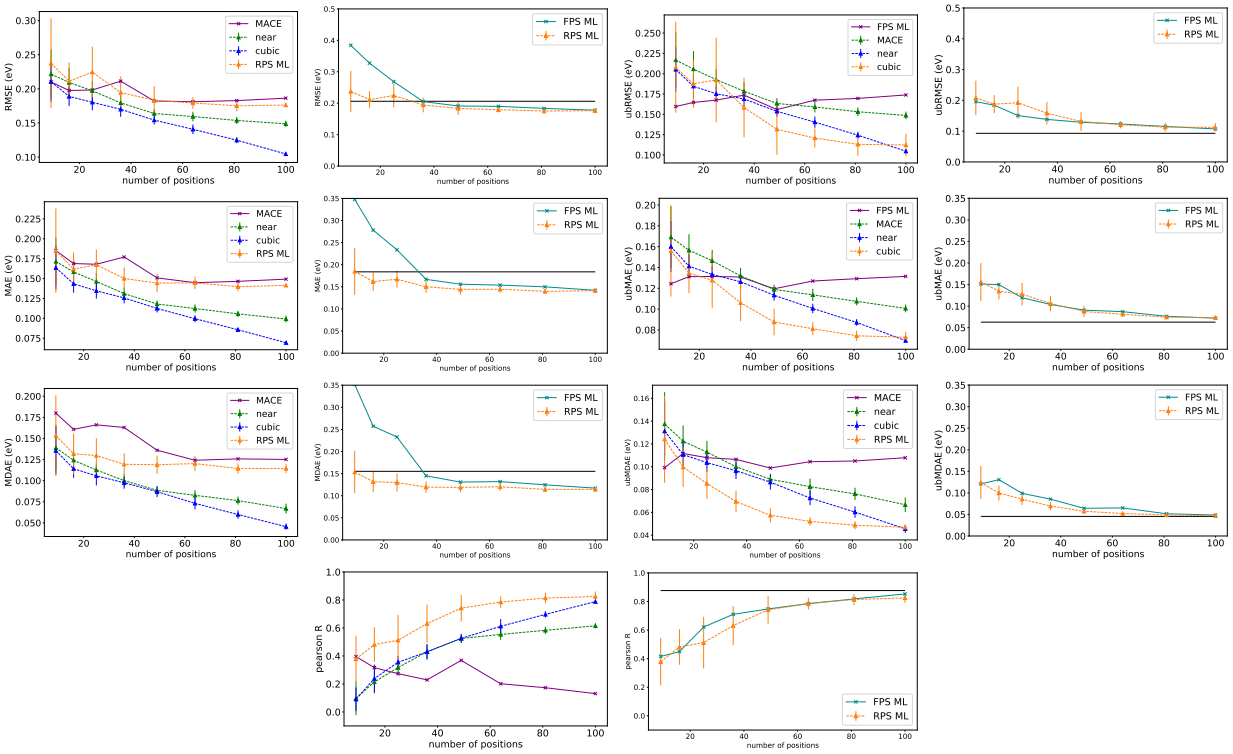

Figure S12: Metrics for **Hydrogen** adsorption. Training is done on E^r_{ads}(**relaxed**). Results are compared with E^r_{ads}(**relaxed**). RMSE (first line), MAE (second line), MAD (third line) and Pearson coefficient (fourth line) as a function of the number of DFT-calculated values in the training set. The sites considered for the training are located on a regularly spaced grid (dashed lines and triangles, labeled by *regular grid*) or are selected by our method (solid lines and crosses, labeled by *FPS*). Results of interpolation methods are represented in green (nearest neighbor interpolation) and in blue (cubic interpolation). Results of the machine learning approach are shown in orange (*regular grid*) and cyan (*FPS*). for the three first line (RMSE, MAE, MAD), results are shown with biased (first and second column) and unbiased (third and fourth column) errors. Horizontal black line represented the value given by the metric considering the difference between E^{nr}_{ads}(**fixed**) and E^r_{ads}(**relaxed**) as the error (see Fig S13

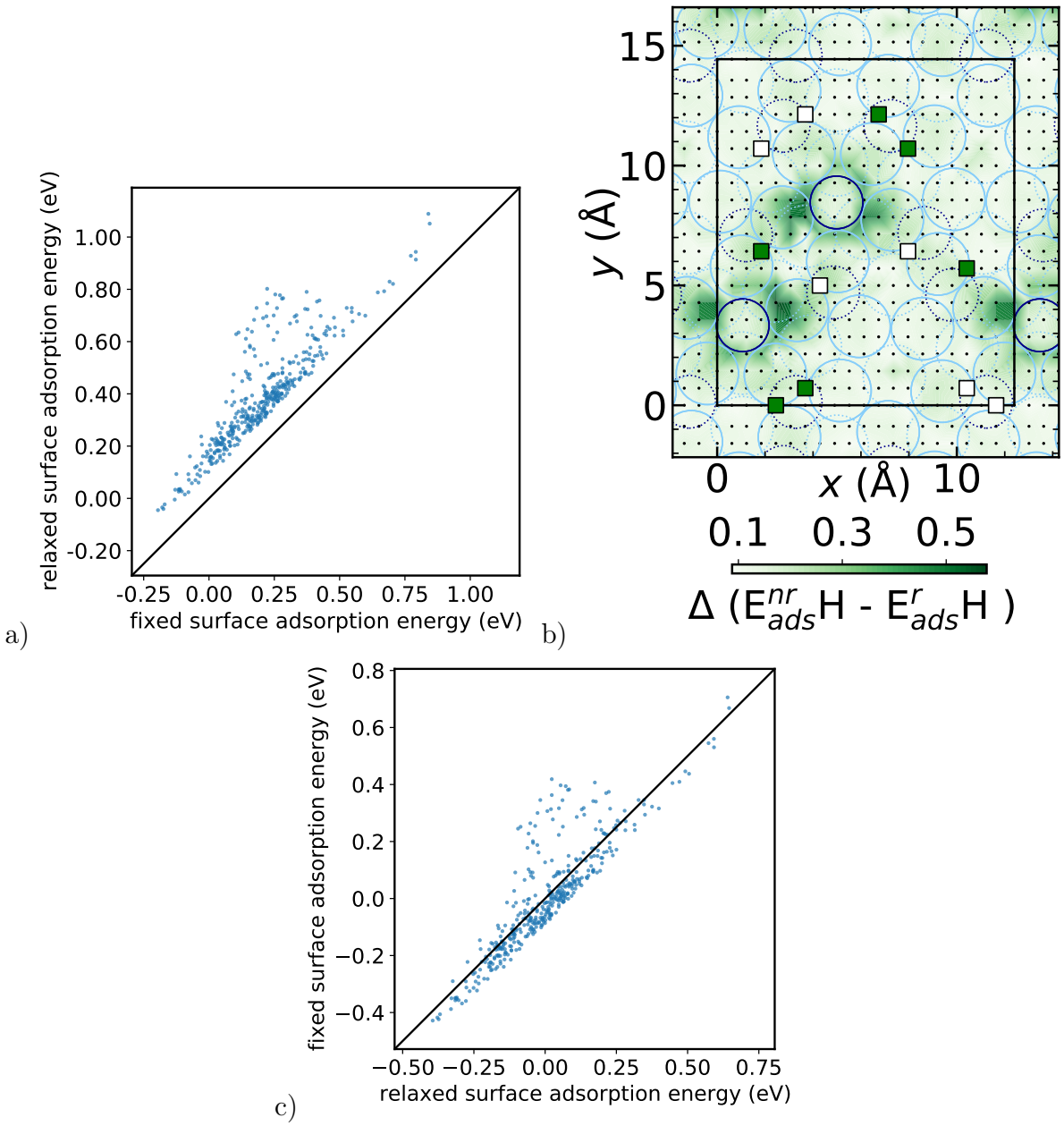

Figure S13: a) comparison between DFT calculations of relaxed and fixed surfaces b) maps of the differences between AEM of DFT calculated relaxed and fixed surfaces. c) Unbiased comparison between DFT calculations of relaxed and fixed surfaces i.e. mean values are subtracted from fixed and relaxed adsorption energies.

S2.2 Prediction of atomic Oxygen adsorption energies

S2.2.1 Adsorption energy maps

A summary of all adsorption energy maps (AEMs) predicted for O adsorption on $Al_13Co_4(100)$ is shown in Fig. S15. At least 100 positions are required to built AEMs that qualitatively mimic the ones calculated with 400 positions, when interpolation methods are used. In contrast, maps are quite well predicted with a low number of positions (in the range [9:36]) when machine learning methods are considered.

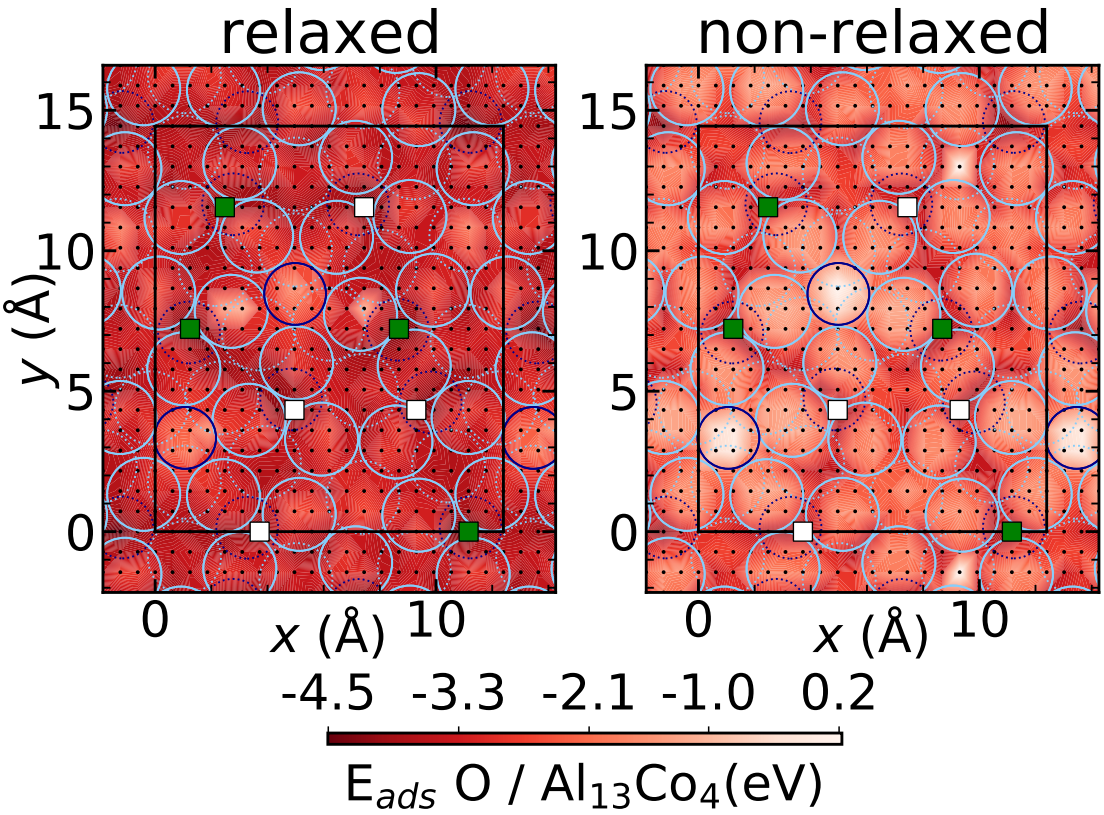

Figure S14: Adsorption energy maps, plotted for E^{nr}_{ads} (top) and E^{r}_{ads} (bottom) by interpolation between 400 DFT optimized values (regular 20×20 grid) for atomic **Oxygen**. The atomic arrangements at the $Al_{13}Co_4(100)$ surface are superimposed. Topmost and subsurface atoms are shown in full and dotted lines, respectively. Color code : Al = light blue, Co = dark blue.

