Electronic Supplementary Material (ESI) for Physical Chemistry Chemical Physics.

Supporting Information

for

Machine learning and DFT investigation of CO, CO_2 and CH_4 adsorption on pristine and defective two-dimensional magnesene

Siby Thomas,[1*] Felix Mayr,[1] Ajith Kulangara Madam[2] and Alessio Gagliardi[1*]

[1]*School of Computation, Information and Technology (SoCIT), Technical University of Munich (TUM), Hans-Piloty-Strasse 1, 85748 Garching, Munich, Germany.*

[2]*Department of Physics, National Institute of Technology Karnataka (NITK), Surathkal, PO: Srinivasnagar - 575025, Mangalore, Karnataka, India*

Email: siby.thomas@tum.de (Siby Thomas); Alessio Gagliardi (alessio.gagliardi@tum.de)

1. Machine learning setup

For training the machine learning (ML) models, we employed the sklearn v1.1.3[1] implementations of Gradient-Boosting, Random-Forest and Kernel-Ridge-Regression. In addition, we used Gradient Boosting as implemented in xgboost v1.7.3.[2] Models were trained on the task of predicting the adsorption energy based on the initial system. For every modeling option, five different 80/20-train-test-splits were chosen (stratified by adsorbents, as well as selecting single adsorbents) and models were built on the training set with 5-fold cross-validation (nested cross-validation). Tabular features were created with JARVIS/matminer[3] and scaled with sklearn's StandardScaler. For SOAP-featurization we used the implementation in the DScribe-library v1.2.2.[4] To featurize the system, we created a global fingerprint for the 2D-Mg and a local fingerprint around the adsorbent C-atom, with a dummy-atom placed at the surface-origin of a surface-normal running through the C-atom. Both were concatenated and fed to the model. Consult our code shared at https://github.com/ThomasSiby/Gas_adsorption_on_2D_materials for more details.

2. Machine learning metrics

To assess the performance of the ML models, different loss functions and metrics such as mean absolute error (MAE) and coefficient of determination (R^2) are used and are evaluated using the relation:

$$MAE = \frac{1}{N}\sum_{i=1}^{N}\left| Y_i - \hat{Y}\right| \tag{S1}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(Y_i - \hat{Y})^2}{\sum_{i=1}^{N}(Y_i - \bar{Y})^2} \tag{S2}$$

where Y_i, $\hat{Y}$ and $\bar{Y}$ are actual, predicted, and mean values of the target feature, and i is the data point at any given instance, and N is the total number of data points.

