Supporting Information:
Reactivity of amorphous carbon surfaces: rationalizing the role of structural motifs in functionalization using machine learning

Miguel A. Caro,[1, 2, *] Anja Aarva,[1] Volker L. Deringer,[3, 4] Gábor Csányi,[3] and Tomi Laurila[1]

[1] *Department of Electrical Engineering and Automation,*
School of Electrical Engineering, Aalto University, Espoo, 02150, Finland
[2] *QTF Centre of Excellence, Department of Applied Physics, Aalto University, Espoo, 02150, Finland*
[3] *Engineering Laboratory, University of Cambridge, Cambridge CB2 1PZ, United Kingdom*
[4] *Department of Chemistry, University of Cambridge, Cambridge CB2 1EW, United Kingdom*

(Dated: August 30, 2018)

This document contains miscellaneous supporting information pertaining our manuscript "*Reactivity of amorphous carbon surfaces: rationalizing the role of structural motifs in functionalization using machine learning*".

[*] mcaroba@gmail.com

I. DETAILS OF THE CLUSTERING ALGORITHM

For our purpose of classification and comparison of atomic sites, we must rely on clustering algorithms that accept the use of a precomputed distance matrix or a similarity matrix (often termed an "affinity matrix" within the ML community). Algorithms that are suited for this purpose are for instance k-means, spectral clustering, or affinity propagation. All these algorithms were tested and were unfortunately found to not be well suited for our problem; one must note that not all motifs appear equally frequently in our a-C systems. For instance, C atoms bonded to only two and even only one neighbors appear less frequently than sp^2 and sp^3 carbons. Therefore, we need a clustering algorithm which can handle clusters of dissimilar sizes. Ideally, we would also wish to specify the number of target clusters to build or, at least, what number of clusters not to exceed. All the aforementioned algorithms failed to deliver this performance and as a matter of fact showed poor agreement between each other and yielded clusterings which did not resonate with chemical intuition. We settled for a variant of k-means: k-medoids. In essence, k-medoids allows us to provide our own metric of distance and relies on the use of medoids instead of centroids, which provide a better exemplary representation of our atomic environments. A medoid is an actual element of the data set (an atomic site in our case), whereas a centroid is a point in the hyper-space of atomic similarities that may or may not be close to any actual sample. Within the k-medoids method, each cluster is built around a medoid (sample) which possesses the minimal average distance from other samples within the same cluster. The procedure is extremely efficient; we used the Numpy-based implementation by Christian Bauckhage [1]. Unfortunately, k-medoids results are very sensitive to initial guesses for which samples to use as medoids. To alleviate this problem, we extended Bauckhage's code to accept more informed initialization than the original one based on random samples. Our modified version of the code is available from GitHub [2]. Briefly, our approach consists on requesting a certain number of initial medoids to be as isolated from other samples as possible, and randomizing the rest according to the total number of clusters to be built. In the present study we ran the algorithm 10 000 times, which should be more than sufficient given the number of elements in the data set. The predicted set of medoids which provide the best intra-cluster coherence is chosen as the final result. Here, we use two different ways to define "incoherence": total and relative. The total intra-cluster incoherence is given by

$$I_{\text{tot}} = \sum_k \sum_{i \in C_k} D_{i,M_k}, \tag{1}$$

where C_k is the kth cluster, i runs through all the samples contained within C_k, and M_k is the medoid of cluster C_k. Minimizing I_{tot} favors the proliferation of even-sized clusters, since it may become affordable to integrate a small number of isolated samples (which could constitute their own cluster attending to chemical intuition) into a larger distant cluster since they contribute little to the total incoherence. To overcome this issue, we also define the relative incoherence as

$$I_{\text{rel}} = \sum_k \frac{1}{n_k} \sum_{i \in C_k} D_{i,M_k}, \tag{2}$$

where n_k is the number of elements within cluster C_k. Minimization of I_{rel} may lead to proliferation of very small clusters with very high internal coherence; this problem can be easily solved with appropriate medoid initialization. In our case we used a mixture of maximally isolated and random medoid initialization.