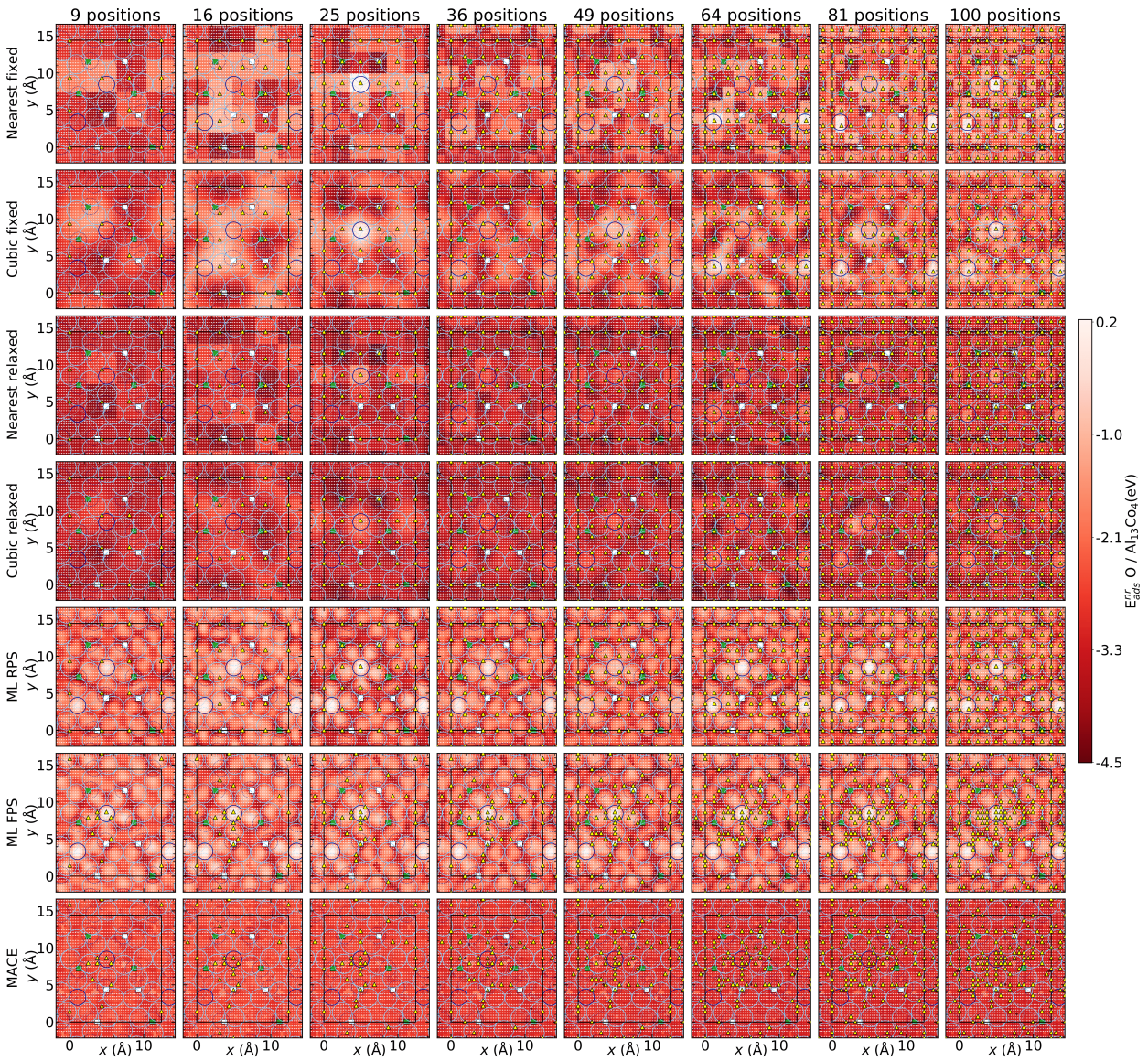

Figure S15: **Oxygen** adsorption energy maps built by considering 9, 25, 36, 49, 64, 91 and 100 (x, y) adsorbate's positions. Machine learning (row 5 and 6 for RPS and FPS with E^{r}_{ads}(**relaxed**) training, row 7 and 8 for RPS and FPS with E^{nr}_{ads}(**fixed**) training), as well as nearest neighbors and cubic interpolations, are used on fixed (row 1 and 2) and relaxed surface (row 3 and 4).

S2.2.2 Error maps and histograms

Figures S16, S17, S18, S19 show maps of normalized residuals (Eq. 5 in the main paper), as well as histograms of errors and scatter plots related to the prediction of H adsorption energies on Al13Co4(100). Training is performed with E^{nr}_{ads} (Figs. S16,S17) or E^{r}_{ads}

(Figs. S18,S19). Histograms for the RPS approach (in blue-green) are more symmetrically distributed. Overall, this illustrates well that there is a risk of inaccurate prediction and improper determination of adsorption sites with the RPS approach. As the training data set increases, the distribution of errors becomes narrower, and the scatter plot aligns more closely with the diagonal. Predictions on fixed surfaces give slightly overestimated adsorption energies, while predictions on relaxed surfaces give slightly underestimated adsorption energies (absolute values).

Figures S20, S21, S22 and S23 show the metrics measured when training is performed on with E_{ads}^{nr} (Figs. S16,S17) or E_{ads}^{r} (Figs. S18,S19). Reference energies are either the ones on the non relaxed surface (Figs. S16,S18) or on the relaxed surface (Figs. S17,S19). The influence of the surface relaxation (following adsorption) on the ML results is illustrated in Fig. S24.

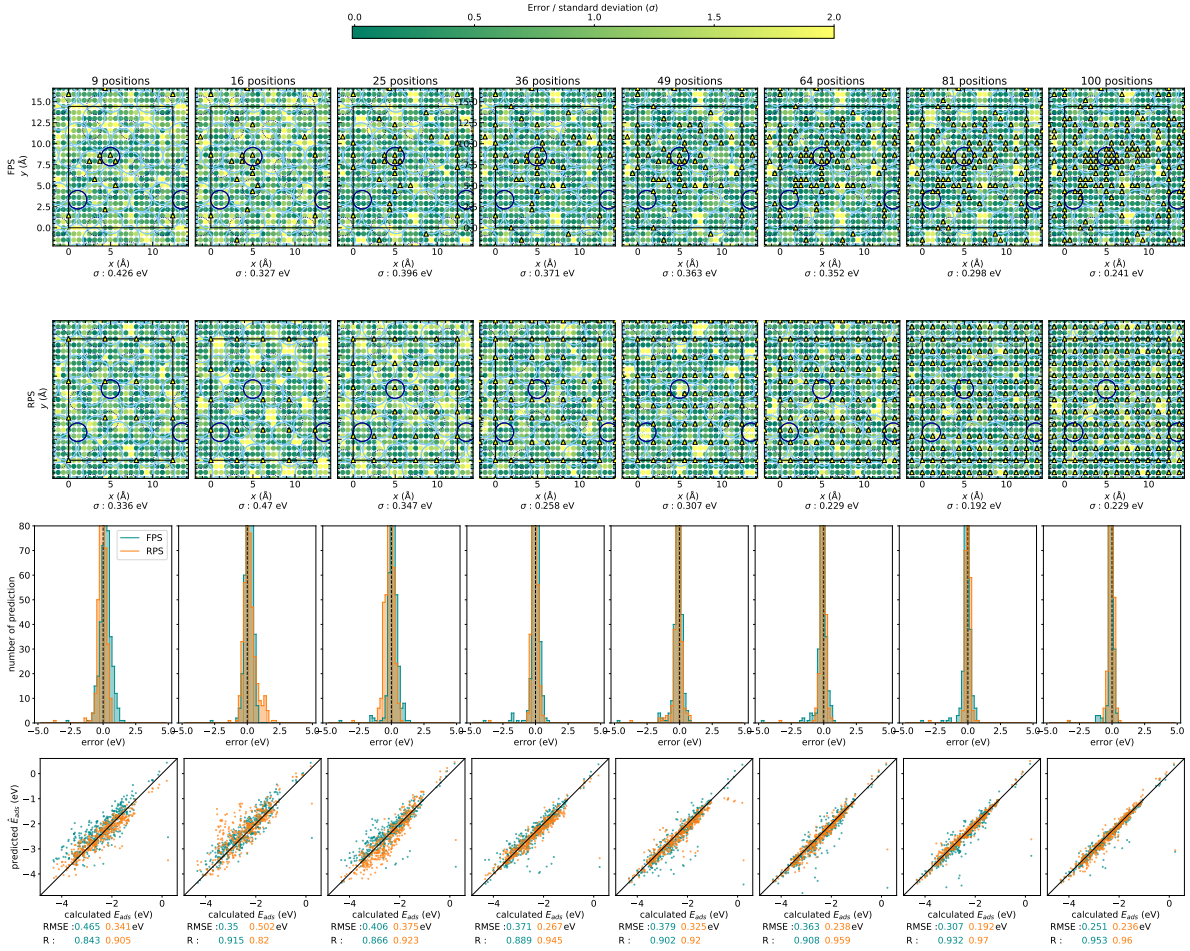

Figure S16: Error maps and histograms for atomic **Oxygen** adsorption energies. Training is done on E_{ads}^{nr}(**fixed**). Results are compared with E_{ads}^{nr}(**fixed**). The positions are selected on regular grids and FPS method.

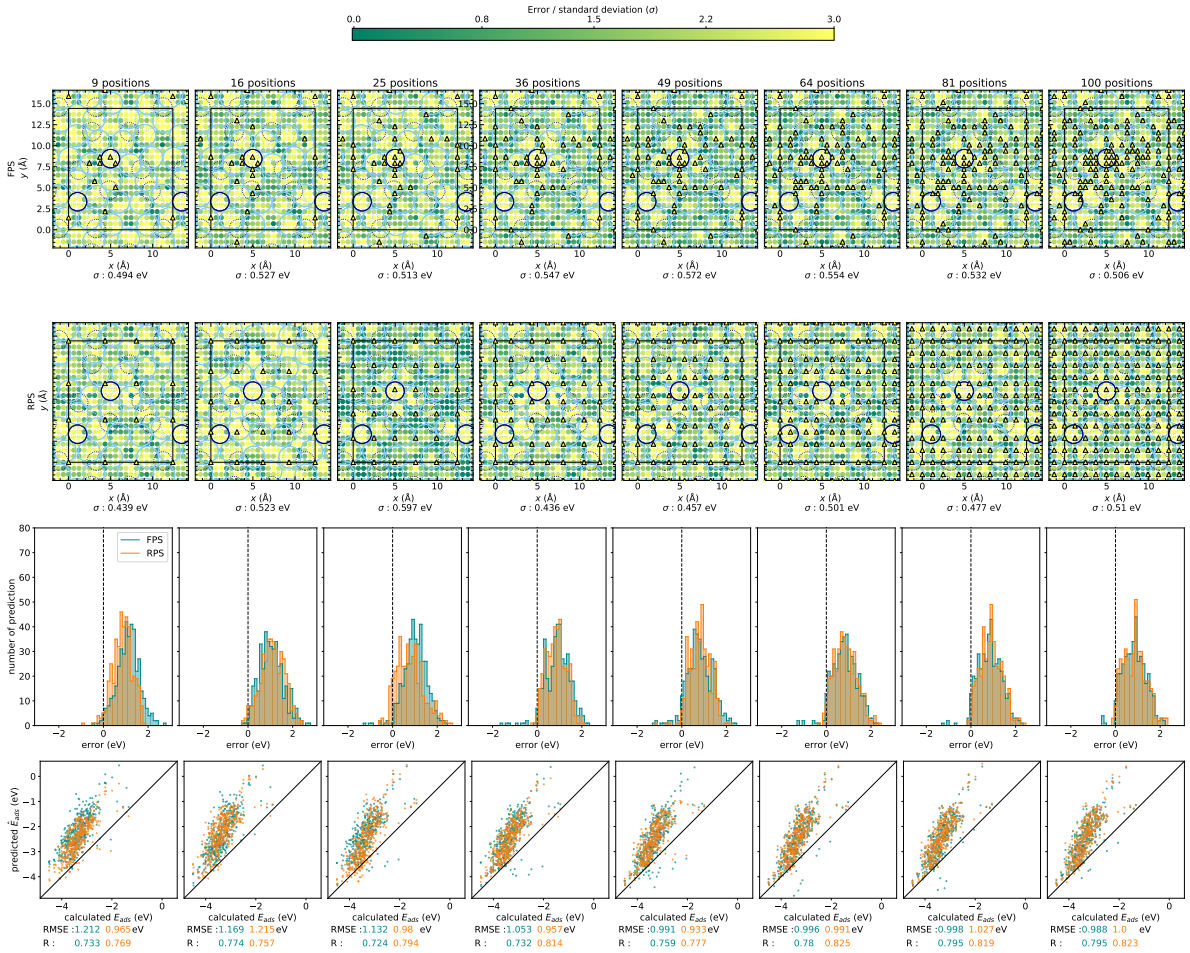

Figure S17: Error maps and histograms for atomic **Oxygen** adsorption energies. Training is done on E_{ads}^{nr}(**fixed**). Results are compared with E_{ads}^{r}(**relaxed**). The positions are selected on regular grids and FPS method.