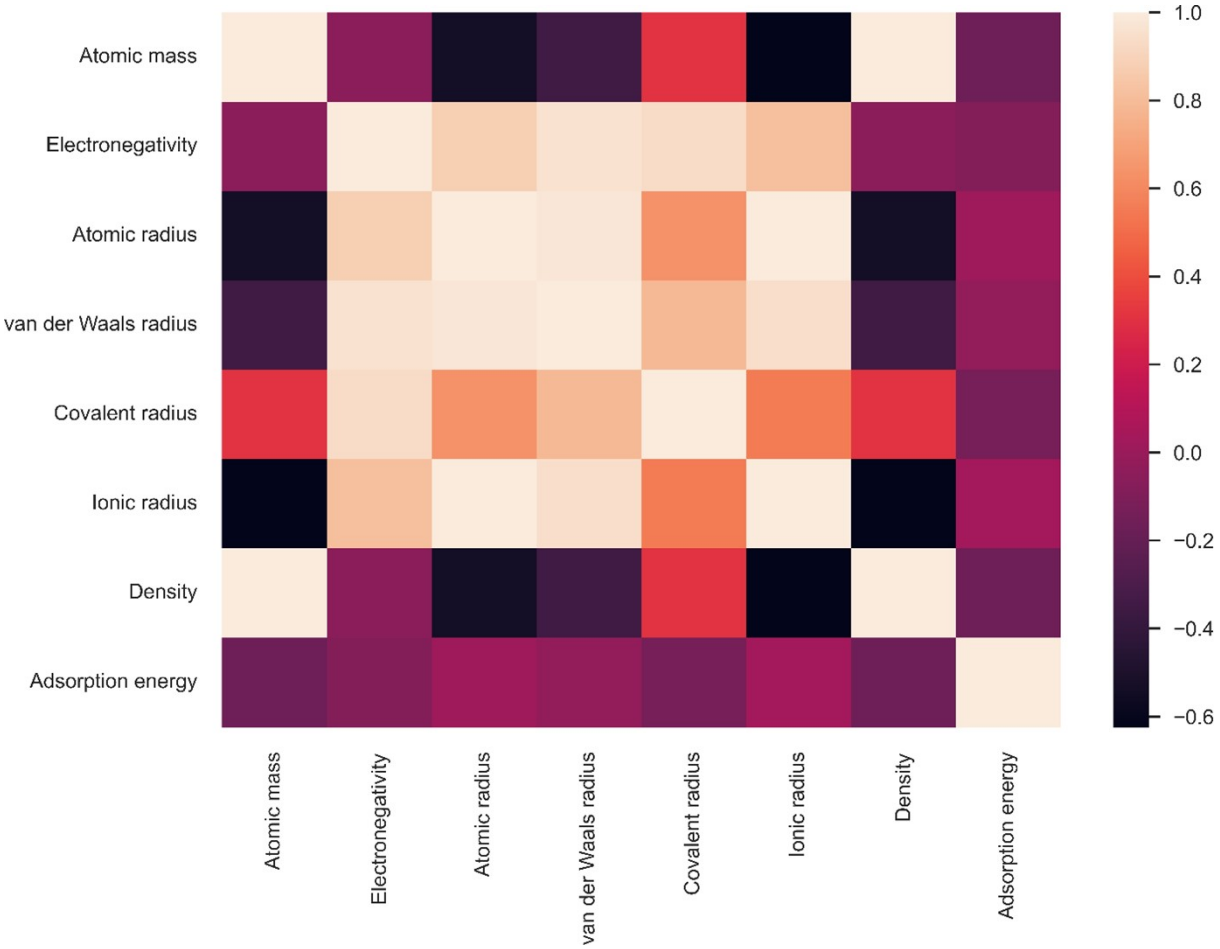

Fig. S1. Pearson's correlation heatmap for the final reduced set of descriptors. The descriptors represent the properties of host 2D-Mg and guest gas molecules. The strength of the correlation is analyzed using the color code associated with the correlation matrix. Here, ivory and black denote high and low correlations between the descriptors.

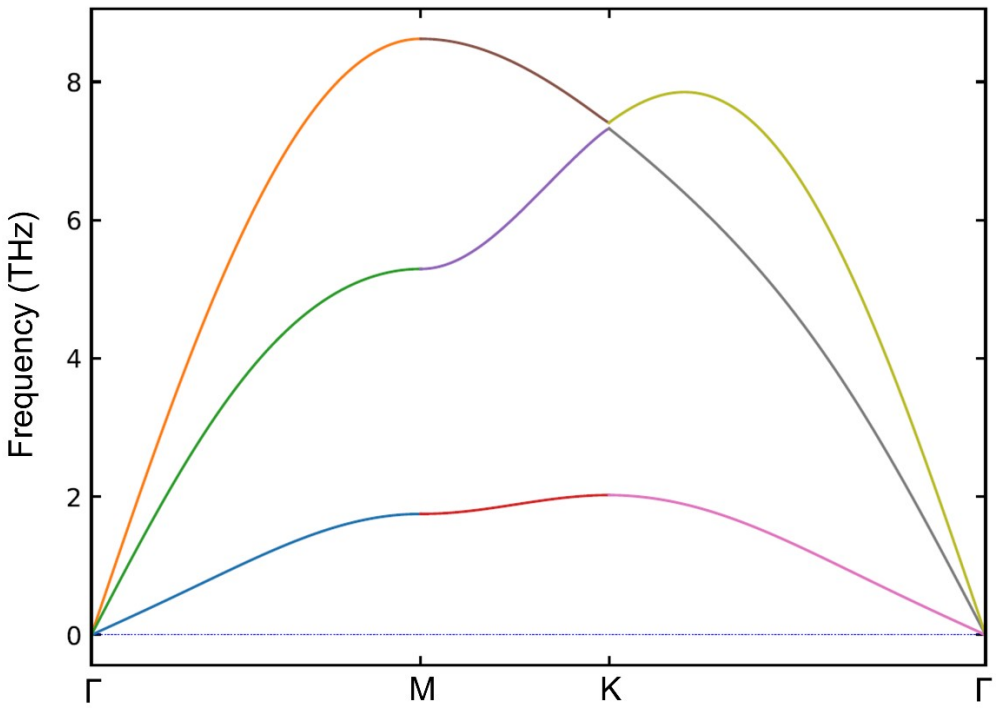

Fig. S2. Phonon dispersion relation of 2D-Mg obtained using the finite displacement method indicating the dynamical stability in the absence of imaginary modes.