All in all, we found that 2 Å SOAP cutoff, together with a maximum of 6 clusters and the use of the relative coherence criterion, provide the best recipe in terms of classifying atomic motifs in a-C in accordance with chemical intuition, as will be shown next.

In Fig. 1 we show the clusterings resulting from applying the different criteria discussed above. We study the representation provided by both total and relative coherence criteria, by requesting the data to be classified into 4, 6 and 8 clusters. On the plots, each data cluster is represented by a different color. For 4 clusters, both criteria lead to unintuitive grouping of sites (red ovals on the figure highlight the issue). This problem is still present for 6 clusters with the total coherence criterion. For 8 clusters, both schemes provide an intuitive representation although at the cost of increased complexity. We find that the relative coherence criterion together with a target 6 clusters provides the best trade-off between complexity (in the sense of minimizing the cluster number) and fulfillment of the chemical intuition requirement.

II. COORDINATES OF THE MEDOIDS

The coordinates of the medoids identified in our manuscript are given in Listing 1, in regular XYZ format (Cartesian coordinates in Å). The central site of the motif is always given centered at (2,2,2).

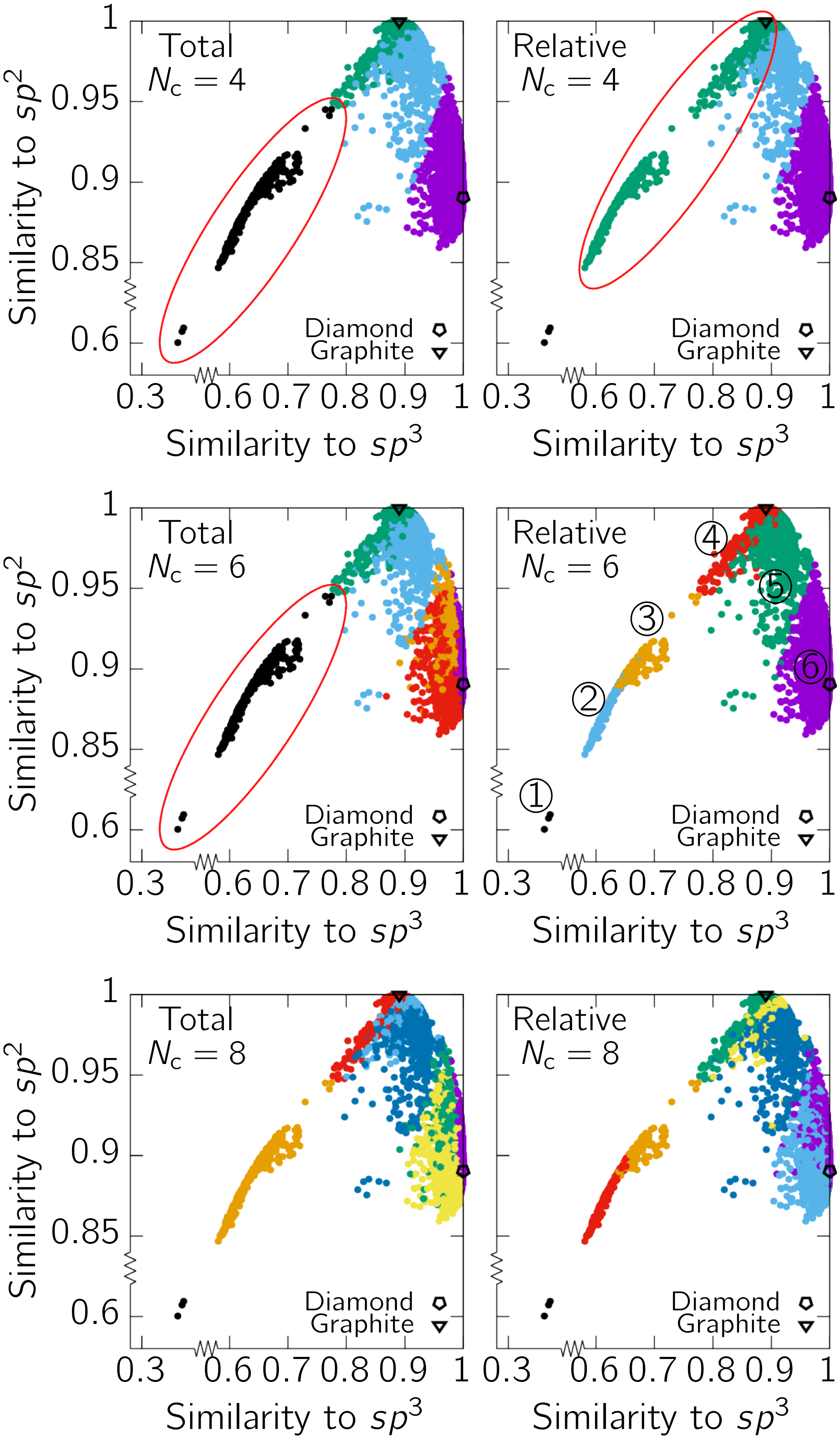

FIG. 1. Results of the clustering analysis with different numbers of target clusters. Atomic sites which belong to the same cluster are represented with same-colored dots. Red ellipses indicate issues related to the algorithm clustering together sites which are too different from each other.

Listing 1. Coordinates of the medoids identified in our manuscript.

```
# Cluster 1:
2
 Lattice=”4.0 0.0 0.0 0.0 4.0 0.0 0.0 0.0 4.0” pbc=”F F F”