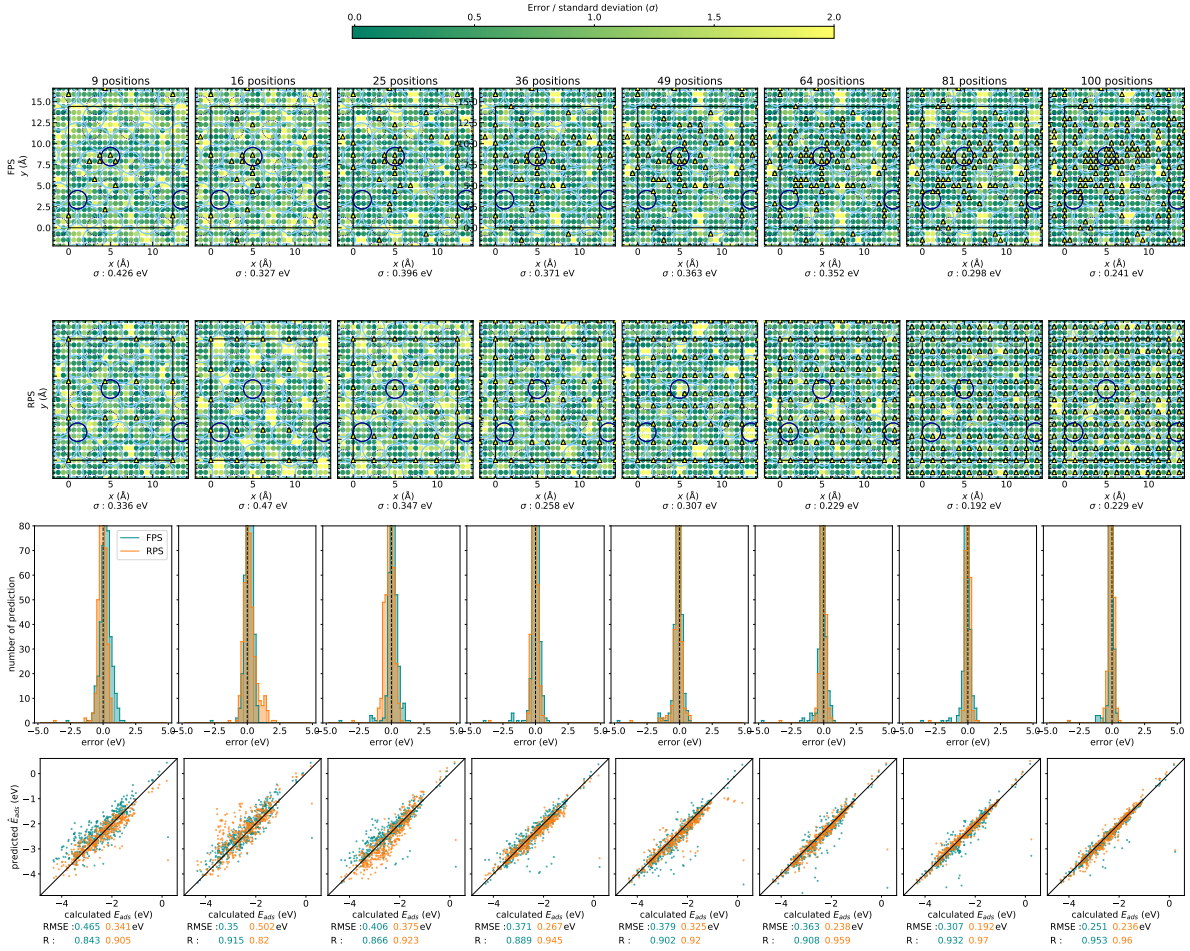

Figure S18: Error maps and histograms for atomic **Oxygen** adsorption energies. Training is done on E^{r}_{ads}(**relaxed**). Results are compared with E^{nr}_{ads}(**fixed**). The positions are selected on regular grids and FPS method.

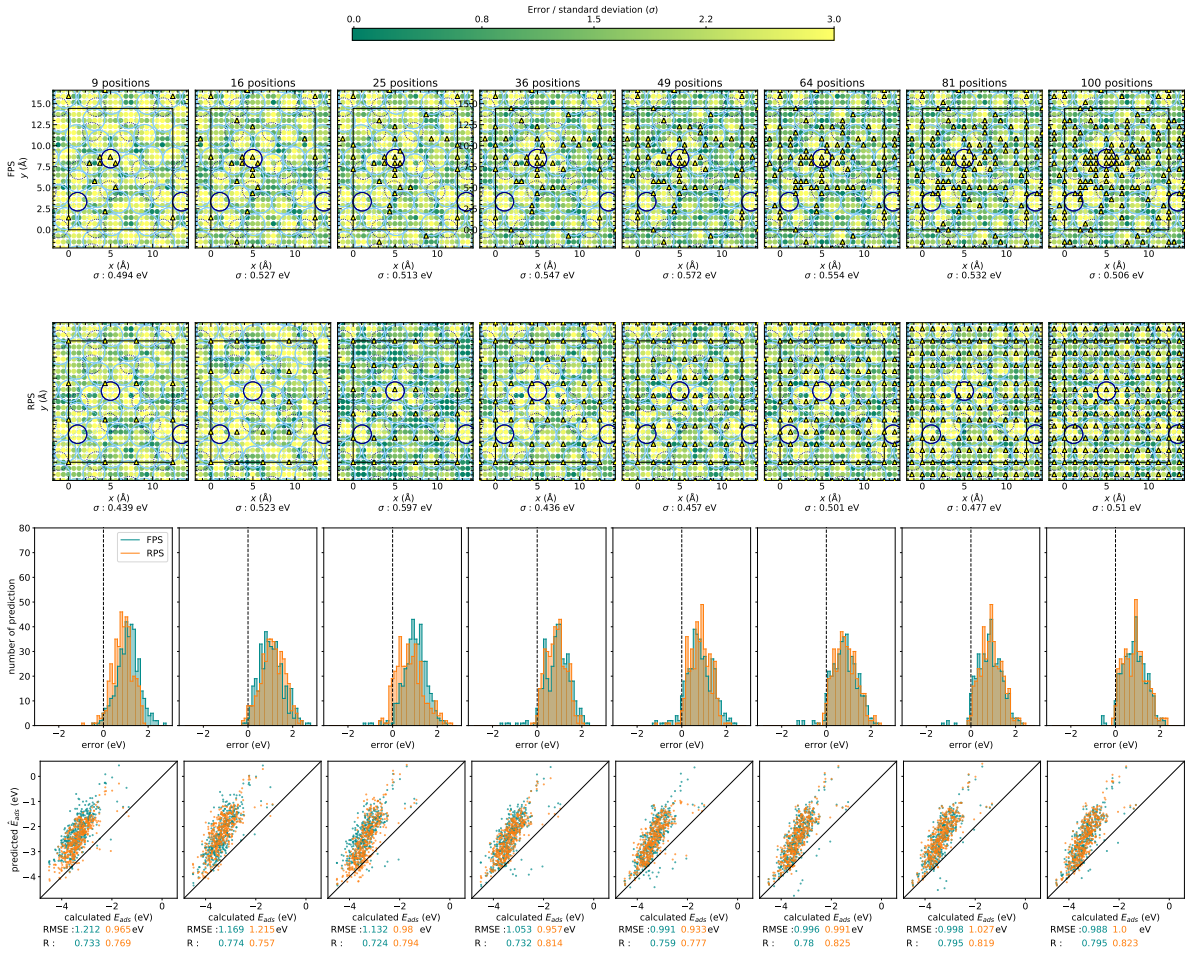

Figure S19: Error maps and histograms for atomic **Oxygen** adsorption energies. Training is done on E^r_{ads}(**relaxed**). Results are compared with E^r_{ads}(**relaxed**). The positions are selected on regular grids and FPS method.

S2.2.3 Metrics

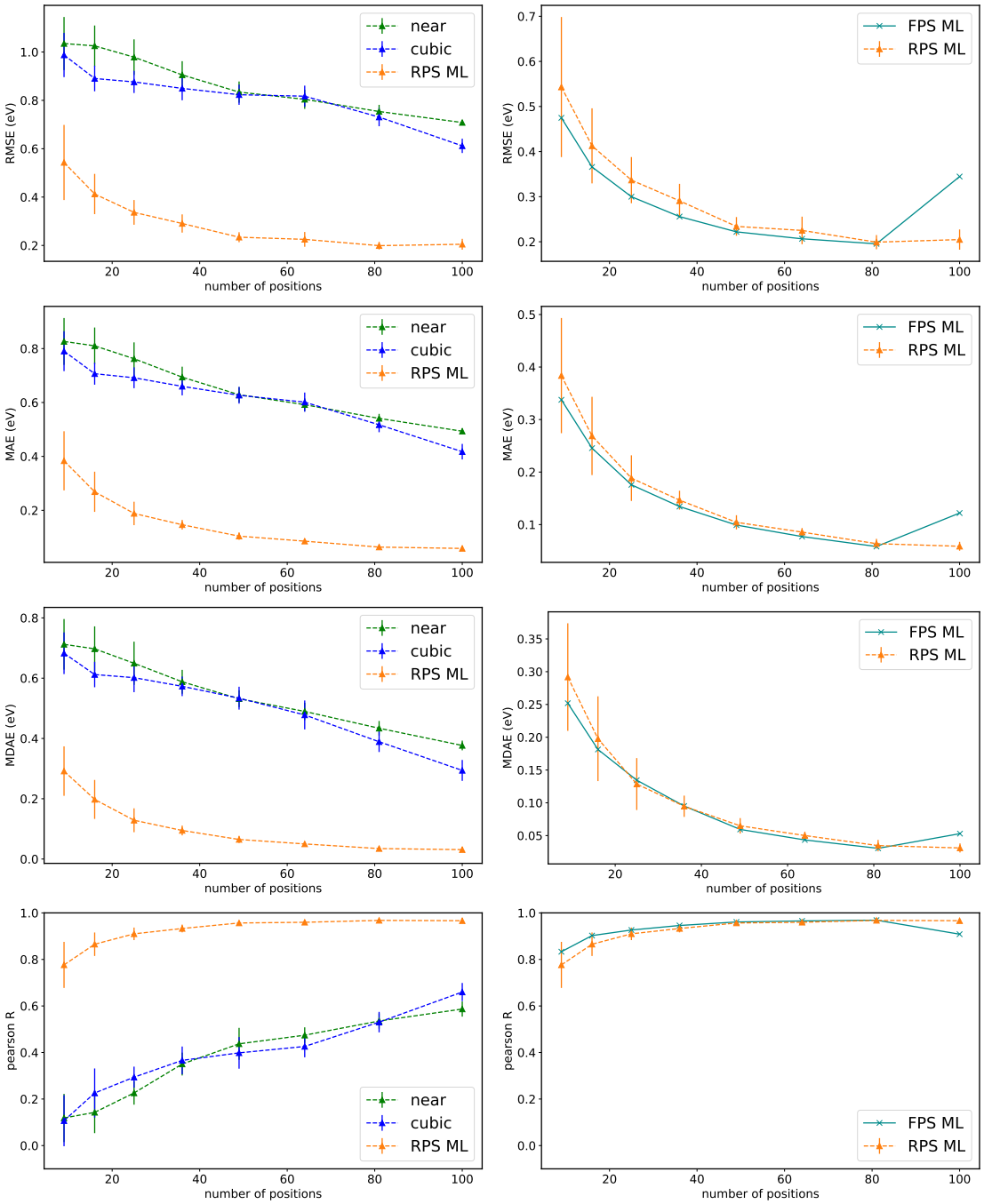

Figure S20: Metrics for **Oxygen** adsorption. Training is done on E_{ads}^{nr}(**fixed**). Results are compared with E_{ads}^{nr}(**fixed**). RMSE (first line), MAE (second line), MAD (third line) and Pearson coefficient (fourth line) as a function of the number of DFT-calculated values in the training set. The sites considered for the training are located on a regularly spaced grid (dashed lines and triangles, labeled by *regular grid*) or are selected by our method (solid lines and crosses, labeled by *FPS*). Results of interpolation methods are represented in green (nearest neighbor interpolation) and in blue (cubic interpolation). Results of the machine learning approach are shown in orange (*regular grid*) and cyan (*FPS*).