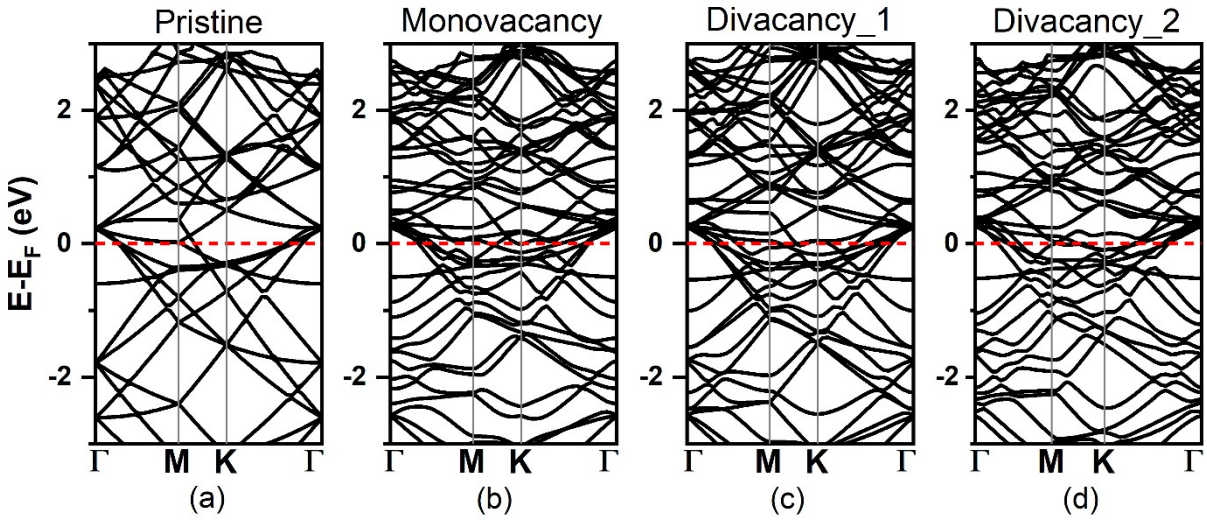

Fig. S3. The electronic band structure of (a) PR, (b) MV, (c) DV1, and (d) DV2 structures along the high-symmetric path Γ-M-K-Γ using PBE functional. The horizontal red dashed line represents the Fermi level and is set to 0 eV.

