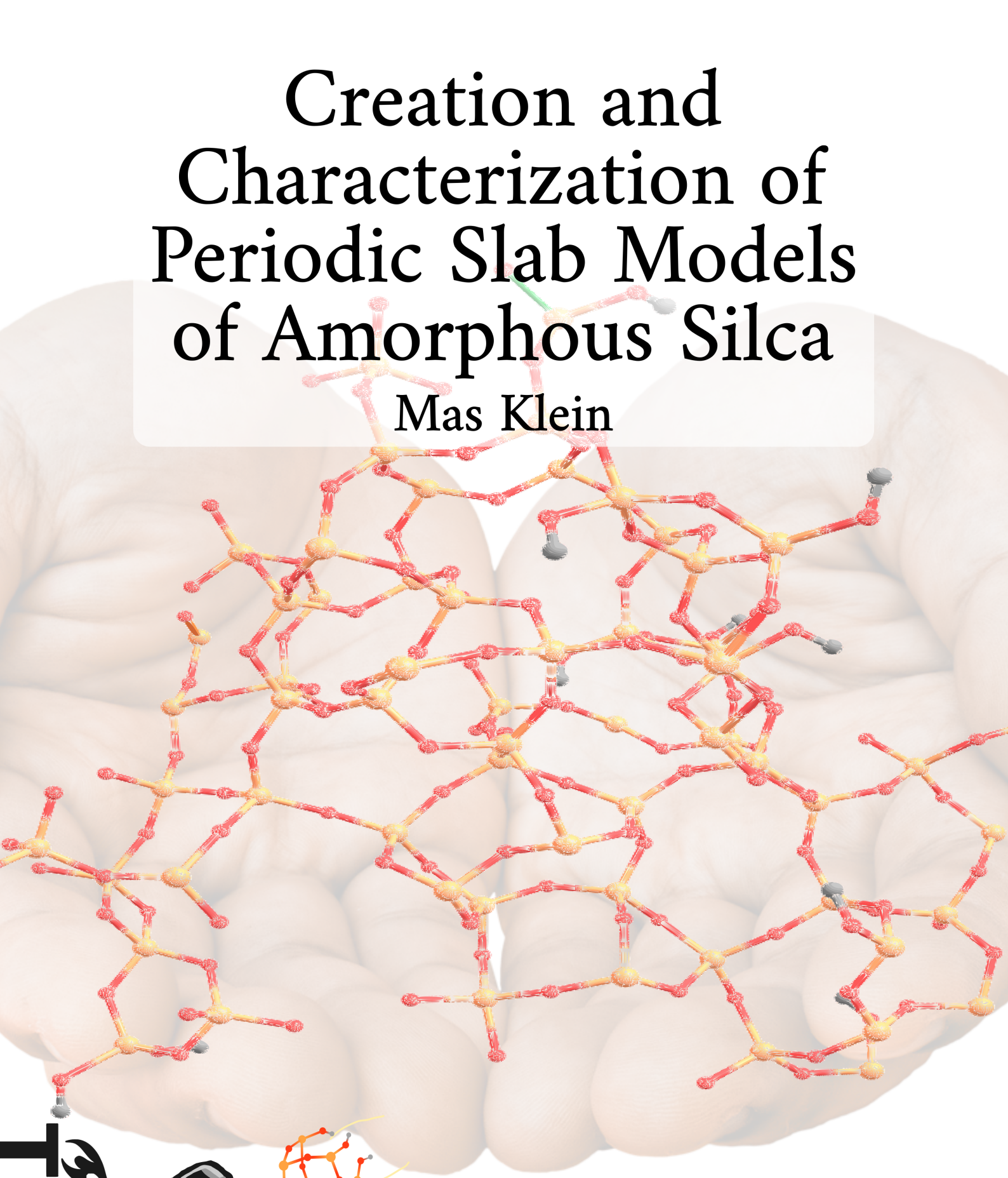
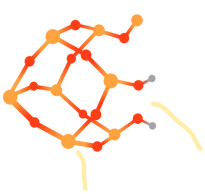
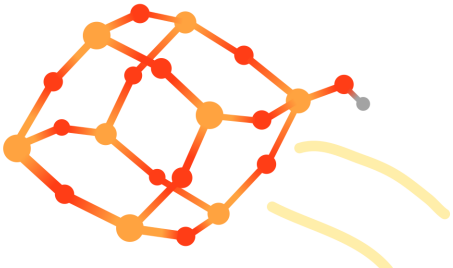
Creation and Characterization of Periodic Slab Models of Amorphous Silca