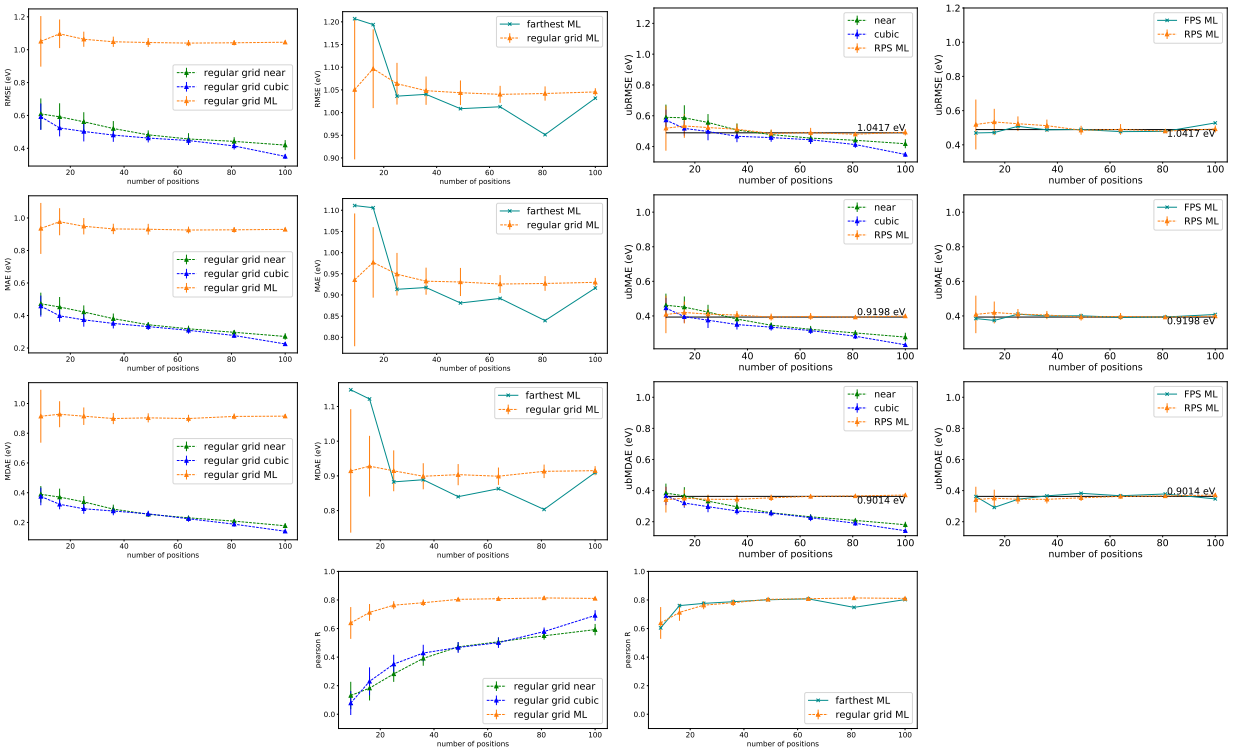

Figure S21: Metrics for **Oxygen** adsorption. Training is done on E_{ads}^{nr}(**fixed**). Results are compared with E_{ads}^{r}(**relaxed**). RMSE (first line), MAE (second line), MAD (third line) and Pearson coefficient (fourth line) as a function of the number of DFT-calculated values in the training set. The sites considered for the training are located on a regularly spaced grid (dashed lines and triangles, labeled by *regular grid*) or are selected by our method (solid lines and crosses, labeled by *FPS*). Results of interpolation methods are represented in green (nearest neighbor interpolation) and in blue (cubic interpolation). Results of the machine learning approach are shown in orange (*regular grid*) and cyan (*FPS*). for the three first line (RMSE, MAE, MAD), results are shown with biased (first and second column) and unbiased (third and fourth column) errors. Horizontal black line represented the value given by the metric considering the difference between E_{ads}^{nr}(**fixed**) and E_{ads}^{r}(**relaxed**) as the error (see Fig S24

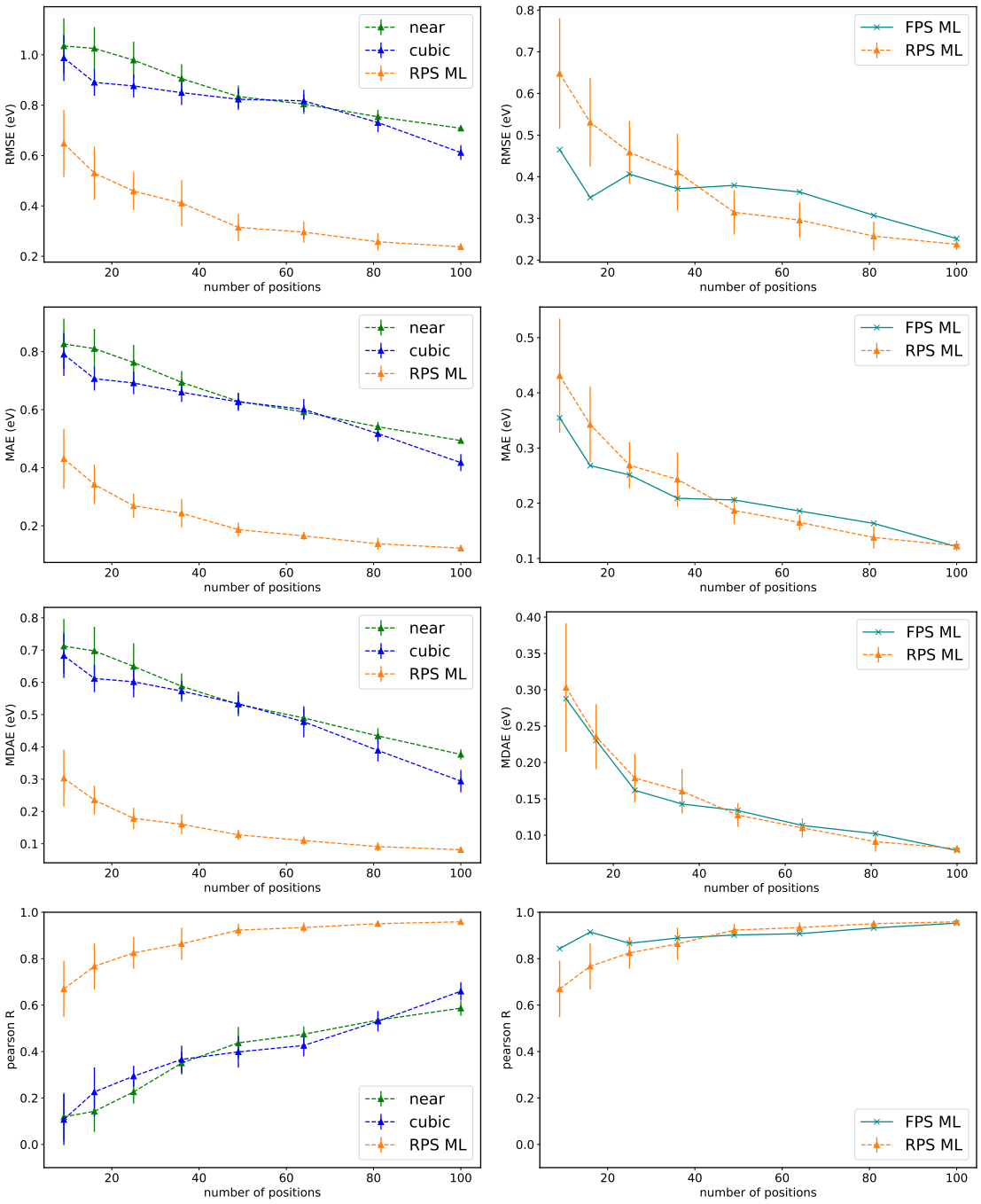

Figure S22: Metrics for **Oxygen** adsorption. Training is done on E^{r}_{ads}(**relaxed**). Results are compared with E^{nr}_{ads}(**fixed**). RMSE (first line), MAE (second line), MAD (third line) and Pearson coefficient (fourth line) as a function of the number of DFT-calculated values in the training set. The sites considered for the training are located on a regularly spaced grid (dashed lines and triangles, labeled by *regular grid*) or are selected by our method (solid lines and crosses, labeled by *FPS*). Results of interpolation methods are represented in green (nearest neighbor interpolation) and in blue (cubic interpolation). Results of the machine learning approach are shown in orange (*regular grid*) and cyan (*FPS*).

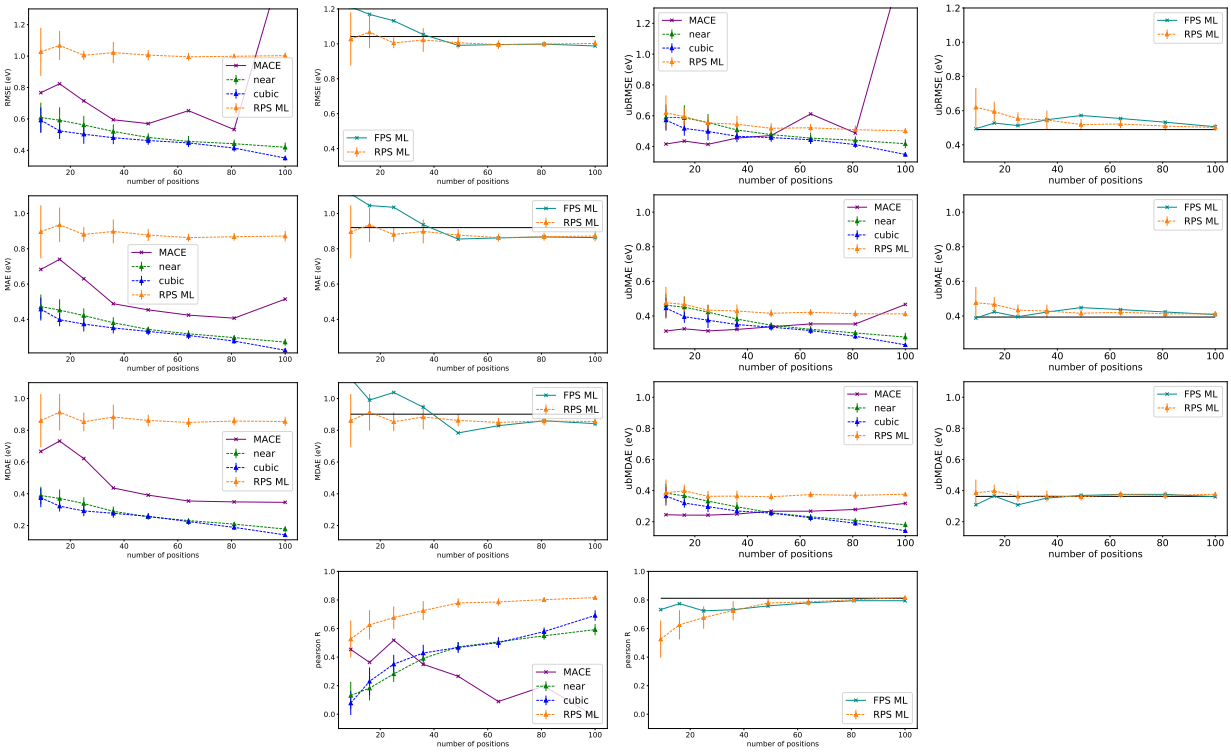

Figure S23: Metrics for **Oxygen** adsorption. Training is done on E^r_{ads}(**relaxed**). Results are compared with E^r_{ads}(**relaxed**). RMSE (first line), MAE (second line), MAD (third line) and Pearson coefficient (fourth line) as a function of the number of DFT-calculated values in the training set. The sites considered for the training are located on a regularly spaced grid (dashed lines and triangles, labeled by *regular grid*) or are selected by our method (solid lines and crosses, labeled by *FPS*). Results of interpolation methods are represented in green (nearest neighbor interpolation) and in blue (cubic interpolation). Results of the machine learning approach are shown in orange (*regular grid*) and cyan (*FPS*). for the three first line (RMSE, MAE, MAD), results are shown with biased (first and second column) and unbiased (third and fourth column) errors. Horizontal black line represented the value given by the metric considering the difference between E^{nr}_{ads}(**fixed**) and E^r_{ads}(**relaxed**) as the error (see Fig S24