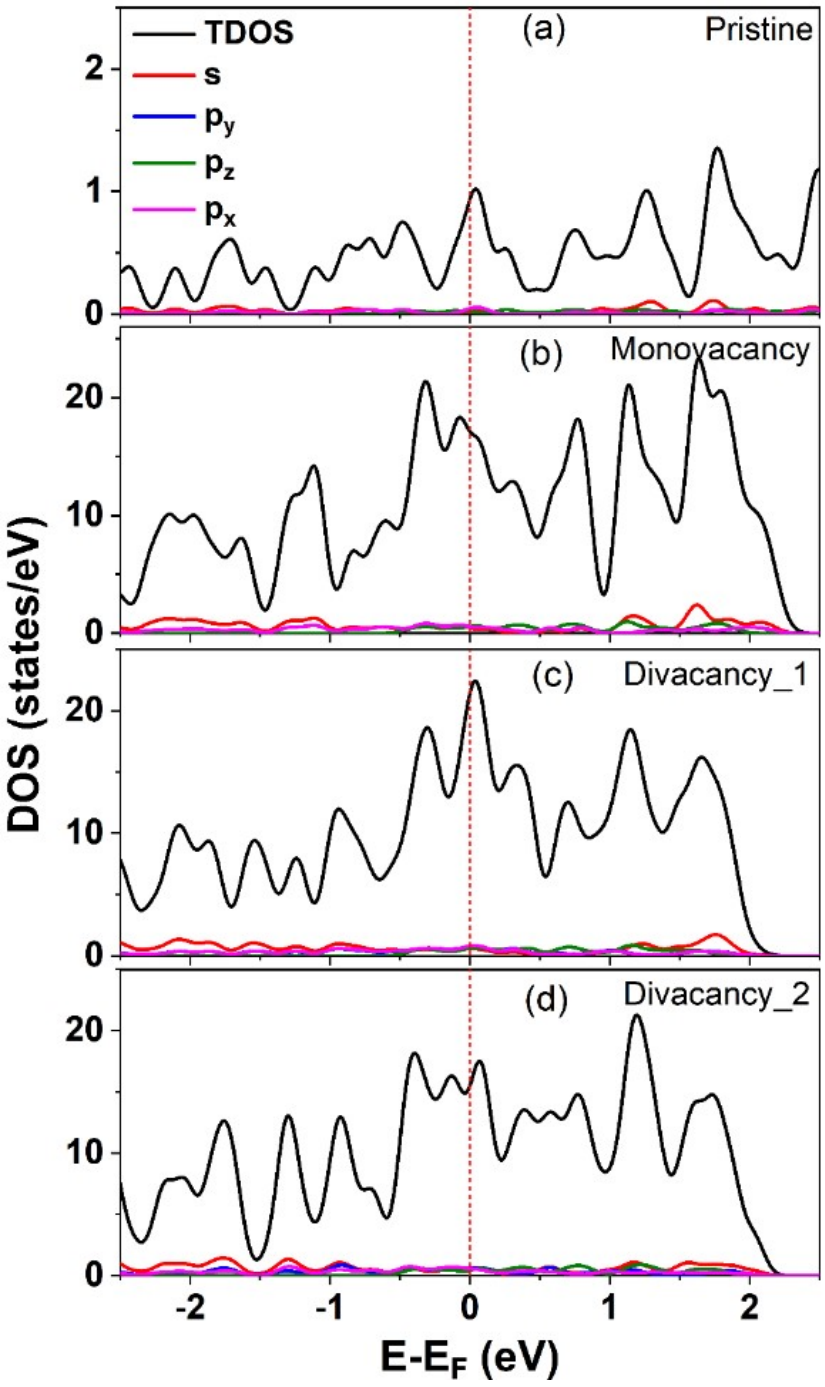

Fig. S4. The total and projected density of states (DOS) of (a) PR, (b) MV, (c) DV1, and (d) DV2 structures were computed using the HSE06 functional. The red dashed vertical line stands for the Fermi level, which is fixed at 0 eV.