C     2.00000000     2.00000000     2.00000000     6
C     2.29687663     2.67777026     0.94624125     6


# Cluster 2:
3
 Lattice=”4.0 0.0 0.0 0.0 4.0 0.0 0.0 0.0 4.0” pbc=”F F F”
C     2.00000000     2.00000000     2.00000000     6
C     0.85233974     2.50752321     1.89310963     6
C     3.03752697     1.39706709     2.69247182     6


# Cluster 3:
3
 Lattice=”4.0 0.0 0.0 0.0 4.0 0.0 0.0 0.0 4.0” pbc=”F F F”
C     2.00000000     2.00000000     2.00000000     6
C     1.36714570     2.76202438     3.01283000     6
C     2.18154943     0.74383237     1.89620746     6


# Cluster 4:
4
 Lattice=”4.0 0.0 0.0 0.0 4.0 0.0 0.0 0.0 4.0” pbc=”F F F”
C     2.00000000     2.00000000     2.00000000     6
C     0.76259816     1.58637350     2.47140027     6
C     3.19947508     1.82510102     2.68586940     6
C     1.72449741     3.11424993     1.02365541     6


# Cluster 5:
4
 Lattice=”4.0 0.0 0.0 0.0 4.0 0.0 0.0 0.0 4.0” pbc=”F F F”
C     2.00000000     2.00000000     2.00000000     6
C     2.55238270     2.72733814     0.77848972     6
C     1.55336964     2.41922688     3.23944008     6
C     2.48320219     0.58494414     2.11089562     6


# Cluster 6:
5
 Lattice=”4.0 0.0 0.0 0.0 4.0 0.0 0.0 0.0 4.0” pbc=”F F F”
C     2.00000000     2.00000000     2.00000000     C
C     0.78845475     1.04857888     2.53432665     6
C     2.75631108     1.55215021     3.30558944     6
C     2.86585723     1.34108586     0.96320102     6
C     1.88348724     3.47487605     1.74732444     6
```

III. NON-PLANARITY/NON-LINEARITY OF MOTIFS

In Fig. 2 we show motif non-linearity and non-planarity h for *sp* and sp^2 sites, respectively. The non-linearity and non-planarity h are defined as the orthogonal distances between the central site and the line (*sp*) or plane (sp^2) defined by its neighbors, respectively.