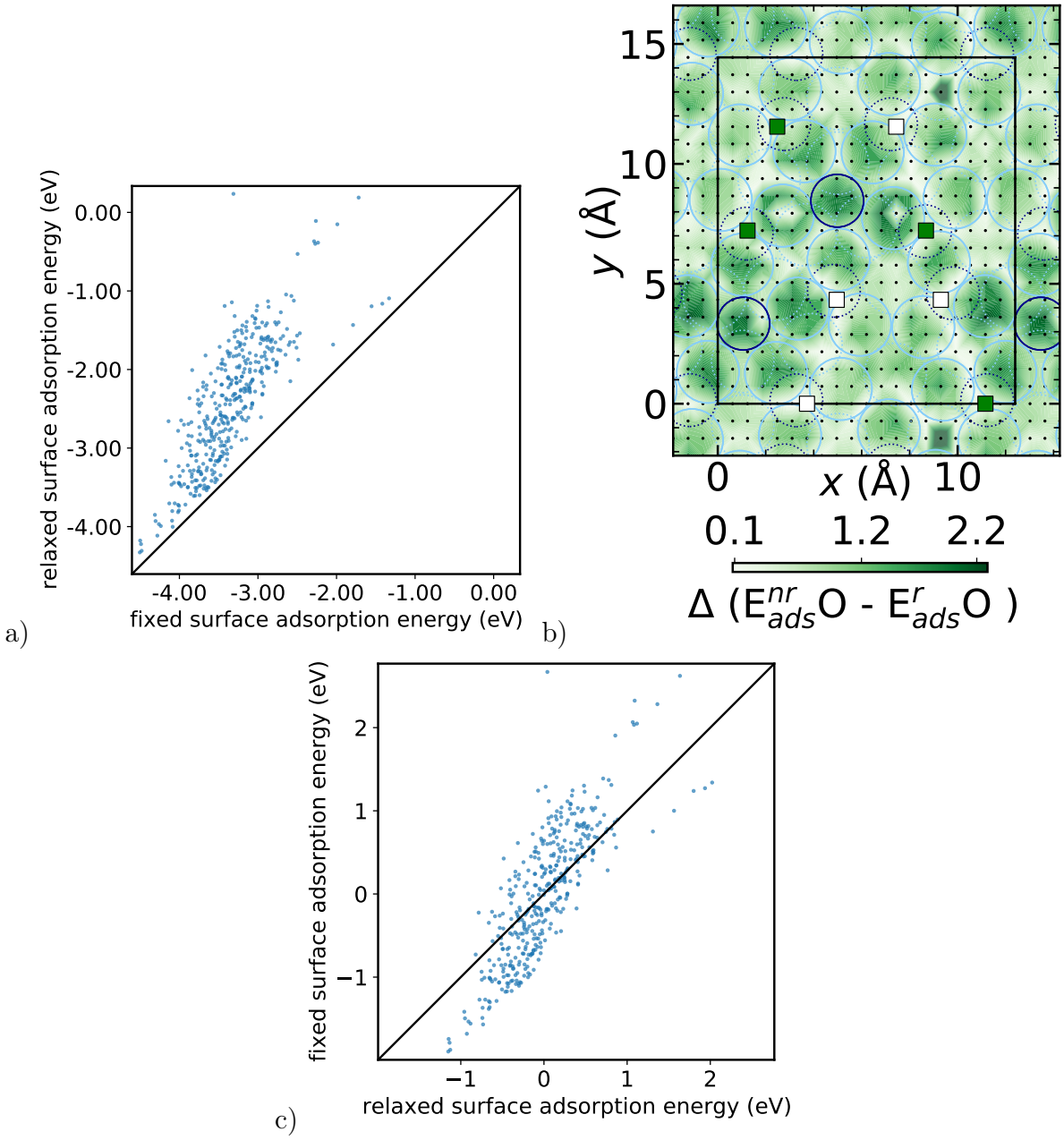

Figure S24: a) comparison between DFT calculations of relaxed and fixed surfaces b) maps of the differences between AEM of DFT calculated relaxed and fixed surfaces. c) Unbiased comparison between DFT calculations of relaxed and fixed surfaces i.e. mean values are subtracted from fixed and relaxed adsorption energies.

S2.3 Prediction of atomic Lead adsorption energies

S2.3.1 Adsorption energy maps

A summary of all adsorption energy maps (AEMs) predicted for O adsorption on $Al_13Co_4(100)$ is shown in Fig. S26. At least 100 positions are required to built AEMs that qualitatively mimic the ones calculated with 400 positions, when interpolation methods are used. In contrast, maps are quite well predicted with a low number of positions (in the range [9:36]) when machine learning methods are considered.

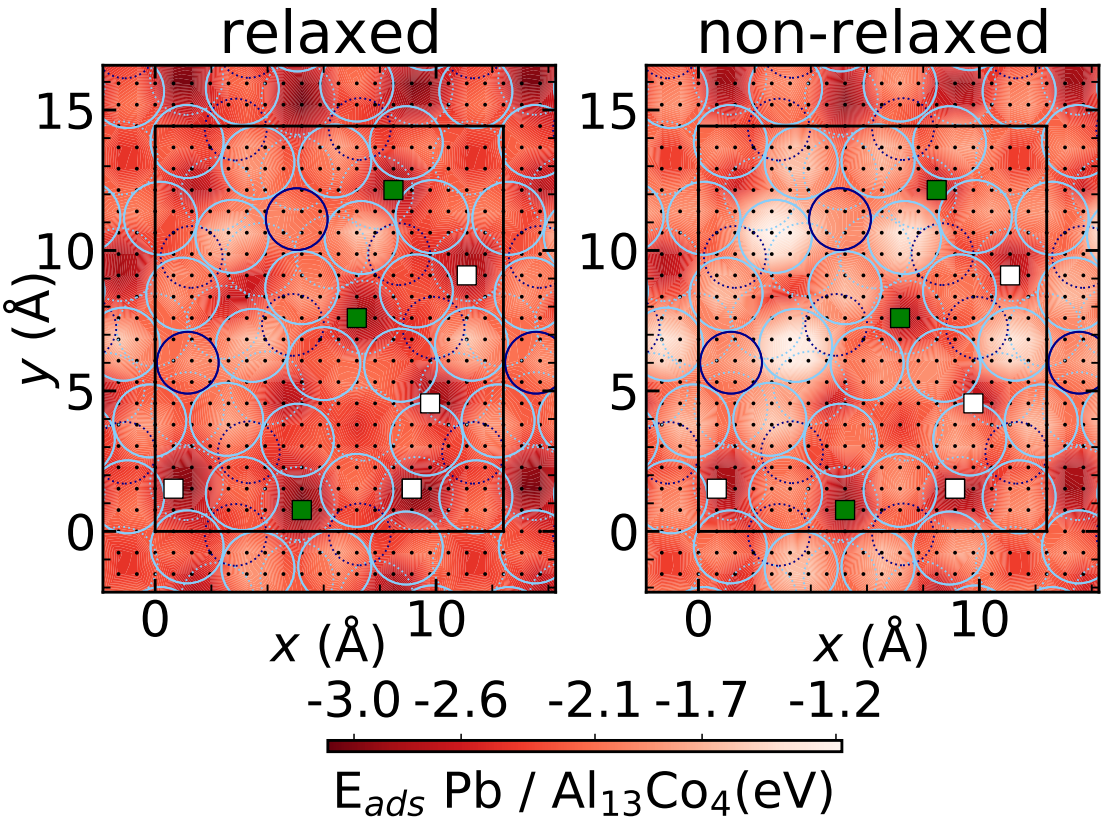

Figure S25: Adsorption energy maps, plotted for E^{nr}_{ads} (top) and E^{r}_{ads} (bottom) by interpolation between 361 DFT optimized values (regular 19×19 grid) for atomic **Lead**. The atomic arrangements at the $Al_{13}Co_4(100)$ surface are superimposed. Topmost and subsurface atoms are shown in full and dotted lines, respectively. Color code : Al = light blue, Co = dark blue.

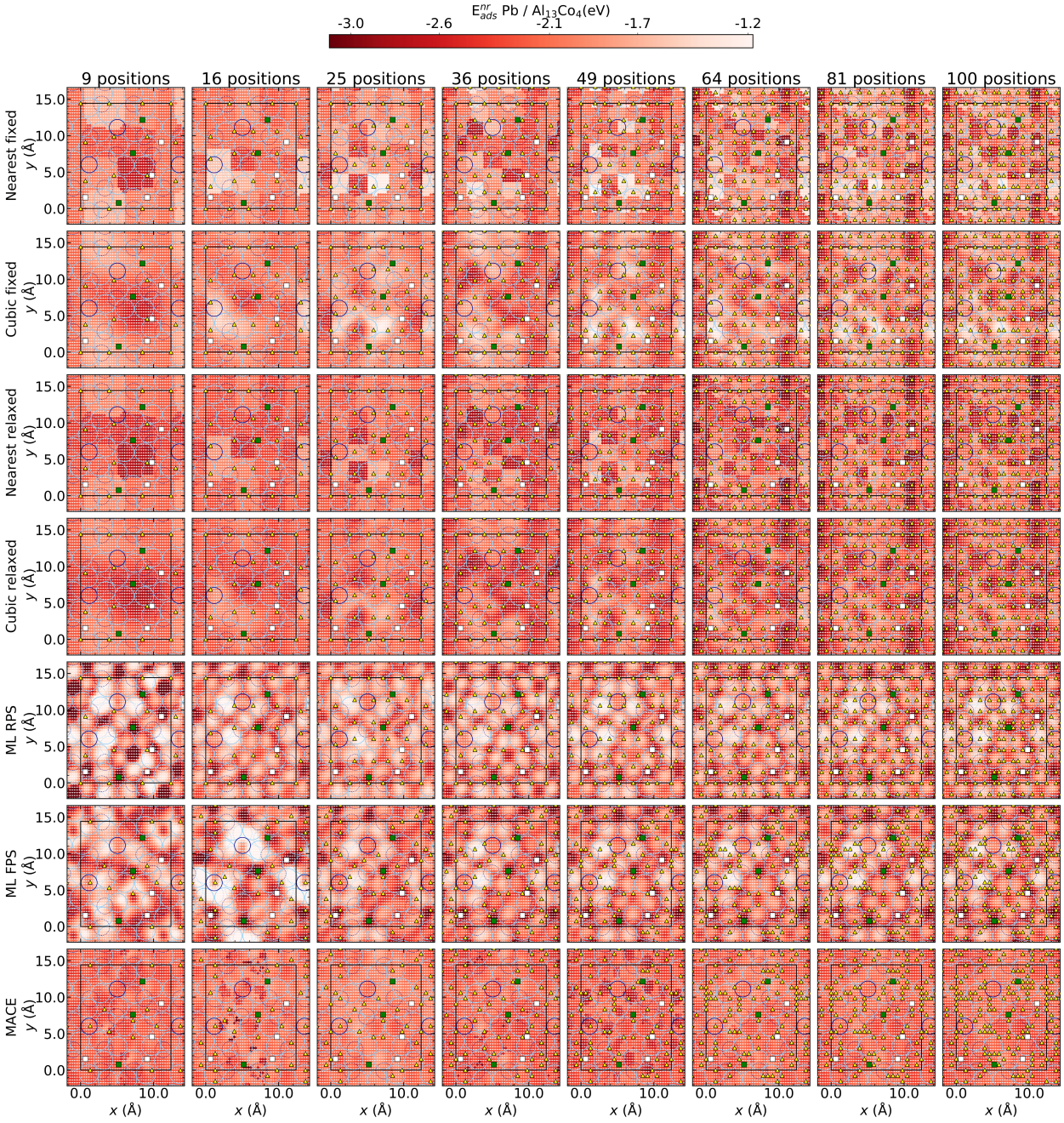

Figure S26: **Lead** adsorption energy maps built by considering 9, 25, 36, 49, 64, 91 and 100 (x, y) adsorbate's positions. Machine learning (row 5 and 6 for RPS and FPS with E^{r}_{ads}(**relaxed**) training, row 7 and 8 for RPS and FPS with E^{nr}_{ads}(**fixed**) training), as well as nearest neighbors and cubic interpolations, are used on fixed (row 1 and 2) and relaxed surface (row 3 and 4).