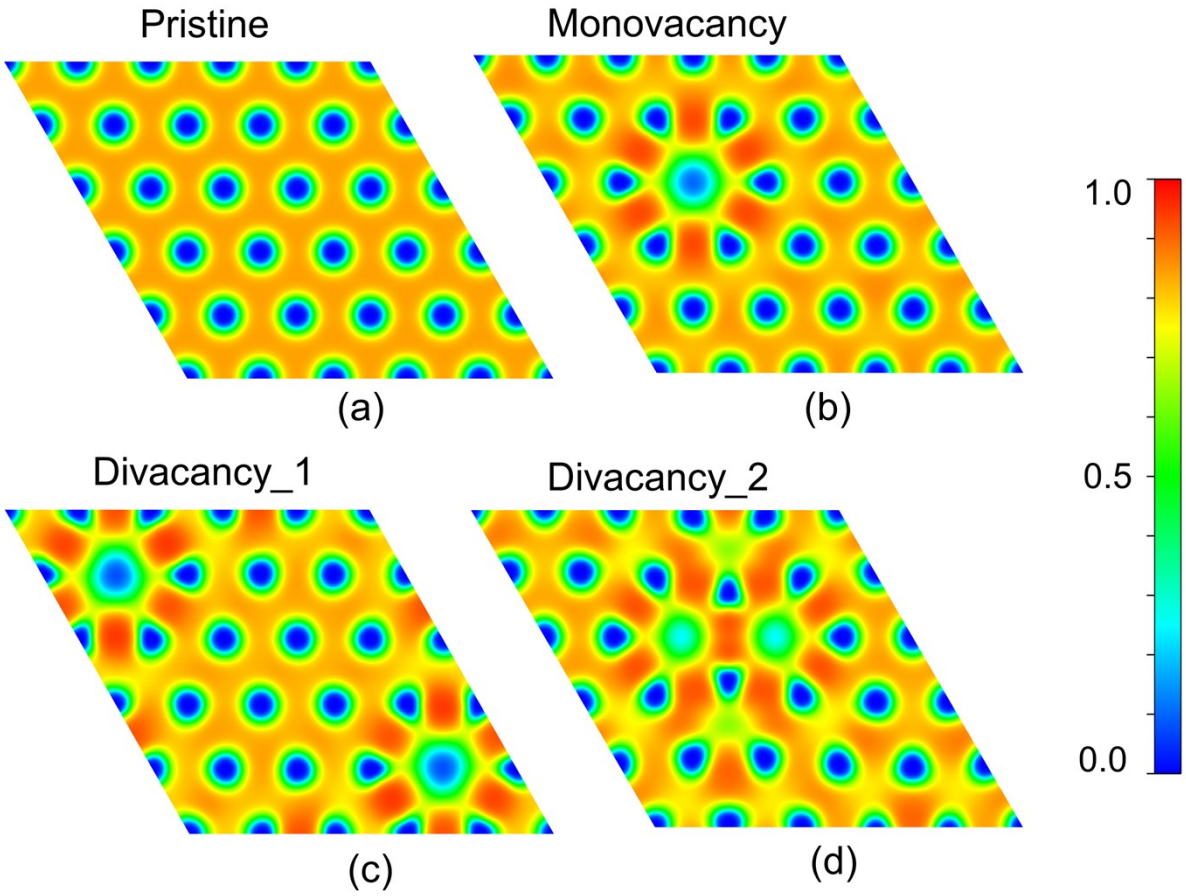

Fig. S5. The electron localization function (ELF) maps of (a) PR, (b) MV, (c) DV1, and (d) DV2 structures with a slice crossing the structural plane. Red in the scaling bar denotes that electrons are highly localized, and blue indicates that electrons are without any localization.

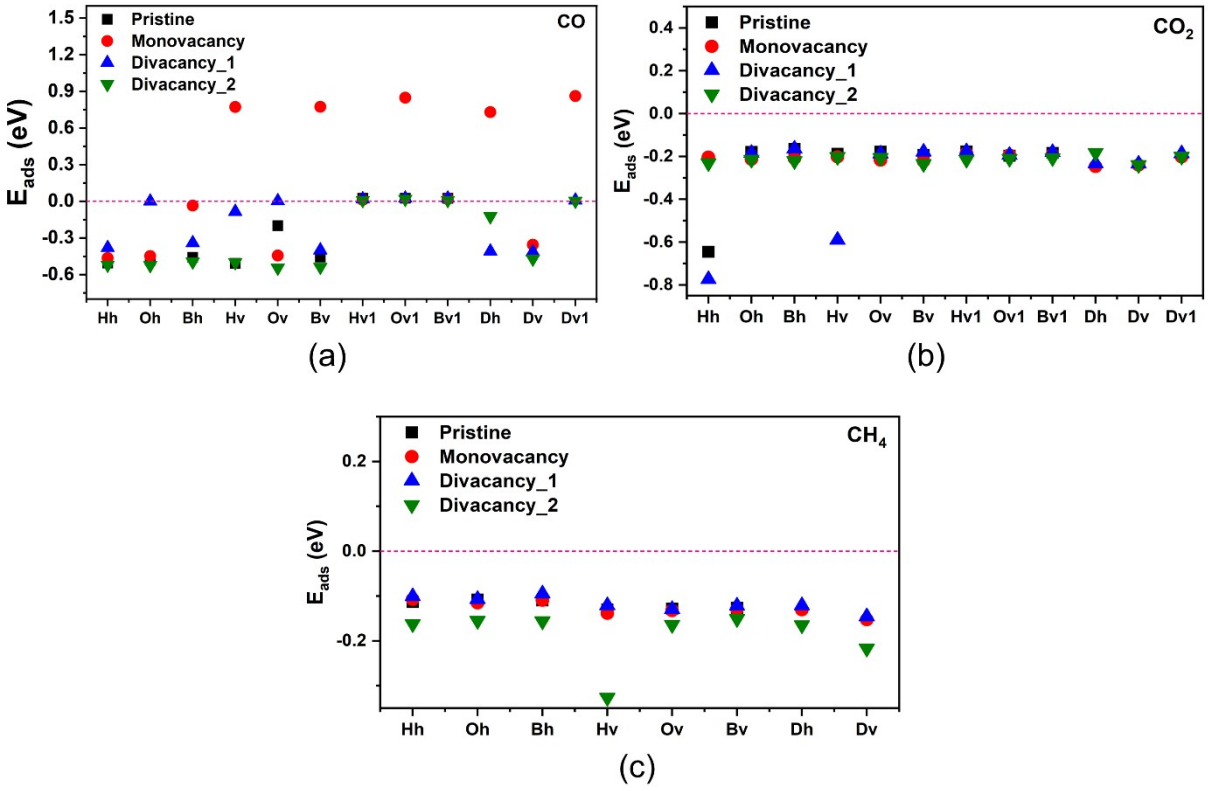

Fig. S6. Computed adsorption energy of CO, CO_2 and CH_4 gas molecules placed at different locations of pristine and defective 2D-Mg.