Mas Klein







TU Delft

Creation and Characterization of Periodic Slab Models of Amorphous Silica

by

Mas Klein

to obtain the degree of Bachelor of Science
at the Delft University of Technology

Performed at:
Faculty of Applied Sciences,
Inorganic System Engineering

Under the supervision of
Prof. Dr. E. A. Pidko
Dr. A. Kolganov

Student number:	3096637, 5661862
Project duration:	April 22, 2024 – July 4, 2024
Thesis committee:	Prof. Dr. E. A. Pidko, TU Delft, supervisor
	Dr. F. C. Grozema, TU Delft

TU Delft Delft University of Technology

ISEgroup ChemE @ TU Delft

Summary

Amorphous silica is a widely used material with many applications. Industrially, it has found common use as a catalyst support or adsorbent. As it is an amorphous material, the lack of long-range periodicity makes it difficult to reason what its surface looks like. As a consequence, when we want to make atomic models there is difficultly determining if they are representative. Furthermore, this difficulty extends to the active sites as there are many different possibilities with different local topologies and varying amounts of strain. This makes the computational modeling of the material a challenge to modern chemistry. This work aims to generate periodic models of amorphous silica of varying roughness and strain and use the topological features of the created models as descriptors for strain. To generate these models, classical molecular dynamics is used to generate bulks and equilibrate surfaces cleaved using a randomly generated stochastic Fourier expansion. DFT is then used to optimize the geometry of the resulting surfaces and their saturated counterparts. The calculated energies are compared to those of the most relaxed states of the substituents the surfaces are composed of.

It was found that the method of cleaving surfaces resulted in varying roughness after re-equilibration and that roughness has a correlation with strain. Varying the roughness had greatest effect on the amount of strained topological features in the model. Algorithmically saturating models showed that strain is generally decreased through the addition of water and strain is most effectively decreased through the removal of two membered rings on the surface. The main result of this study is that, using purely topological features, the strain of a model can be predicted using a multivariate linear regression. Using the coordination of O atoms, average bond lengths and angles as descriptors, multivariate linear regression was found to result in an R^2 of 0.925.

Contents

Summary i

1 Introduction 1

2 Theory 4
2.1 Classical Molecular Dynamics . . . 4
2.1.1 Classical mechanics . . . 4
2.1.2 Force Fields and Motion . . . 5
2.1.3 Many-Particle Systems . . . 6
2.1.4 Fluctuations in Temperature and Pressure . . . 6
2.1.5 Simulated Annealing . . . 8
2.2 Density Functional Theory . . . 8
2.2.1 The Kohn-Sham Equations . . . 9
2.2.2 Basis Sets . . . 10
2.2.3 Dispersion Correction . . . 11
2.2.4 Semi-empirical DFT . . . 12
2.3 ROBERT . . . 12

3 Methods 13
3.1 Computational Details . . . 13
3.1.1 Classical Molecular Dynamics . . . 13
3.1.2 DFT Geometry Optimizations . . . 14
3.1.3 Semiempirical DFT Geometry Optimizations . . . 14
3.2 Cleaving of Amorphous Bulk . . . 14
3.3 Algorithmic Saturation of Amorphous Surfaces . . . 15
3.4 Determination of Strain . . . 15

4 Results and Discussion 17
4.1 Creation of surfaces . . . 17
4.1.1 Roughness of Generated Models . . . 17
4.1.2 Topology of Generated Models . . . 18
4.1.3 Saturation of Generated Models . . . 20
4.2 Strain . . . 21
4.2.1 Strain of Dry Surfaces . . . 21
4.2.2 Strain of Saturated Surfaces . . . 23
4.2.3 Descriptors for the Strain of Dry Surfaces . . . 25
4.2.4 Screening Semi-empirical DFT . . . 29

5 Conclusion 31
5.1 Conclusion . . . 31
5.2 Outlook . . . 32

6 Acknowledgements 33

Bibliography 34

A Declaration of use of AI 38

B Remaining Topological Information 39

C Single Variable Linear Regressions 42

D Full statistical models Reports 46

List of Figures

1.1 Examples of amorphous silica models a) cluster-model bowl (taken from Caricato)[1] b) cluster-model crystal (from Goldsmith et al.)[2] c) periodic slab (taken from Comas-Vives)[3] 2