S2.3.2 Error maps and histograms

Figures S27, S28, S29, S30 show maps of normalized residuals (Eq. 5 in the main paper), as well as histograms of errors and scatter plots related to the prediction of H adsorption energies on Al13Co4(100). Training is performed with E_{ads}^{nr} (Figs. S27,S28) or E_{ads}^{r} (Figs. S29,S30). Histograms for the RPS approach (in blue-green) are more symmetrically distributed. Overall, this illustrates well that there is a risk of inaccurate prediction and improper determination of adsorption sites with the RPS approach. As the training data set increases, the distribution of errors becomes narrower, and the scatter plot aligns more closely with the diagonal. Predictions on fixed surfaces give slightly overestimated adsorption energies, while predictions on relaxed surfaces give slightly underestimated adsorption energies (absolute values).

Figures S31, S32, S33 and S34 show the metrics measured when training is performed on with E_{ads}^{nr} (Figs. S31, S32) or E_{ads}^{r} (Figs. S33, S34). Reference energies are either the ones on the non relaxed surface (Figs. S31, S33) or on the relaxed surface (Figs. S32, S34). The influence of the surface relaxation (following adsorption) on the ML results is illustrated in Fig. S35.

Figure S27: Error maps and histograms for atomic **Lead** adsorption energies. Training is done on E^{nr}_{ads}(**fixed**). Results are compared with E^{r}_{ads}(**relaxed**). The positions are selected on regular grids and FPS method.

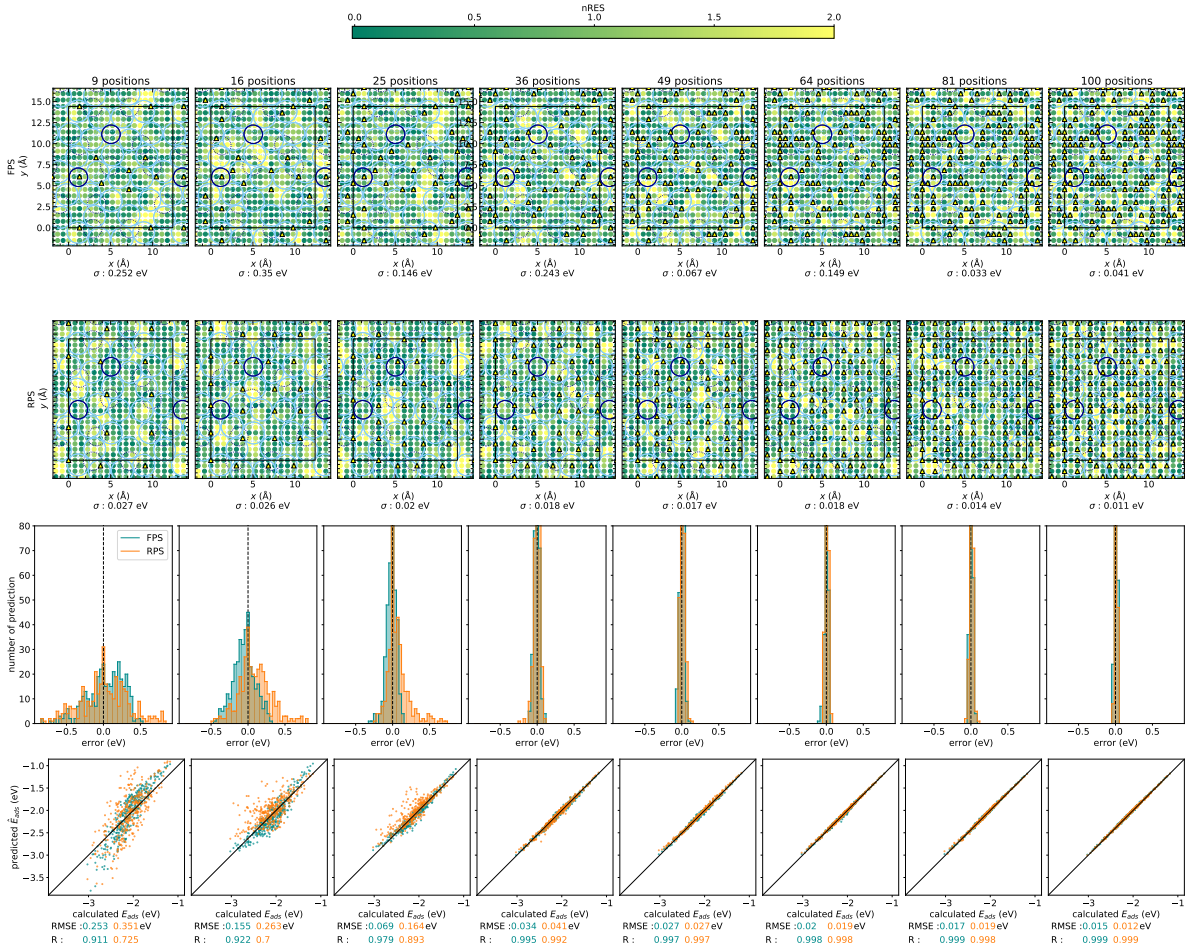

Figure S28: Error maps and histograms for atomic **Lead** adsorption energies. Training is done on E_{ads}^{nr}(**fixed**). Results are compared with E_{ads}^{nr}(**fixed**). The positions are selected on regular grids and FPS method.

Figure S29: Error maps and histograms for atomic **Lead** adsorption energies. Training is done on E^r_{ads}(**relaxed**). Results are compared with E^{nr}_{ads}(**fixed**). The positions are selected on regular grids and FPS method.

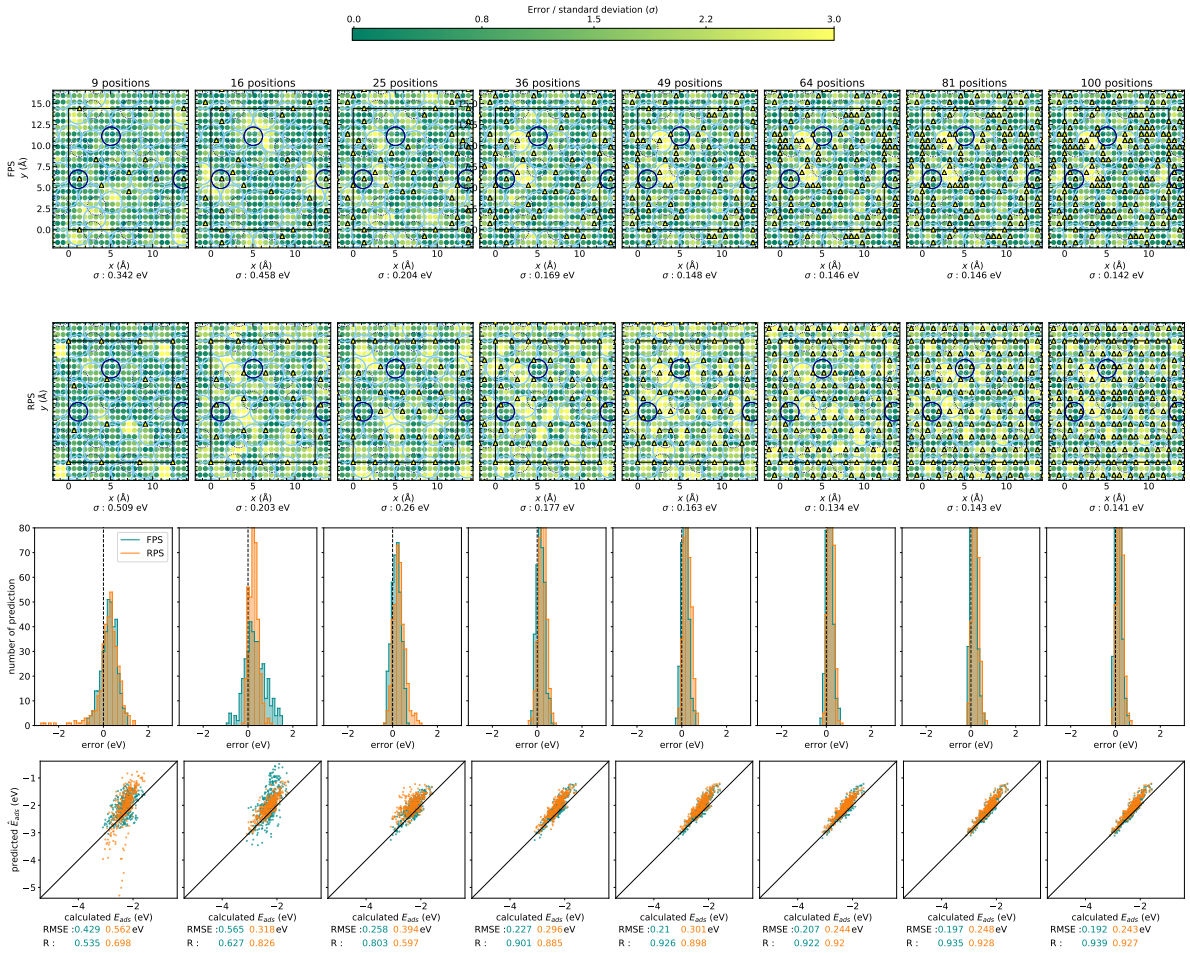

Figure S30: Error maps and histograms for atomic **Lead** adsorption energies. Training is done on E^r_{ads}(**relaxed**). Results are compared with E^r_{ads}(**relaxed**). The positions are selected on regular grids and FPS method.