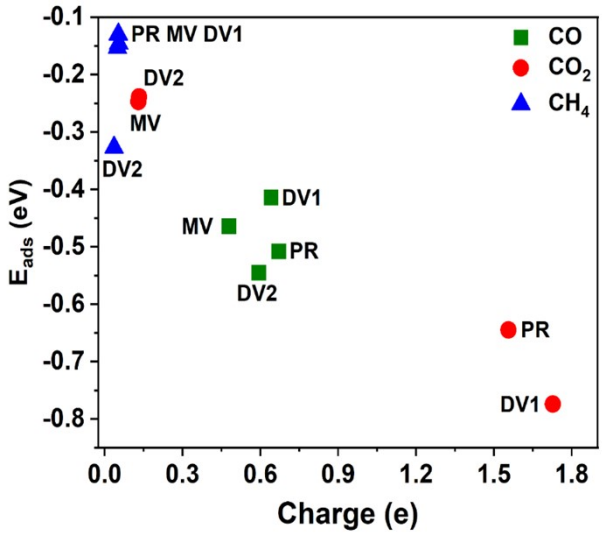

Fig. S7. Adsorption energy versus charge analysis for pristine and defective 2D-Mg with the presence of CO, CO_2 and CH_4 gas molecules.

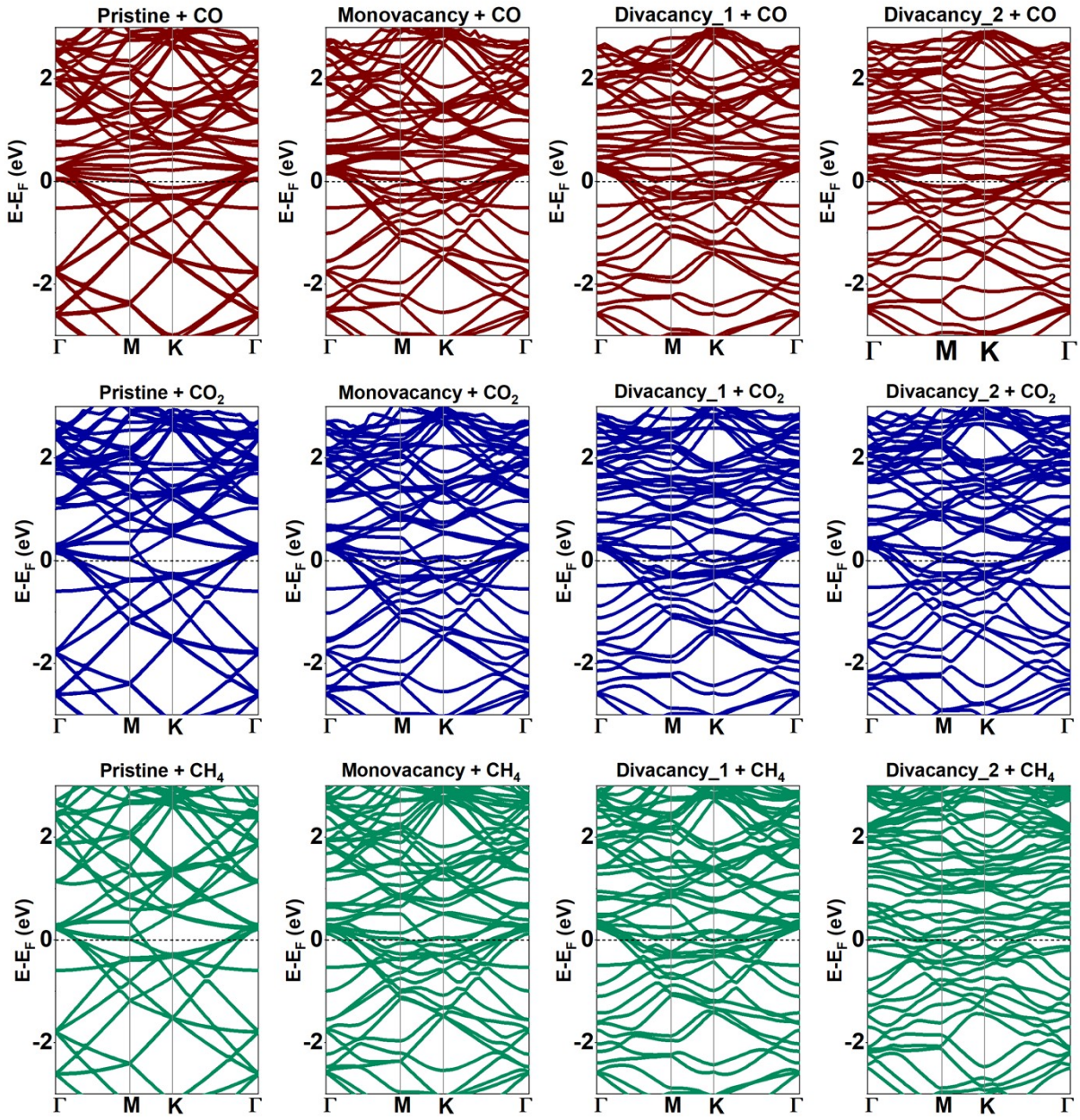

Fig. S8. Calculated electronic band structures of pristine and defective 2D-Mg when CO, CO_2 and CH_4 gas molecules are adsorbed on the surface. The band structures are computed using the PBC functional and the horizontal black dashed line represents the Fermi level.

