Machine Learning Prediction of Defect Formation Energies in a-SiO_2

Diego Milardovich[1], Markus Jech[1], Dominic Waldhoer[1,2], Michael Waltl[1,2], and Tibor Grasser[1]

[1]Institute for Microelectronics, Technische Universität Wien,
Gußhausstraße 27–29, 1040 Vienna, Austria
[2]Christian Doppler Laboratory for Single-Defect Spectroscopy in Semiconductor Devices,
Gußhausstraße 27–29, 1040 Vienna, Austria
E-mail: [milardovich | jech | waldhoer | waltl | grasser]@iue.tuwien.ac.at

***Abstract*—Due to its stochastic nature, the calculation of defect formation energies in amorphous structures is a CPU-intensive task. We demonstrate the use of machine learning to predict defect formation energies to significantly minimize the number of required calculations. Different combinations of *descriptors* and machine learning algorithms are used to predict the formation energies of hydroxyl E' center defects in amorphous silicon dioxide structures. The performance of each combination is analyzed and compared to results obtained from direct *ab initio* calculations.**

I. Introduction

Ab initio methods, particularly those based on density functional theory (DFT), are routinely used in material science to calculate electronic and structural properties. However, these methods have the disadvantage of being highly computationally expensive. This disadvantage limits the use of ab initio methods to small systems (on the order of few hundred atoms) and short time scales (on the order of tens of ps).

The design of modern electronic devices heavily relies on reliability considerations of new fabrication processes and emerging materials (e.g. 2D materials). In this regard, DFT studies can provide valuable information on underlying defects and physical mechanisms impacting the device reliability. Even so, the high computational costs often prohibit a direct employment of DFT within process or device simulations, particularly for amorphous structures like gate dielectrics, where statistical data has to be gathered.

A promising solution is offered by using machine learning (ML) to reduce the computational demands required to study certain aspects of device reliability. In this work, we study the possibility of using ML models to calculate the formation energies of hydroxyl E' center defects in amorphous silicon dioxide (a-SiO_2) structures [1]. Studying the behavior of these defects, especially their formation during device processing, is of great importance for the development of modern microelectronics, since they are suspected to be responsible for bias temperature instability (BTI) and random telegraph noise (RTN) in MOS transistors [2–4].

Since the formation of hydroxyl E' centers depends on the availability of hydrogen, the concentration of these defects is not directly accessible in DFT or ML models. However, it could be derived from a kinetic Monte Carlo (KMC) process model [5], coupled with ML-based on-the-fly prediction of formation energies as a function of the local environment.

II. Methodology

In order to build a ML model which is able to predict certain electronic properties, it is first necessary to represent the atomistic structure in a way which is compatible with the selected ML algorithm. Such mathematical representation of the structure is called a *descriptor*. The conceptual workflow from structures to the final predictions is shown in Fig. 1.

In this study, we prepared 16 different a-SiO_2 structures, each with a total of 216 atoms. These structures were created with LAMMPS [6] by using the ReaxFF [7] force field and the melt-quench technique, as described in [8]. An example of these structures is shown in Fig. 2. Within these structures, 1271 hydroxyl E' centers were created, as shown in Fig. 3. Their formation energies were extracted using DFT; the calculations were performed using the PBE functional [9] in the CP2K software package [10]. The resulting distribution of formation energies can be seen in Fig. 4.

Three popular ML models were used to predict the formation energy of the hydroxyl E' centers: neural network (NN), kernel ridge regression (KRR) and decision tree (DT), as implemented in the scikit-learn package [11]. These models were combined with 2 local descriptors common in the literature: *smooth overlap of atomic positions* (SOAP) [12] and *atom-centered symmetry functions* (ACSF) [13], implemented in the Python package DScribe [14].

It is clear from Fig. 5 that there is an inverse correlation between the formation energy of an hydroxyl E' center defect and the length of the Si-O bond which has to be broken in order to form the defect. In other words, the formation energy is influenced by directly accessible geometric quantities (e.g., bond-lengths and bond-angles). Based on this observation, we propose a simple geometry-based descriptor: *bond-lengths and bond-angles* (BLBA).