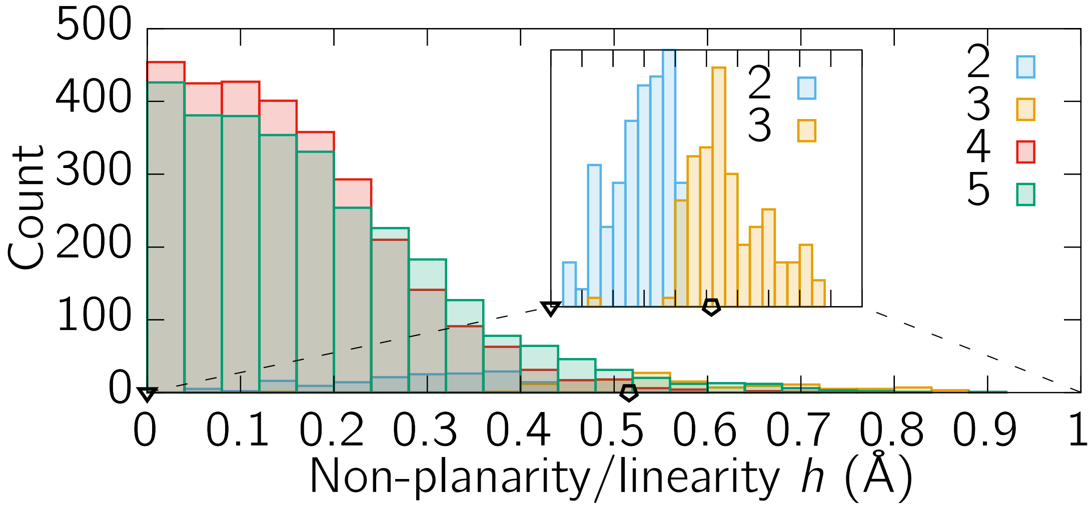

FIG. 2. Motif non-planarity (sp^2) and non-linearity (sp).

IV. INTEGRATED LOCAL DENSITY OF STATES

The local density of states (LDOS), averaged over all sites belonging to the same cluster, is shown in Fig. 3. The interval from -3 to 3 eV was used to integrate the LDOS. In the paper we show that this integrated value performs well as descriptor for adsorption characteristics.

V. KERNEL OPTIMIZATION FOR A GAP ML MODEL OF ADSORPTION ENERGY

The kernels used to measure the degree of similarity between two atomic sites rely on several parameters that can strongly influence the kernel performance. For instance, the parameter used to "smear" the atomic density within the SOAP approach [3], σ_{atom}, can make a SOAP descriptor "sharp" or "fuzzy". Sharp descriptors are good at interpolating the properties of atomic sites which resemble very closely the training configurations, but will be bad at interpolating (and, obviously, extrapolating) far away. The following is a list of all of the parameters which affect kernel performance (where we have excluded the spherical harmonics parameters for the SOAP expansion):

- σ_{atom}, "atom sigma" [units: Å]. It controls the smearing of the atomic density in the generation of the SOAP representation of the atomic density.
- r_{c}, SOAP cutoff radius [units: Å]. It defines the region around an atomic site within which SOAP can "see" the atomic neighborhood. The information about the environment outside the cutoff sphere is lost.
- σ, the parameter used for Tikhonov regularization when constructing a GAP model [4] [unitless]. It represents noise in the data.
- N_{t}, number of samples in the training set. Usually, the more the better, although many models will not improve beyond a certain number of training configurations.
- ζ, SOAP kernel exponent [unitless]. This number, typically equal to 4, is used to make the SOAP kernel more or less sharp. The higher the number the sharper the SOAP kernel.
- σ_n, smearing parameter of the LDOS moment Gaussian kernel [units: electrons × eV^n]. We use $n_{\text{max}} + 1$ of these parameters, where n_{max} depends on the definition of the LDOS kernel [see Eq. (7) of main manuscript].