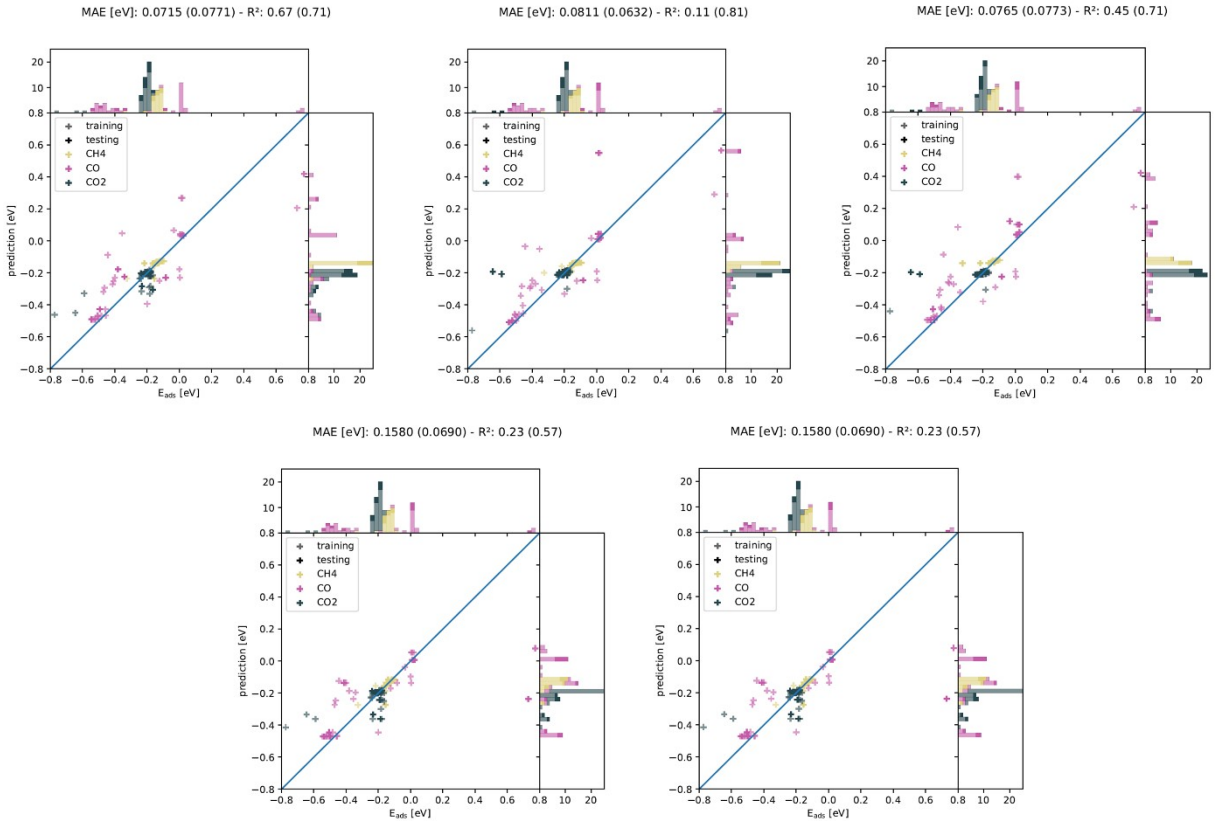

Fig. S9. Parity plots for cross-validated models with different train-test-splits, as run to get a perception of the actual, expected prediction performance (see Table 5, nested cross-validation in the Manuscript). The stated values above the plot show scores for both the testing set as well as the training set in parentheses. The given images show five different Random Forest models using SOAP featurization.