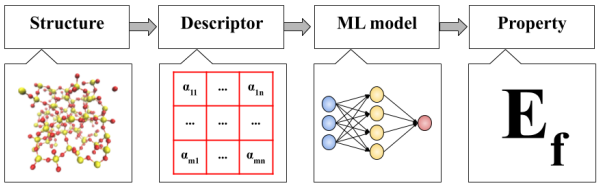


Fig. 1: Workflow to predict a property from an atomistic structure. The structure must be represented by a descriptor. Then, a ML model can be trained and subsequently be used to predict the desired property in new atomistic structures.

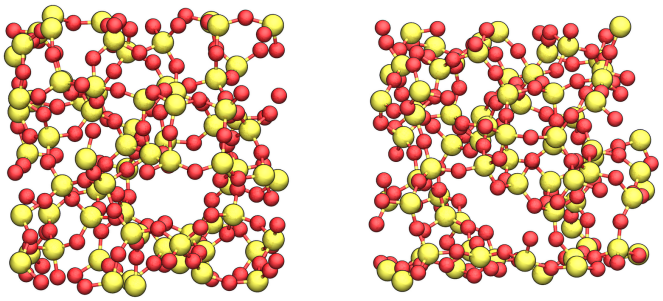
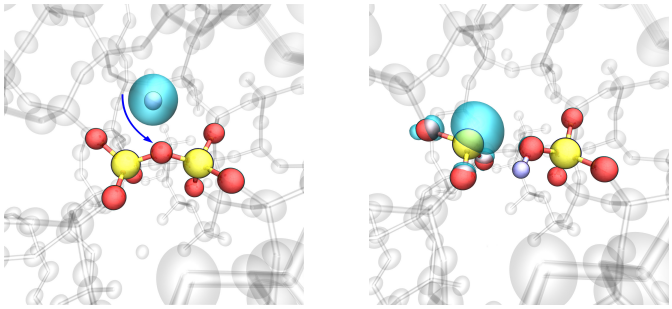


Fig. 2: Example of two amorphous a-SiO_2 structures used in this work to train and test the ML models. Each structure contains 216 atoms and a total of 16 such structures are used in this work.



(a) (b)

Fig. 3: Hydrogen interaction with the a-SiO_2 matrix (a) can lead to formation of a hydroxyl group, as well as breakage of a strained Si-O bond, resulting in a hydroxyl E' center defect (b). The blue bubbles show the spin-density associated with interstitial hydrogen and the hydroxyl E' center defect, respectively.

Our BLBA descriptor is constructed according to the following steps:

1) Create the neighbor list of N^{th} order. The starting point of this list is the Oxygen atom for which we would like to predict the formation energy of an hydroxyl E' center defect. The first step is to list the atoms bonded to this Oxygen atom. Then, every step consists in listing the atoms bonded to the atoms found in the previous step, until a maximum number of iterations, N, is reached (in our study, N = 3).

2) Extract the bond-lengths for every pair of bonded atoms in the neighbor list created in step 1. Use these values to create a bond-lengths vector. The bond lengths within one iteration are ordered in descending order. Note that the values from different iteration steps are kept separated to retain information about the distance to the defect site.

3) Extract the bond-angles between triplets in the neighbor list created in step 1. Use these values to create a bond-angles vector. As for the bond-lengths, the angles are arranged in descending order.

4) The BLBA descriptor is created by merging the bond-lengths vector and bond-angles vector, in this order. Therefore, this descriptor is a vector which consists of the properly ordered bond-lengths and bond-angles of the neighbourhood of the possible defect site.

Although our BLBA descriptor cannot describe the local environment in the same level of detail of SOAP and ACSF descriptors, it displays very valuable properties: since it contains only physical values which are relative to the possible defect site, our descriptor is invariant to spatial translations and rotations of the coordinate system. Moreover, since its components are ordered, it is also invariant with respect to permutation of indexes. Finally, it is highly compact, i.e. it contains sufficient information to be used for the prediction of the formation energies of hydroxyl E' center defects, while keeping its size and complexity to a minimum.
The performance of our BLBA descriptor will also be considered, as a demonstration of the potential of simpler geometry-based descriptors. The 1271 hydroxyl E' center defects were randomly divided into a training and a testing data set, with a ratio of 4:1. Every permutation of descriptor and ML model was trained with the training set and used to predict the values for the testing set. In every case, the mean absolute error (MAE) was calculated between the predictions and the targets.