In Fig. 4 we show the Monte Carlo results of parameter optimization. The RMSE of each model (that is, of each combination of parameters) is one dot on the graph. The lowest lying (lowest error) models define the convex hull, and delimit the model accuracy that can be achieved with the constraints imposed by constructing the kernel (and other aspects of the GAP model) in some particular way. The best model is given by the combination of parameters which yield the lowest error. From the graph one can see how we first optimized the SOAP parameters and coarse-grained the LDOS parameters, followed by fine tuning of optimal SOAP+LDOS parameters.

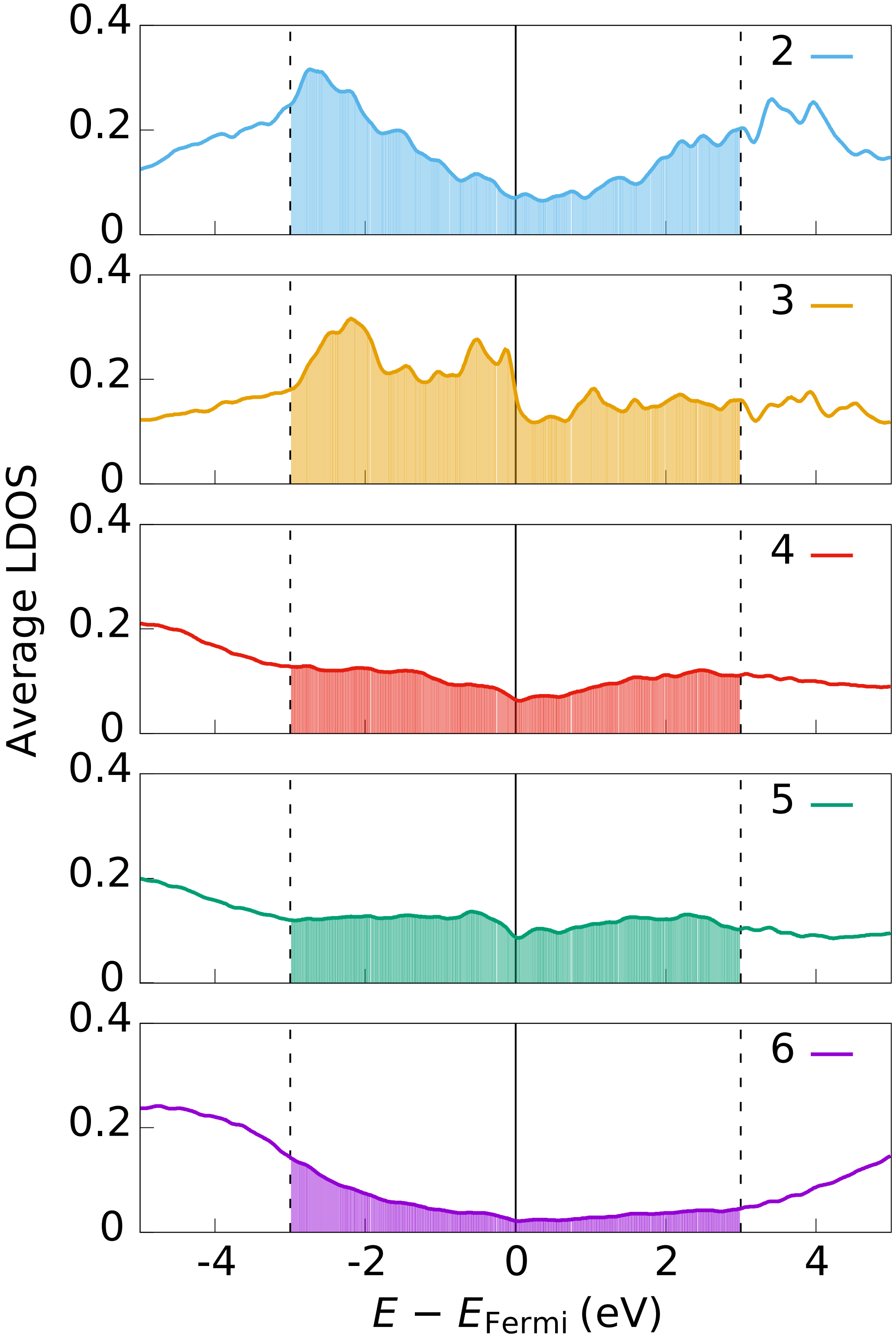

FIG. 3. Average local density of states (LDOS) of each cluster. Vertical dashed lines delimit the integration interval. Cluster 1 is excluded from the plot due to poor sampling.