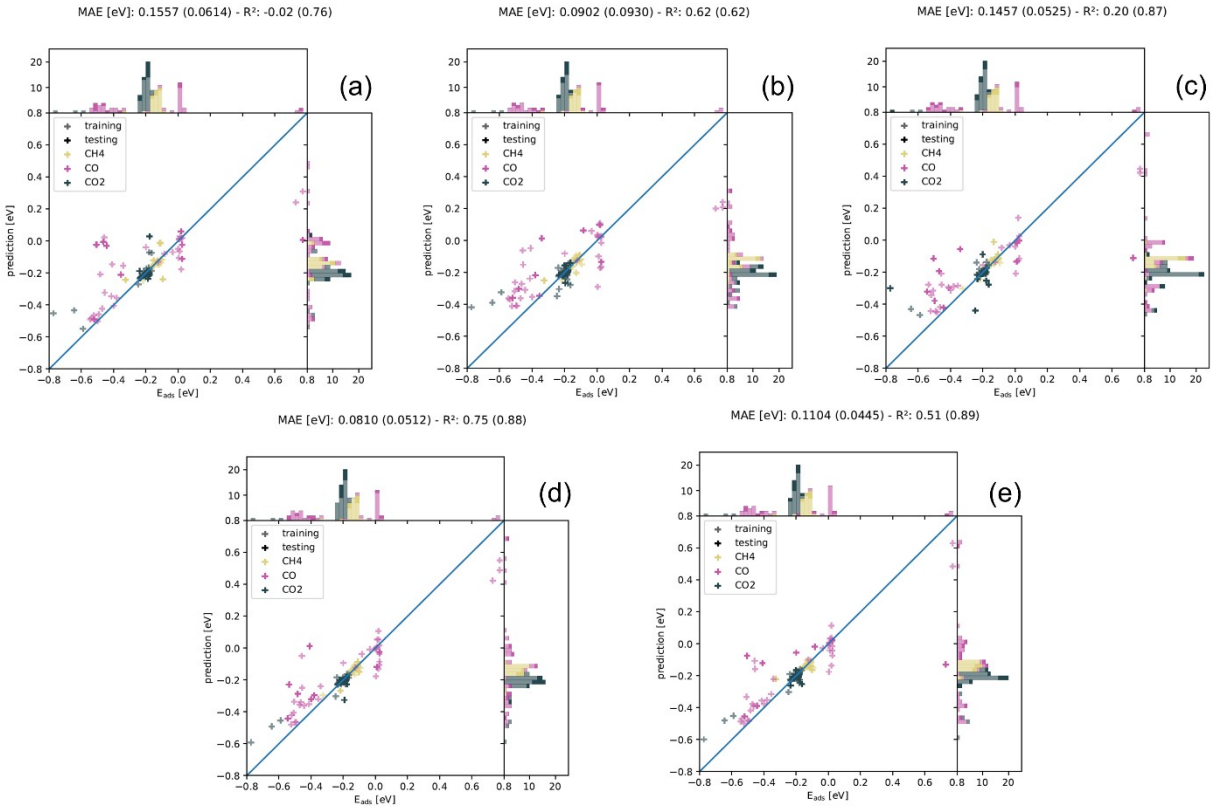

Fig. S10. Parity plots for cross-validated models with different train-test-splits, as run to get a perception of the actual, expected prediction performance (see Table 5, nested cross-validation in the Manuscript). The stated values above show scores for both the testing set as well as the training set in parentheses. In addition to the parity plot, distributions of true and predicted values are shown color-coded by the adsorbent species. The given images show five different Random Forest models using the statistical features.

Notes and References

1 F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot and É. Duchesnay, *Journal of Machine Learning Research*, 2011, **12**, 2825–2830.

2 T. Chen and C. Guestrin, in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Association for Computing Machinery, New York, NY, USA, 2016, pp. 785–794.

3 K. Choudhary, K. F. Garrity, A. C. E. Reid, B. DeCost, A. J. Biacchi, A. R. Hight Walker, Z. Trautt, J. Hattrick-Simpers, A. G. Kusne, A. Centrone, A. Davydov, J. Jiang, R. Pachter, G. Cheon, E. Reed, A. Agrawal, X. Qian, V. Sharma, H. Zhuang, S. V. Kalinin, B. G. Sumpter, G. Pilania, P. Acar, S. Mandal, K. Haule, D. Vanderbilt, K. Rabe and F. Tavazza, *npj Comput Mater*, 2020, **6**, 1–13.

4 L. Himanen, M. O. J. Jäger, E. V. Morooka, F. Federici Canova, Y. S. Ranawat, D. Z. Gao, P. Rinke and A. S. Foster, *Computer Physics Communications*, 2020, **247**, 106949.