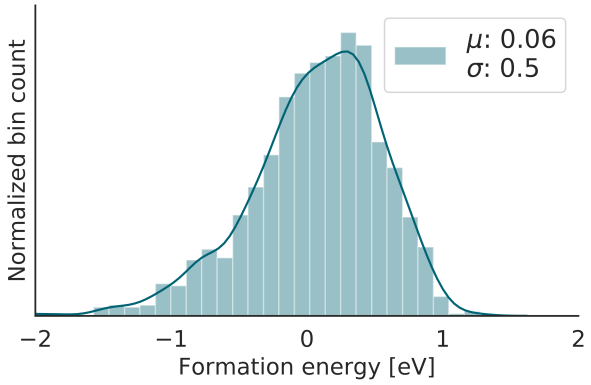


Fig. 4: Formation energies of hydroxyl E' center defects in a-SiO_2, obtained with DFT calculations, together with the mean (μ) and standard deviation (σ). The broad distribution is due to the amorphous nature of the structures.

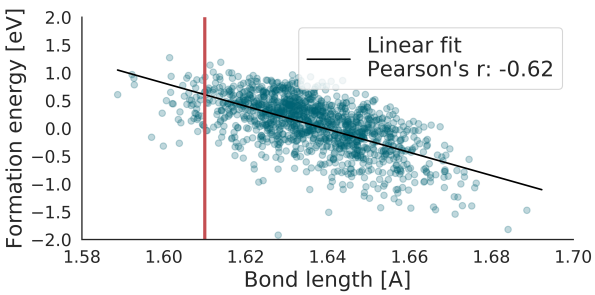


Fig. 5: Correlation between formation energies and bond-lengths. The formation energy decreases for larger bond lengths, so defects preferably form at strained Si-O bonds. The red line indicates the Si-O bond-length in alphaquartz [15].

III. Results and Discussion

The results of this study are summarized in Fig. 6, where the prediction errors for the formation energies in the testing set are presented, together with its MAE, for every combination of descriptor (SOAP, ACSF and BLBA) and ML model (NN, KRR and DT) studied in this work.
Ultimately, the worst results were obtained with the ACSF descriptor in conjunction with a DT model, which presents a MAE of $0.39\,\mathrm{eV}$. On the other hand, the best results were obtained with the SOAP descriptor in conjunction with a NN model, which presents a MAE of $0.26\,\mathrm{eV}$. This was expected, since SOAP describes the local environment in a higher level of detail. In our study, SOAP required 108 parameters to describe the local environment of the atom of interest, while ACSF required 36 and BLBA 17. A logical explanations is that the information contained in these extra parameters allows the SOAP descriptor to achieve more accurate predictions of formation energies of defects in the atomistic structures.
Overall, BLBA performed similarly to SOAP and ACSF. However, it is important to note the following differences:

- BLBA is more compact than SOAP and ACSF, since it is able to properly describe the relevant local environment with a significantly lower amount of information.

- SOAP and ACSF have several additional parameters particularly designed to represent the chemical surroundings. This makes them more accurate, but it also means that the user requires a deeper understanding of the underlying atomistic nature. Moreover, such necessity to adjust parameters which depend on the specific atomistic structure makes them less accurate when this information is not available. On the contrary, BLBA does not require to adjust any parameter (with the exception of N, the number of iterations, but this adjustment does not require any previous knowledge of the atomistic structure). Therefore, BLBA might perform better in atomistic structures of which there is a limited previous knowledge.

- Every element of BLBA represents a physical property of the atomistic structure (since each element of the descriptor represents a particular bond-length or bond-angle). This does not only make it simpler than SOAP and ACSF, it also gives BLBA a much higher interpretability.

- Our BLBA descriptor performed slightly better than the SOAP and ACSF descriptors when combined with a KRR or DT model. We believe that this could be explained given the fact that our BLBA descriptor is able to describe the local atomistic environments with a lower amount of parameters. Therefore, it performs better with reduced data-sets, by avoiding overfitting. However, more complex ML models, such as our NN model, might make use of the extra information provided by the SOAP descriptor to increase the formation energy predictions accuracy.