2.1 Illustration of Jacob's ladder . . . 10
2.2 Radial profile of the electron density of the Si atom after 0 ionizations (black), 1 ionization (red), 2 ionizations (green), and 3 ionizations (blue). Adapted from Garcia[4] . . . 11

3.1 Visualization of algorithm for saturating the surface. . . . 16

4.1 Mean-square displacement roughness compared to alpha of varying initial bulk thicknesses for a) models of *N* = 3 b) models of *N* = 1, 2 . . . 18
4.2 Probability distributions for a) Si-O-Si bond angles (°), b) O-Si-O bond angles (°), c) all Si-O bond lengths (Å), across all models of *N* = 3 . . . 20
4.3 Number of two-membered (blue) and three-membered (orange) rings averaged across surfaces of given α . . . 20
4.4 structures of Si (orange), O (red) and H (grey). ^{5}Si found in saturated structures where a) the silanol is attached to the ^{5}Si and b) the silanol is shared between Si atoms. . . . 21
4.5 Total strain of models *N* = 3 units for given values of α . . . 22
4.6 Range of $\Delta E_{per\ Si-O}$ for models of sizes *N* = 1, 2, 3, 4 unit cells thick . . . 23
4.7 $\Delta E_{per\ Si-O}$ of selected structure as the surface is saturated by the algorithm . . . 23
4.8 2-MR of a) generation 10, before saturation, and b) generation 11, after saturation. Si (orange), O (red), H (grey) . . . 24
4.9 $\Delta E_{per\ Si-O}$ after adding 9 and 19 water molecules algorithmically (blue) and manually breaking two-membered rings (orange) . . . 24
4.10 $\Delta E_{per\ Si-O}$ of saturated surfaces plotted against strain of corresponding dry surface for *N* = 2 (orange) and 3 (blue) unit cells . . . 25
4.11 Predicted $\Delta E_{per\ Si-O}$ from the multivariate linear regression model against the calculated from DFT $\Delta E_{per\ Si-O}$ . . . 27
4.12 Linear trend between the number of average bond angle and the strain per bond for the given model. . . . 28
4.13 Correlation between DFTB calculated electronic energy and DFT calculated electronic energy for dry surfaces of a) 3 units cells and b) 2 unit cells thick . . . 29
4.14 Correlation between DFTB calculated electronic energy and DFT calculated electronic energy for saturated surfaces of a) 3 units-cells and b) 2 unit cells thick . . . 30

List of Tables

3.1 Buckingham, Lennard-Jones forcefield parameters and assigned charges of atoms . . . 13

4.1 Average percentage of defects for surfaces of *N* = 3 19
4.2 Coefficients, t-values, significance of t-values for multivariate linear regression describing $\Delta E_{per\ Si-O}$ of all dry surfaces involving given independent variables. Original statsmodel reports can be found in figure D.1 26
4.3 Coefficients for multivariate linear regressions describing $\Delta E_{per\ Si-O}$ of dry surfaces of specified thickness involving given independent variables. Original statsmodel reports can be found in appendix D 29

1

Introduction

In chemistry there is a desire to know the molecular structure of what is being studied as it allows for discussion about molecular, physical properties, or reactivity.[5–7] Moreover, through the power of computational modeling these properties can be predicted or determined through simulation and calculation.[5,8] Partially through these reasons, x-ray crystallography is considered one of the preeminent structural analysis techniques as there is little ambiguity in structure if the spacial positioning of all atoms in a sample can be determined.[9] For crystals this allows for the construction of representative models through identifying a unit that repeats periodically to give a definable long-range structure. This is a unit cell of the material. However, there are materials for which this long-range periodicity cannot be defined as there is none. These are named amorphous materials.

One such example is silica (SiO_2). Fortunately. it does have structures where there is known periodic order. Examples include the polymorphs of SiO_2, α-quartz[10] and β-cristobalite,[11] or zeolites.[12] Yet, the forms of SiO_2 which are encountered daily, like the glass in our windows, or that which are used in our agriculture[13] or medicine,[14] are amorphous. Applying the structures of the perfect crystals is possible, and done within research[2], however the produced results deviate from experimentally determined values.[1,2]. This is why viable amorphous models of SiO_2 are desired. Especially when considering larger systems, the structural uncertainty of these amorphous materials grows drastically. Hence, despite the clear practical use and multitudes of applications, understanding the structure and surface of amorphous materials remains a challenge to modern chemistry.