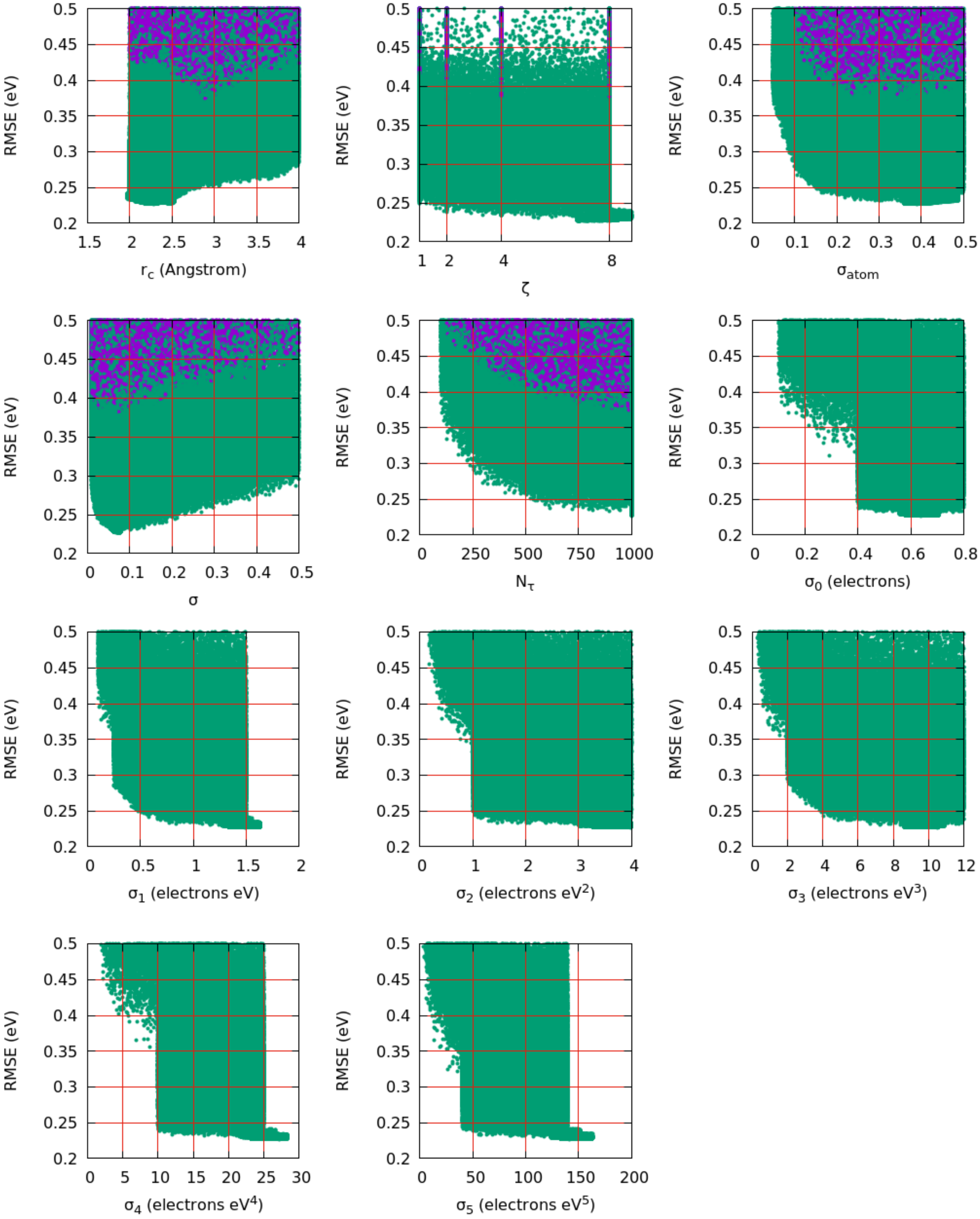

FIG. 4. Performance of different GAP ML models for adsorption energy prediction which can be built for the H-probe data. N_t samples drawn from half of the data are used for training and all the samples in the other half are used for testing. SOAP-only models are purple data points and SOAP+LDOS models are green data points.