S2.3.3 Metrics

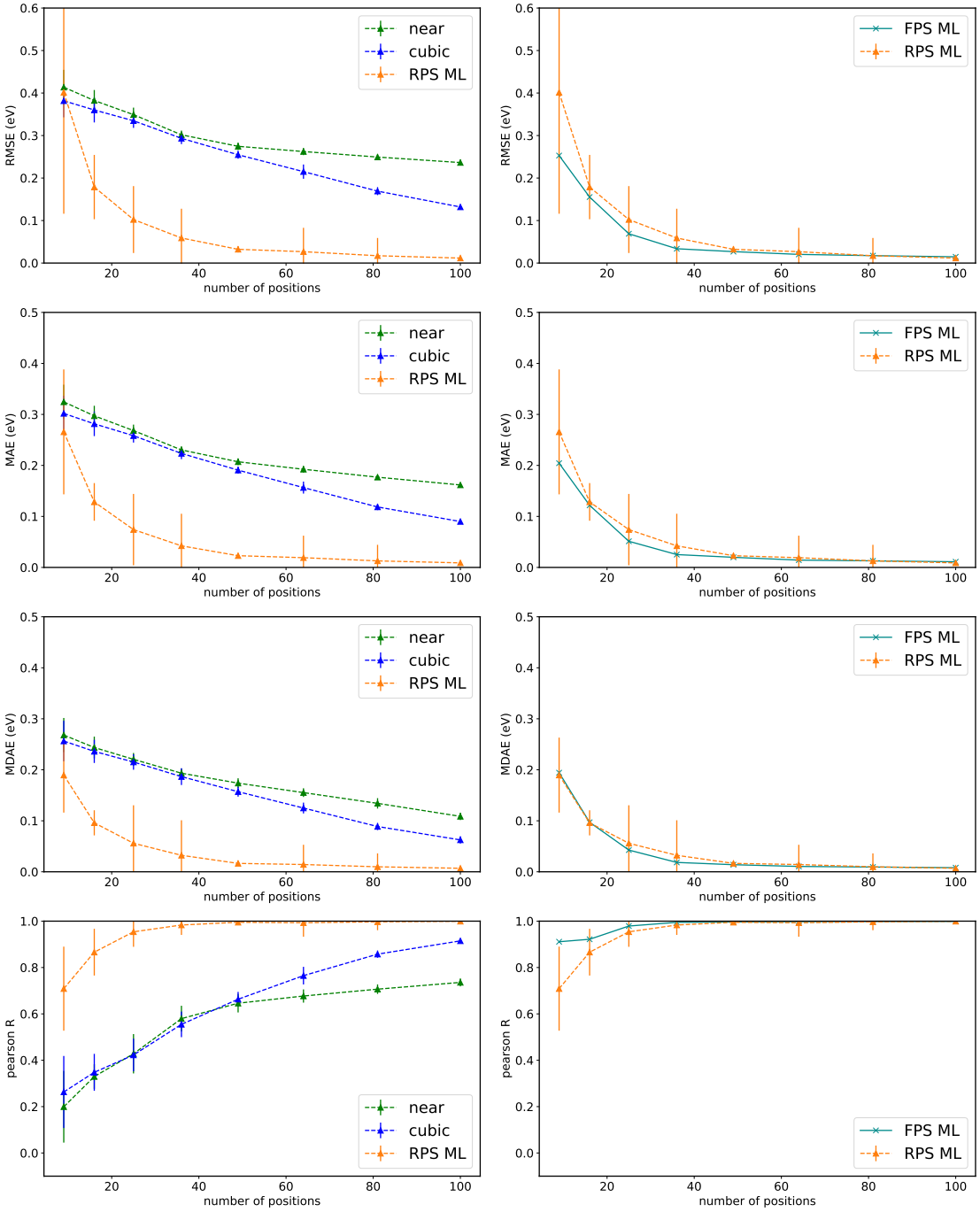

Figure S31: Metrics for **Lead** adsorption. Training is done on E_{ads}^{nr}(**fixed**). Results are compared with E_{ads}^{nr}(**fixed**). RMSE (first line), MAE (second line), MAD (third line) and Pearson coefficient (fourth line) as a function of the number of DFT-calculated values in the training set. The sites considered for the training are located on a regularly spaced grid (dashed lines and triangles, labeled by *regular grid*) or are selected by our method (solid lines and crosses, labeled by *FPS*). Results of interpolation methods are represented in green (nearest neighbor interpolation) and in blue (cubic interpolation). Results of the machine learning approach are shown in orange (*regular grid*) and cyan (*FPS*).

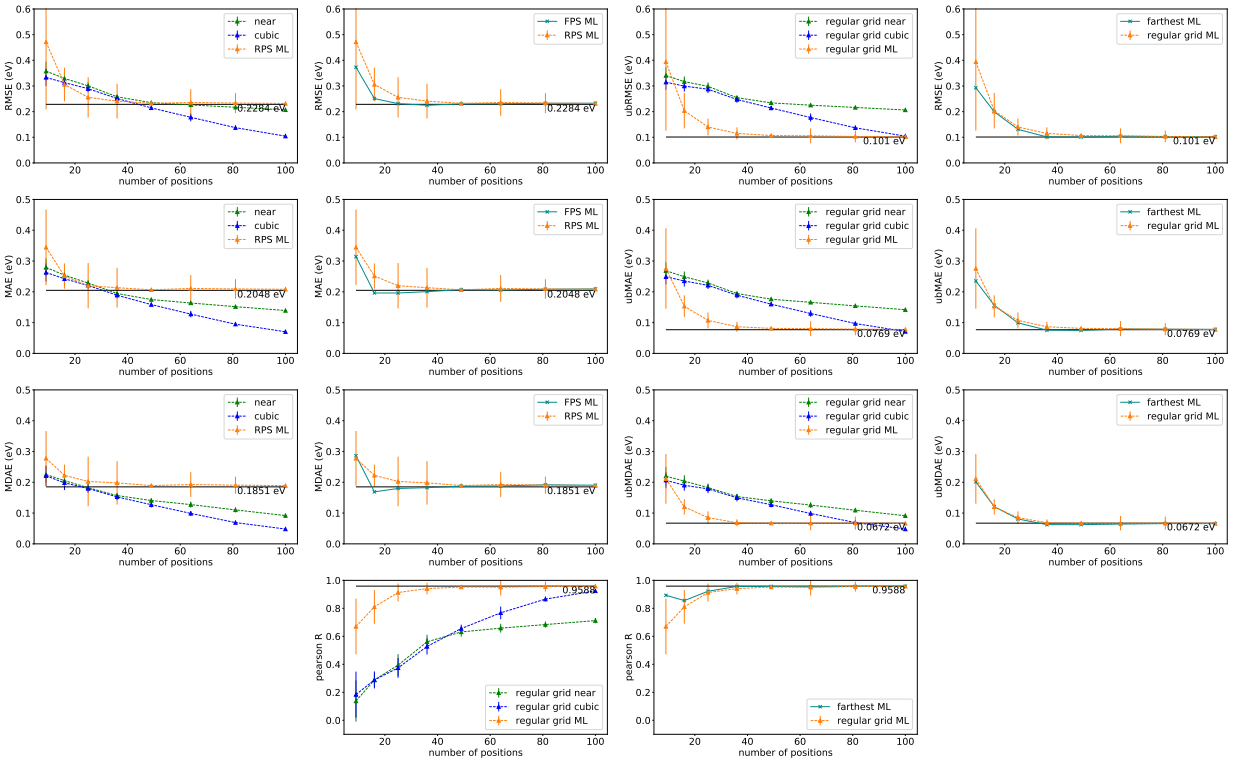

Figure S32: Metrics for **Lead** adsorption. Training is done on E_{ads}^{nr}(**fixed**). Results are compared with E_{ads}^{r}(**relaxed**). RMSE (first line), MAE (second line), MAD (third line) and Pearson coefficient (fourth line) as a function of the number of DFT-calculated values in the training set. The sites considered for the training are located on a regularly spaced grid (dashed lines and triangles, labeled by *regular grid*) or are selected by our method (solid lines and crosses, labeled by *FPS*). Results of interpolation methods are represented in green (nearest neighbor interpolation) and in blue (cubic interpolation). Results of the machine learning approach are shown in orange (*regular grid*) and cyan (*FPS*). for the three first line (RMSE, MAE, MAD), results are shown with biased (first and second column) and unbiased (third and fourth column) errors. Horizontal black line represented the value given by the metric considering the difference between E_{ads}^{nr}(**fixed**) and E_{ads}^{r}(**relaxed**) as the error (see Fig S35

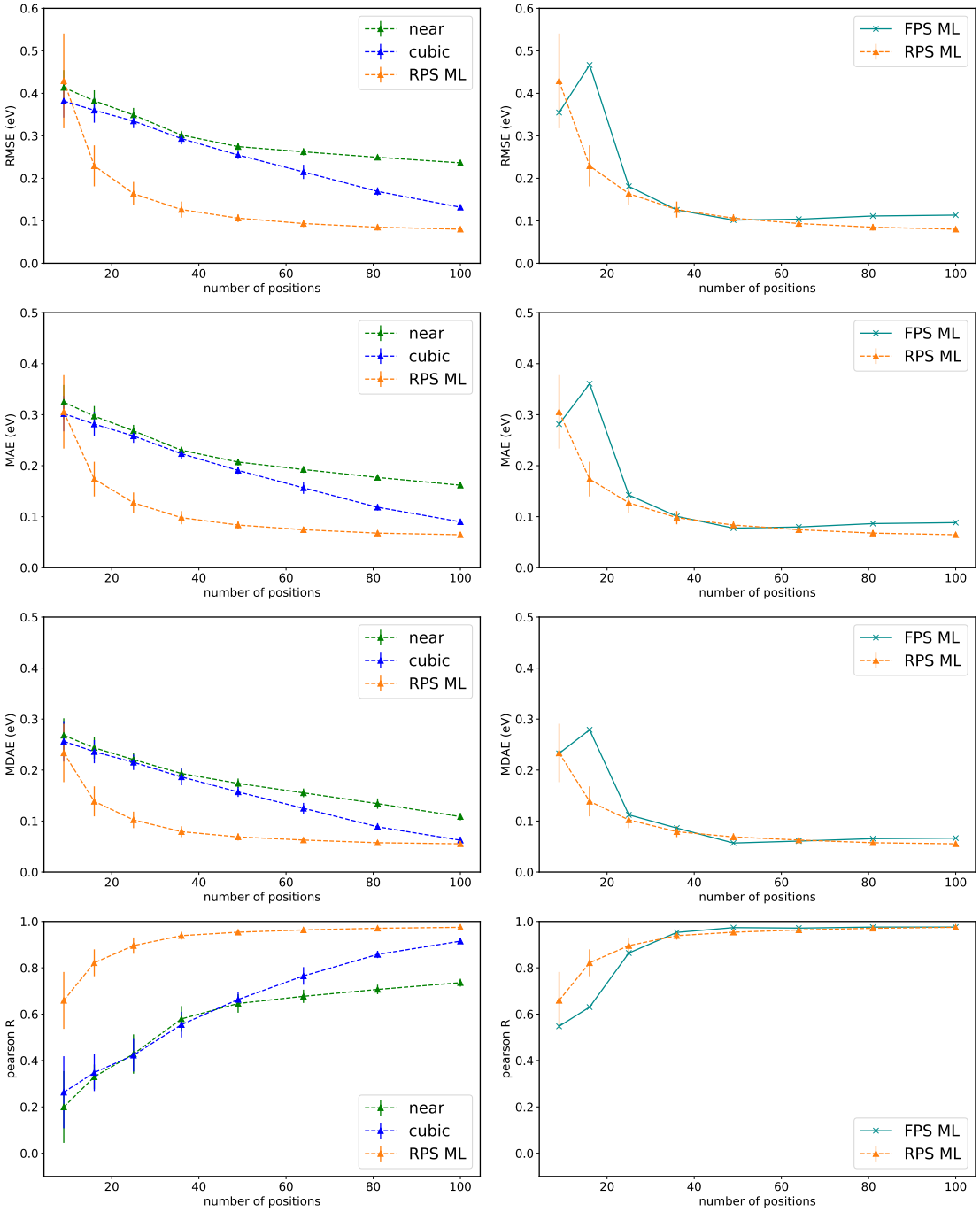

Figure S33: Metrics for **Lead** adsorption. Training is done on E^{r}_{ads}(**relaxed**). Results are compared with E^{nr}_{ads}(**fixed**). RMSE (first line), MAE (second line), MAD (third line) and Pearson coefficient (fourth line) as a function of the number of DFT-calculated values in the training set. The sites considered for the training are located on a regularly spaced grid (dashed lines and triangles, labeled by *regular grid*) or are selected by our method (solid lines and crosses, labeled by *FPS*). Results of interpolation methods are represented in green (nearest neighbor interpolation) and in blue (cubic interpolation). Results of the machine learning approach are shown in orange (*regular grid*) and cyan (*FPS*).