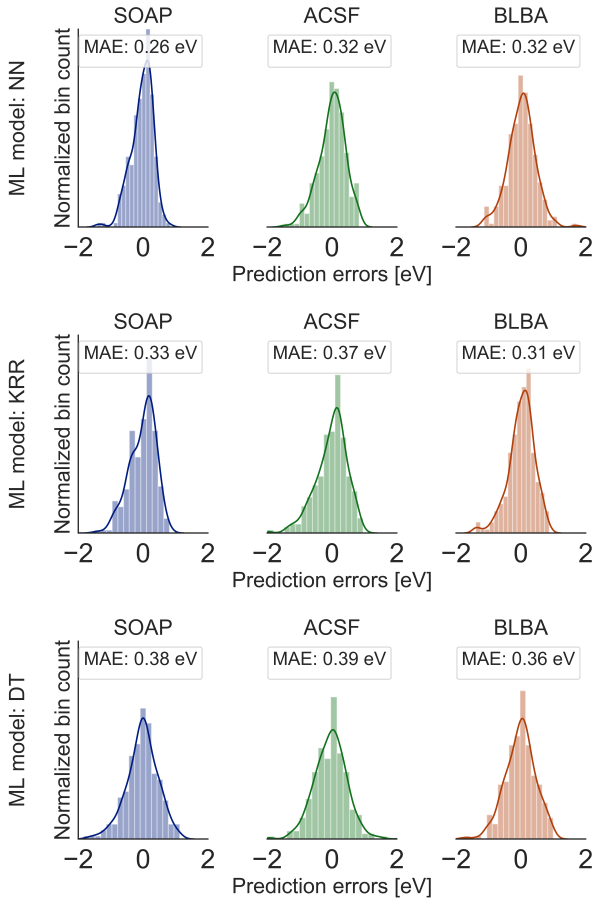
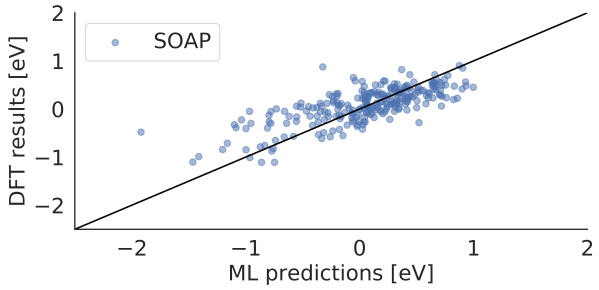
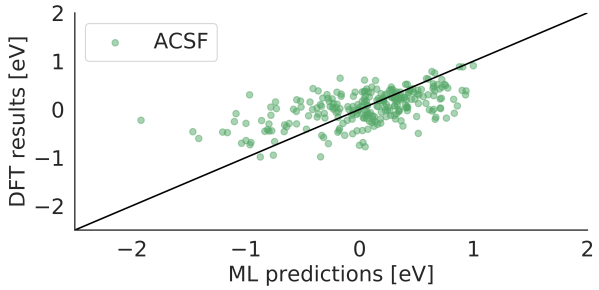


Fig. 6: Error distributions and MAE in the prediction of formation energy of hydroxyl E' center defects in a-SiO_2 for the different combinations of descriptors (SOAP, ACSF and BLBA) and ML models (NN, KRR and DT).

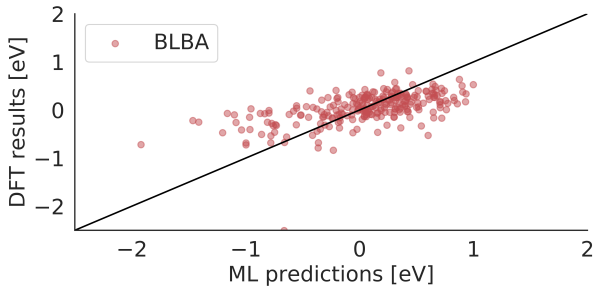
Even though there are noticeable differences in the accuracy of the defect formation energy predictions, it is important to remember that the predictions of all the solutions studied in this work are considered to be accurate enough for our practical application.
In our particular case, we are interested in using a combination of a descriptor and a ML model to predict the formation energies of defects in a-SiO_2. Such predictions will be combined with an KMC method in future works, in order to assess the likelihood of defects formation in specific conditions. In other words, this work can be considered as a first step towards a more sophisticated ultimate goal: to predict reaction barriers by using a combination of a KMC method and a ML model.
It is also important to remember that the accuracy of the predictions is highly dependent on the atomistic structures we consider and the electronic property we want to extract or predict from them. Therefore, the results obtained in this work do not necessarily imply that the prediction accuracy of these descriptors would be unchanged if we apply them to different materials or different electronic properties from the one considered here. In case of applying a descriptor-based ML solution to predict a certain electronic property, it is crucial to consider different combinations of descriptors and ML models and to analyze the results obtained from them before choosing the best combination.