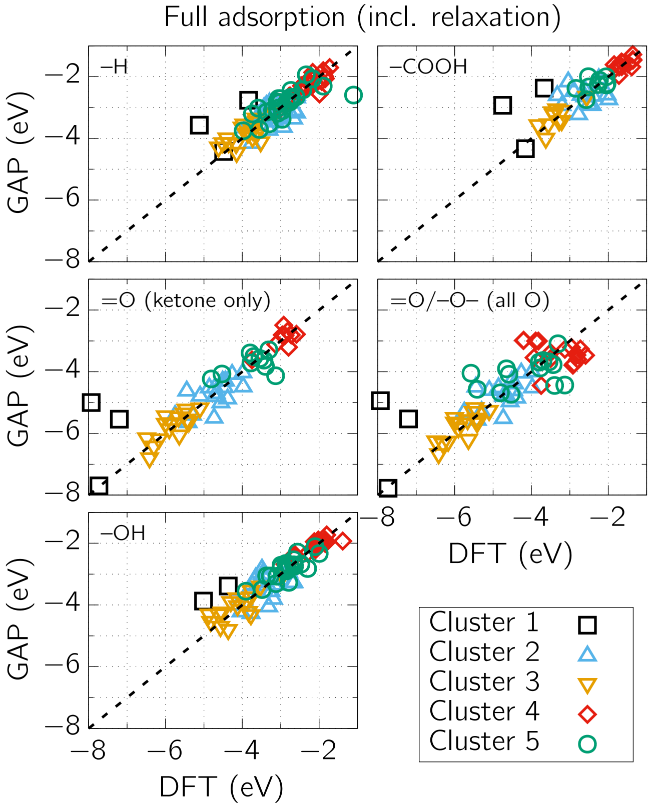

FIG. 5. SOAP+LDOS GAP models for adsorption energy prediction on a-C surface sites, including cluster 1 sites.

TABLE I. Performance (error estimates) of the GAP ML models for adsorption of different functional groups on a-C surface atomic motifs.

	All motifs		Excl. cluster 1	
	MAE (meV)	RMSE (meV)	MAE (meV)	RMSE (meV)
–H	248	364	227	313
–COOH	295	443	243	316
=O	329	558	261	338
=O/–O–	468	686	417	556
–OH	261	345	239	303

VI. EFFECT OF UNDERCOORDINATED ATOMS

The one-fold coordinated atoms in cluster 1 are highly unstable structural defects (3 out of 10 800 sites) in our a-C surfaces. It is quite difficult to train a ML model which can predict accurate adsorption energies for these sites and all other sites *simultaneously*. Removing these sites the errors are reduced considerably. Training a model which can predict the energies of those defective sites correctly would require more data in that particular region of feature space. To illustrate this issue, in Fig. 5 and Table I we show the predictions of our best SOAP+LDOS ML model including cluster 1 sites.

[1] C. Bauckhage, *Numpy/scipy Recipes for Data Science: k-Medoids Clustering*, Tech. Rep. (University of Bonn, 2015).

[2] M. A. Caro. Fork of C. Bauckhage's k-Medoids Python implementation for enhanced medoid initialization. http://github.com/mcaroba/kmedoids (accessed August 30, 2018).

[3] A. P. Bartók, R. Kondor, and G. Csányi, "On representing chemical environments," Phys. Rev. B **87**, 184115 (2013).

[4] A.P. Bartók and G. Csányi, "Gaussian approximation potentials: A brief tutorial introduction," Int. J. Quantum Chem. **115**, 1051–1057 (2015).