Within industrial catalysis, amorphous SiO_2 is a common support upon which the catalyst can be dispersed.[15–17] To aid with understanding, computational modeling of the catalytic system is done to gain insight into the energy barriers, intermediates, and pathways that form the chemical reaction.[18–21] There are challenges specific to the creation of a single model but also related to the active site of the reaction. The challenge related to the creation of a model is two-fold: 1) the exact structure of amorphous SiO_2 is not known, so what is a representative model? 2) How does one create a representative reduced-model as there is no long-range periodic structure in the system and eon can only have so many atoms present in the simulation before it becomes too computationally expensive? As for those related to the active site, as there is no regularity in the structure, the active site of the catalytic reaction can plausibly take a on myriads of different local structures. This issue is further amplified when considering that there is likely no one single active site which is dominant in the contribution to the macroscopically observed rate.[2] Frameworks do exist to take into account the different contribution to the overall rate of each individual active site.[22] Nevertheless, to be applied they still rely on being able to identify, generate, and model a range of active sites. With these considerations, being able to

generate models of amorphous SiO_2 which can be viably used in computational modeling will improve understanding of the molecular and physical properties of the system and allow for more accurate modeling of the reactions happening on its surface.

The modeling of amorphous SiO_2 currently has two approaches: cluster type models and periodic slabs.[23] Cluster-type models consist of a fragment of silica which is cut from a larger, solid, surface with the dangling bonds saturated by hydrogen and assuming that this will be representative. The main advantages of this method is that they are, one, computationally less expensive and, two, easier to generate than periodic slabs. To create these structures one popular method is to cut out a ”bowl” of non-hydrogen atoms from a larger amorphous bulk and saturate the dangling atoms (figure 1.1a),[1] or to use a fragment of silica such as silsesquioxane molecules[2] (figure 1.1b). Through modeling the system in this fashion, non-physical relaxations happen at the edges as there are no interactions with a bulk beyond the atoms explicitly modelled. This is where periodic slabs (figure 1.1c) differ. Periodic slabs employ periodic boundary conditions (PBC) such that the structure is, essentially, infinite in the specified directions through using virtual copies. This prevents the artificial relaxation that cluster models fall victim to. These PBC’s also make them computationally more expensive compared to cluster models and as such limit the size of the system which can reasonably be modeled given specific computational resources. The model must also be made large enough such that there is no artificial interaction between the catalyst and its virtual copies.[2]

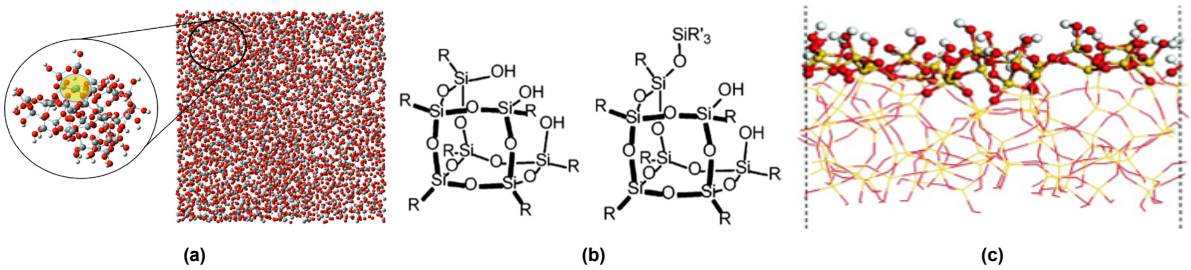


Figure 1.1: Examples of amorphous silica models a) cluster-model bowl (taken from Caricato)[1] b) cluster-model crystal (from Goldsmith et al.)[2] c) periodic slab (taken from Comas-Vives)[3]