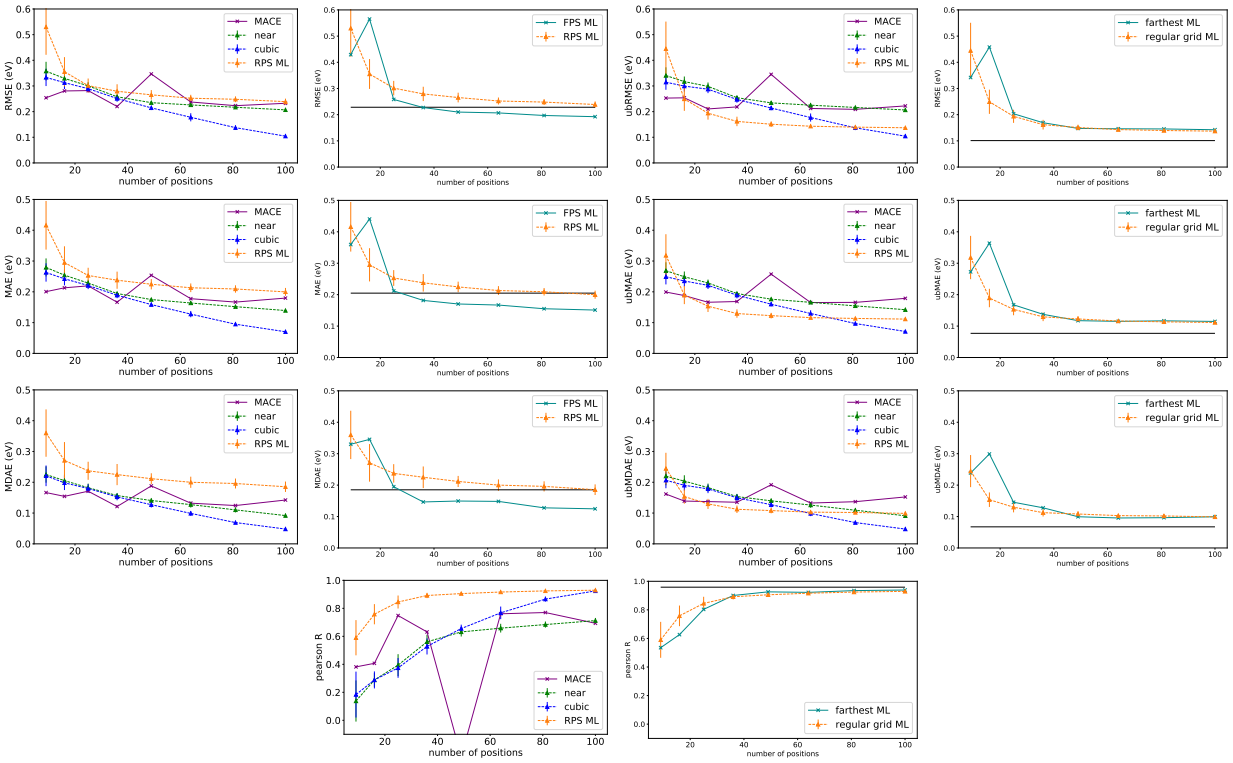

Figure S34: Metrics for **Lead** adsorption. Training is done on E^r_{ads}(**relaxed**). Results are compared with E^r_{ads}(**relaxed**). RMSE (first line), MAE (second line), MAD (third line) and Pearson coefficient (fourth line) as a function of the number of DFT-calculated values in the training set. The sites considered for the training are located on a regularly spaced grid (dashed lines and triangles, labeled by *regular grid*) or are selected by our method (solid lines and crosses, labeled by *FPS*). Results of interpolation methods are represented in green (nearest neighbor interpolation) and in blue (cubic interpolation). Results of the machine learning approach are shown in orange (*regular grid*) and cyan (*FPS*). for the three first line (RMSE, MAE, MAD), results are shown with biased (first and second column) and unbiased (third and fourth column) errors. Horizontal black line represented the value given by the metric considering the difference between E^{nr}_{ads}(**fixed**) and E^r_{ads}(**relaxed**) as the error (see Fig S35

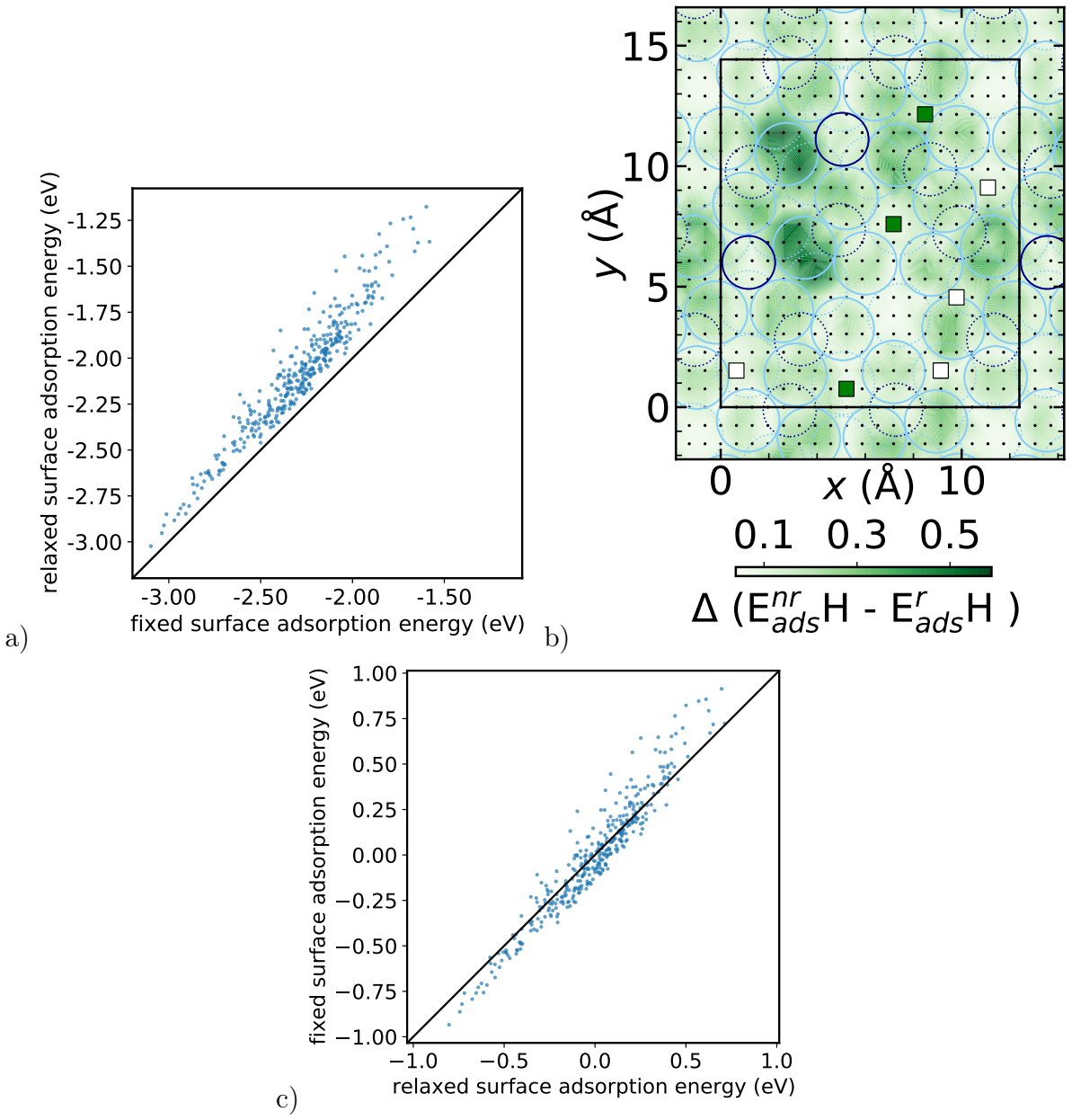

Figure S35: a) Comparison between DFT calculations of relaxed and fixed surfaces. b) maps of the differences between AEM of DFT calculated relaxed and fixed surfaces. c) Unbiased comparison between DFT calculations of relaxed and fixed surfaces i.e. mean values are subtracted from fixed and relaxed adsorption energies.

S2.4 The $\Lambda(n)$ metric for all adsorbates

Table S4: Values of $\Lambda(n)$ measured for the H, O and Pb adsorbates on $Al_{13}Co_4(100)$.

n	9	16	25	36	49	64	81	100
Hydrogen								
FPS	3	20	0	0	0	0	0	0
RPS *mean*	4.84	3.68	5.62	3.52	1.30	0.76	0.63	0.5
std deviation	11.52	4.97	6.40	4.68	2.62	1.58	1.73	0.5
Near *mean*	4.37	4.64	10.5	10.67	7.59	4.08	3.05	4.5
std deviation	8.59	10.2	13.12	11.75	9.96	5.45	4.87	4.46
Cubic *mean*	6.28	4.44	7.75	8.01	5.03	4.72	3.73	3.25
std deviation	9.84	5.82	5.41	8.61	5.8	6.67	4.72	2.77
Oxygen								
FPS	10	0	0	0	0	2	2	0
RPS *mean*	2.78	2.32	0	0.86	0.09	0.24	0.05	0
std deviation	6.57	4.86	0	1.79	0.54	0.86	0.35	0
Near *mean*	7.51	11.12	10.19	10.01	9.78	12.4	11.935	8.75
std deviation	9.60	12.82	11.39	14.11	11.45	9.52	12.5	5.80
Cubic *mean*	10.85	10.56	16.00	11.15	8.34	7.36	6.62	6.50
std deviation	11.19	10.02	11.54	12.49	9.21	5.28	6.01	5.22
Lead								
FPS	0	0	0	0	0	0	0	0
RPS *mean*	5.35	1.59	0.17	0.07	0	0	0	0
std deviation	5.68	2.52	0.65	0.37	0	0	0	0
Near *mean*	5.06	3.43	3.11	2.96	2.76	1.99	3.06	2.67
std deviation	9.12	4.63	3.88	3.84	3.64	2.89	3.97	3.56
Cubic *mean*	6.53	6.38	4.68	2.73	1.97	2.12	2.06	2.09
std deviation	9.07	9.59	6.67	3.45	2.06	2.20	1.76	1.67

References

(1) Bartok, A. P.; Kondor, R.; G. Csanyi, G. On representing chemical environments. *Phys. Rev. B* **2013**, *87*, 184115.

(2) Himanen, L.; ; Jäger, M.; Morooka, E. V.; Canova, F. F.; Ranawat, Y. S.; Gao, D. Z.; Rinke, P.; Foster, A. S. DScribe: Library of descriptors for machine learning in materials science. *Computer Physics Communications* **2020**, *247*, 106949.

(3) DeSandip, A. P.; Bartok, G. C.; Ceriotti, M. Comparing molecules and solids across structural and alchemical space. *Phys. Chem. Chem. Phys.* **2016**, *18*, 13754.

(4) Deringer, V. L.; Bartók, A. P.; Bernstein, N.; Wilkins, D. M.; Ceriotti, M.; Csányi, G. Gaussian process regression for materials and molecules. *Chemical Reviews* **2021**, *121*, 10073–10141.

(5) Pihlajamäki, A.; Malola, S.; Kärkkäinen, T.; Häkkinen, H. Orientation adaptive minimal learning machine: application to thiolate-protected gold nanoclusters and gold-thiolate rings. *arXiv preprint arXiv:2203.09788* **2022**,

(6) Rossi, R. J. *Mathematical Statistics: An Introduction to Likelihood Based Inference*; New York: John Wiley and Sons, 2018; Chapter 5, p 227.

(7) Entekhabi, D.; Reichle, R. H.; Koster, R. D.; Crow, W. T. Performance Metrics for Soil Moisture Retrievals and Application Requirements. *Journal of Hydrometeorology* **2010**, *11*, 832 – 840.

(8) Kandaskalov, D.; Fournée, V.; Ledieu, J.; Gaudry, E. Adsorption Properties of the o-$AL_{13}Co_4$(100) Surface Towards Molecules Involved in the Semi-Hydrogenation of Acetylene. *J. Phys. Chem. C* **2014**, *118*, 23032–2304.