(a)



(b)



(c)

Fig. 7: Correlation between the DFT calculation results and the ML-based predictions using (a) SOAP, (b) ACSF and (c) BLBA descriptors (in all cases, combined with a neural network). All descriptors perform similarly. However, SOAP shows a slightly better correlation in general, particularly for negative values.

In order to allow a clear comparison between our BLBA descriptor and the well-established SOAP and ACSF descriptors, a correlation between the DFT results and the predictions made by using each of these descriptors is shown in Fig. 7. In all cases, the descriptors were combined with a NN model, since it was the ML model which showed the best results in the previous section.

It can be noted that there is a clear linear correlation between the DFT results and the predictions made by all the descriptors. The fact that all distributions are roughly centered around the identity line shows there is no considerable systematic error in the predictions and no strong overfitting of the ML model.

The SOAP descriptor yields slightly better results than ACSF and BLBA, particularly for negative formation energies. However, all descriptors are accurate enough for our practical application.

IV. Conclusions

The study of reliability in modern electronic devices requires the use of simulation techniques in order to assess the effect of new materials and changes in the fabrication processes.

The direct study of these reliability issues posed by new materials and changes in the fabrication processes using DFT is currently a computational challenge due to the high costs. Therefore, computationally cheaper solutions must be found in order to substitute the direct application of DFT and other ab initio methods.

In this context, there are several practical applications in which such DFT calculations could be aided or even replaced by computationally inexpensive ML models combined with well-established descriptors, as shown in this work and in other examples in the literature [16–18].

Novel descriptors can be developed for specific applications, in the same way in which our BLBA descriptor was developed to be used in the ML-based prediction of formation energies of hydroxyl E' center defects in a-SiO_2. This approach could prove particularly useful in the study of amorphous materials, where large statistics are needed. Moreover, this solution could be applied to other defects, as well as to new materials, providing an enormous potential to aid the development of novel electronic devices.

Finally, apart from providing computationally inexpensive solutions to practical problems in the study of modern electronic devices reliability, descriptor-based ML solutions show an enormous potential to provide a deeper understanding of the mechanisms by which defects are formed and of the relationships between atomistic structure properties and the formation of such defects.

V. Acknowledgments

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 871813, within the framework of the project Modeling Unconventional Nanoscaled Device FABrication (MUNDFAB).

References

[1] Al-Moatasem El-Sayed et al. *Phys. Rev. Lett.*, 114, 2015.
[2] Tibor Grasser et al. *IEEE International Electron Devices Meeting*, pages 21.1.1–21.1.4, 2014.
[3] Yannick Wimmer et al. *Proc. R. Soc*, 472: 20160009, 2016.
[4] Wolfgang Goes et al. *Microelectron. Reliab.*, 87:286–320, 2018.
[5] Antonino La Magna et al. *Phys. Status Solidi*, 216, 2018.
[6] Steve Plimpton. *J. of Comp. Phys.*, 117:1–19, 1995.
[7] Adri C. T. van Duin et al. *J. Phys. Chem.*, 105:9396–9409, 2001.
[8] Al-Moatasem El-Sayed et al. *Solid State Electron.*, 109:68–71, 2013.
[9] John P. Perdew et al. *Phys. Rev. Lett.*, 77, 1996.
[10] Jürg Hutter et al. *WIREs Comput Mol Sci*, 4:15–25, 2013.
[11] Fabian Pedregosa et al. *Journal of ML Res.*, 12:2825–2830, 2011.
[12] Albert P. Bartók et al. *Phys. Rev. B*, 87, 2013.
[13] Jörg Behler. *Journal of Chem. Phys.*, 134, 2011.
[14] Lauri Himanen et al. *Comp. Phys. Com.*, 106949, 2019.
[15] T. Demuth et al. *J. Phys.: Condens. Matter*, 11:3833–3874, 1999.
[16] Felix Faber et al. *Int. Journal of Quantum Chem.*, 115:1094–1101, 2015.
[17] Shin Kiyohara et al. *Science Advances*, 2(11), 2016.
[18] Atsuto Seko et al. *Phys. Rev. B*, 95, 2017.