Machine Learning Prediction of Defect Structures in Amorphous Silicon Dioxide

Diego Milardovich¹, Markus Jech¹, Dominic Waldhoer¹,², Al-Moatasem Bellah El-Sayed¹,³ and Tibor Grasser¹

¹Institute for Microelectronics, Technische Universität Wien, Gußhausstraße 27–29, 1040 Vienna, Austria
²Christian Doppler Laboratory for Single-Defect Spectroscopy in Semiconductor Devices, Gußhausstraße 27–29, 1040 Vienna, Austria
³Nanolayers Research Computing, Ltd., 1 Granville Court, Granville Road, London N12 0HL, United Kingdom
E-mail: [milardovich | jech | waldhoer | el-sayed | grasser]@iue.tuwien.ac.at

II. METHODOLOGY

The first step towards an efficient ML based prediction is a consistent training data-set. We prepared 16 different defect-free a-SiO₂ models, each containing a total of 216 atoms. These structures were created by utilizing the melt-quench technique [9] within the molecular dynamics engine LAMMPS [10] in conjunction with the classical force field ReaxFF [11]. Examples of these structures are shown in Fig. 1. A total number of 1271 unique hydroxyl E⁺ centers were selectively created by placing a hydrogen atom in the vicinity of a bridging oxygen atom, followed by a subsequent geometrical relaxation. An elongated (weaker) Si-O bond will break and a hydroxyl (OH) group forms with a remaining defective silicon dangling bond in its neutral configuration, as shown in Fig. 2.

Fig. 1: Example of two a-SiO₂ structures containing 216 atoms each, used in this work to train and test the ML model. Yellow depicts silicon atoms, while red depicts oxygen atoms.

Fig. 2: A hydrogen atom is placed in the vicinity of a bridging oxygen atom (a) and the structure is subsequently relaxed. Breaking of an elongated Si-O bond leads to the formation of a hydroxyl E⁺ center (b), which is a silicon dangling bond facing a hydroxyl (OH) group.
Calculations were performed using the PBE functional [12] in the CP2K software package [13], which is a computationally expensive process. Typically, 4 nodes with 48 cores each require several hours to complete a single geometry relaxation. In total, the data-set is composed of 1271 hydroxyl E\textsuperscript{\textprime} centers and the respective defect free host structures. Once the data-set is created, the next step is to represent these structures in a way compatible with the selected ML algorithm. Such a mathematical representation of the structures (i.e., a vector or matrix) is called a descriptor. There is a wide range of available descriptors, ranging from local to global representations and including structural and/or electrostatic contributions, such as atom-centered symmetry functions (ACSF) [14]. Ewald sum matrix [15] and smooth overlap of atomic positions (SOAP) [16]. Among the different options, we chose a structural representation using the SOAP descriptor, since this descriptor showed the best performance in our previous work, in which we used a ML model to predict the formation energies of hydroxyl E\textsuperscript{\textprime} centers in a-SiO\textsubscript{2} structures [17]. However, given the flexibility of our approach, any other descriptor, even individually designed ones, can be used within the proposed workflow.

Out of the 16 a-SiO\textsubscript{2} structures, 15 are used to train the model and one is left for testing purposes. This translates into 1188 hydroxyl E\textsuperscript{\textprime} center structures in the training set and 83 in the testing set. The atoms within a 3 Å cutoff radius at every defect site, which on average contains about 9 atoms, are used to construct a local descriptor of the defect environment, utilizing the SOAP descriptor, as implemented in the Python package DSCribe [18]. To analyze the data, we employ a kernel ridge regression (KRR) model [19], as implemented in the scikit-learn package [20]. This is a cost-efficient model, which performed well in our previous study [17]. In this work, we extend its application to the prediction of defect structures, instead of just their formation energies.

Our ML model is trained to find the relationship between the SOAP descriptions of the hydroxyl E\textsuperscript{\textprime} centers and their host defect free a-SiO\textsubscript{2} structures. Once the ML model is trained, it is capable of predicting the formation of defects and their structural properties in new structures. The procedure is as follows:

1) **Prediction**: The environments of the surrounding atoms of interest in the defect-free structure are represented with the SOAP descriptor. This matrix is used as an input for the ML model, to predict the SOAP descriptor of the resulting defect complex.

2) **Optimization**: An optimization method is used to adjust the coordinates of the atoms in the defect free structure until its descriptor matrix matches the descriptor matrix predicted by the ML model for the defect structure. In this particular application, the Nelder-Mead method, with a convergence criteria of 5% of the initial loss function value.

3) **Validation**: The final step of the process is to validate the results. This is done by comparing the ML predicted defect structure with the equivalent target structure produced by DFT relaxation. We use the geometrical distance between the ML predicted and DFT relaxed defect site as a measure for the prediction quality.

The entire proposed workflow is graphically summarized in Fig. 3. As stated above, the same generic workflow could be used in the prediction of other defects in new materials.

### III. RESULTS AND DISCUSSION

In order to provide a clear illustration of the application of our proposed framework, the workflow shown in Fig. 3 is applied to the prediction of a particular defect, namely a hydroxyl E\textsuperscript{\textprime} center in an initially defect-free a-SiO\textsubscript{2} structure.
Fig. 4: (a) Prediction of the formation of a hydroxyl $E'$ center in an a-SiO$_2$ structure. The local environments of the atoms around the oxygen atom of interest are described with a matrix of SOAP descriptors (1). This matrix is used as input for our ML model (2), to predict the SOAP description matrix of the defective site (3). An optimization method is used to adjust the coordinates of the atoms in the input structure, in order to minimize the mean squared error (MSE) between the optimized and predicted SOAP matrices, and hence forms the defect in the a-SiO$_2$ structure. The bottom right figure shows how the difference between the initial and predicted defect structure descriptor matrices reduces as the optimization progresses. The final result (5) is the predicted defect structure. (b) Zoom to the predicted structure. (c) Superposition between the ML predicted defect structure (color) and its DFT-relaxed equivalent (shadow) around the defect site.

Details of our application and the individual steps are shown in Fig. 4. A bridging O atom in the top right of the structure is selected as the precursor configuration for the hydroxyl $E'$ center. Subsequently, the SOAP descriptor matrix of the surroundings of the O atom is computed with the parameters: $r_{cut} = 3.0$ Å and $n_{max} = l_{max} = 4$, as shown in Fig. 5. The resulting matrix is used as an input for our ML model, which was previously trained with the training data-set. Our ML model predicts the expected SOAP description matrix of the defective structure at the site of interest.

Once this prediction is made, a loss function is defined as the mean squared error (MSE) between the predicted SOAP and the defect free structure with one additional H atom placed in the direct vicinity, 1 Å, of the bridging atom, to form the defect. An optimization method is used to adjust the coordinates of the respective atoms in the input structure until this loss function reduces below a certain threshold value. The optimization method selected in this case is the Nelder-Mead method, implemented in the Python package SciPy [21]. However, given the flexibility of our approach, other optimization methods, including optional bounds or (non-) linear constraints can be used instead. The final result is the predicted structure of the defect, as shown in the bottom-left of Fig. 4.

Fig. 5: Defect free structure descriptor matrix. Each row of this matrix is the SOAP vector representation of the local environment of a specific atom. Only those atoms within a certain cutoff distance from the central oxygen atom (C.O.A.) of interest, are considered.

The complete prediction and optimization process took 0.5 seconds on a typical desktop machine (Intel Core i7 2.2 GHz and 8 GB of RAM).

Given the stochastic nature of the amorphous network, the accuracy and efficiency of our proposed framework must be analyzed on a statistical scale. We therefore benchmarked it against a full set of data of an a-SiO$_2$ model, the testing data-set. Nevertheless, the respective DFT calculations are available and were used as a reference by computing the distance vector of atoms within 6 Å around the defect site. The results are
scaled w.r.t. the number of atoms involved, in order to obtain a normalized quantity.

The loss function to be minimized in the optimization process was defined as the MSE between the SOAP matrices of the input and predicted structures (loss function A). A second loss function was defined by adding a penalty if the distance between the hydrogen and central oxygen atoms deviated from 1 Å (loss function B). The process of predicting the hydroxyl $E^r$ center defect structures in the testing data-set was repeated under identical conditions for both loss functions.

The results can be seen in Fig. 6, which shows the distribution of the deviations of the ML based approach compared to the DFT results across the testing data-set. For the 83 structures in the testing set, the average distance between the ML and DFT structures is 0.461 Å/atom for loss function A and 0.285 Å/atom for loss function B. This shows that the framework works without any user knowledge about the system. However, with some detailed information and knowledge, the individual modules can be fine-tuned to provide even more accurate predictions.

**Fig. 6:** Distribution of the deviations of the ML predicted defect structures compared to its equivalent DFT-relaxed structures, together with its mean value. Atoms within a 6.0 Å cutoff distance from the central oxygen atom are considered.

### IV. Conclusions

There is a wide range of applications in which machine learning (ML) based techniques can provide computationally inexpensive, but accurate solutions, as shown by several examples in the literature [4], [5], [15]. In this work, a solution was developed to study the formation of hydroxyl $E^r$ center defects in amorphous silicon dioxide ($\text{SiO}_2$). The results clearly demonstrate a competitive level of accuracy, while being computationally inexpensive when compared to DFT. The presented approach benefits from being highly modular, meaning that its components are interchangeable with other descriptors, ML models and optimization algorithms, depending on the individual problem. Hence, it can be easily extended and adapted to other defect species in $\text{SiO}_2$, or even different materials, such as hafnium dioxide ($\text{HfO}_2$).

Our approach is of particular importance for defect studies in situations where the application of DFT and other ab initio methods are too expensive, such as defect studies in large simulation cells and investigations on a large statistical scale. Furthermore, applications with a demand of on-the-fly calculations of certain data, such as kinetic Monte Carlo (KMC) simulations [22], benefit from the presented framework. Currently, reaction rates in such simulations have to be predefined. However, with our approach, structural and energetic information can be calculated almost instantaneously.

### 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 (MUNDIFAB). The authors acknowledge support from the Christian Doppler Laboratory for Single-Defect Spectroscopy in Semiconductor Devices and the Vienna Scientific Cluster (VSC).

### References