The most common method employed when generating a periodic slab is referred to as a ”melt and quench” and in practice is a form of simulated annealing.[3,24–26] This entails melting a crystal of SiO_2 containing the desired amount of Si and O atoms to a high temperature, commonly above 4000 K, then cooling the system at a specified rate, commonly 1 K/ps, to approximately room temperature. This generates a distorted crystal structure which is subsequently cleaved followed by the saturation of the dangling bonds. The simulation for this procedure is carried out though the use of classical molecular dynamics (CMD). Cleaving the surfaces is commonly done on the basis of a flat plane defined by chosen Miller indices.[3] This method of cleaving has the drawbacks of artificially reducing the surface area of the model and having no way with which the surface roughness can be tuned. To remedy this, two methods have been proposed. One by Wimalasiri et al.[26] which uses the rate of cooling to influence the surface topology. Another by Nguyen and Laird[24] which uses a stochastic Fourier expansion to cleave a rough surface for the periodic slab. The main issues when employing the former method is that the variation which can be achieved is relatively limited and it does not work for potentials which become unstable at extreme temperatures.[26] The latter, however, shows promise as it does not fall to the aforementioned short-comings and the surface roughness is controlled through a empirical parameter which can be specified.[24]

Beyond simply generating the SiO_2 structures, the hydroxylation is a necessary step for the creation of an accurate model. Within certain industrial applications, highly dehydroxylated silica (circa. 1 OH/nm^2) is formed due to the high temperatures employed or simply through the synthesis of the

support.[3] However for computational modeling, silanol groups must be intentionally added or removed, preferably in a matter that leaves the model representative of the larger system. Creating these hydroxylated models has been achieved previously through manually saturating molecules then systematically removing silanol groups which look like they could react to leave the surface as water.[3] After each subsequent removal, DFT geometry optimizations are done. Put all together, this method leads to a fair amount of manual labor, computational time, and human bias. Nguyen and Laird[24] introduced silanol groups through adding bulks of water on top of and below the dry model and using a reactive force-field, ReaxFF,[27] to simulate how the water will react with the surface. In principle, obtaining a specific silanol density is possible through simply choosing a frame which has reached the desired silanol density. An algorithmic method of saturating and functionalizing the surface is employed by Wimalasiri et al.[26] They employed a probing method which finds Si-O bonds on the surface of their models and algorithmically added H_2O across said bond. Their method could either be altered to start with the longest Si-O bonds found on the surface or randomly selecting bonds and attempting to functionalize them through the use of a Monte Carlo method. One drawback of this method is that, as currently described in its implementation, once a bond is found to be able to be broken, it is. There is no comparison between the breaking of different bonds and which will lead to the more stable structure. While this may not necessarily lead to less-than physically realistic surfaces for higher silanol density surfaces, it might for lower; this in the sense that more strained bonds are more likely to be functionalized in practice.

The generation of surfaces with varying amounts of strain has also not been extensively studied. In this context, strain can be explained from the viewpoint of classical molecular modeling. When explicitly defining bonds for a CMD simulation, bonds and the angles which they make between each other are commonly described as harmonic oscillators.[8] The potential energy stored in these bonds and bond angles is defined relative to the equilibrium position. So, the energy calculated is an energy penalty for not being in its most relaxed state, it is the strain in the system. Experimental data has provided evidence that the reactivity of a SiO_2 surface can be increased through straining Si-O bonds.[28] Furthermore, Ab-initio computational studies have shown that a Si-O bond is not reactive until its angle has been brought far enough from its equilibrium state[29] and CMD studies using ReaxFF have shown that the hydroxylation of the Si-O bond is more favorable in a locally strained environment.[30] Thus, the importance of strain for a model's overall reactivity is understood however the generation of new models and quantifying their strain, or even showing what possible descriptors can be used to infer a model's overall strain, has not been explored as of yet.

This study aims to develop and build upon current methods of generating realistic models for amorphous SiO_2 surfaces and attempting to quantify their strain using topological descriptors. Semiempirical methods of electronic structure determination will also be screened. To do this, surfaces of varying roughness will be generated using a method like that of Nguyen and Laird.[24] and functionalized. Two methods of functionalization will be used: reactive force-fields or genetic algorithm. For both functionalized and dry surfaces descriptors for the strain of the model will be explored. Models of varying size will also be generated to test if the methods proposed can be applied independent of the desired size of model. This report is structured as follows: The theoretical background of the computational methods used to gather data is explained. This entails the background of both classical molecular dynamics and density functional theory. After this the methods used to generate data are described and subsequently results from said data are presented and discussed. Ultimately, conclusions from the results of the research and outlook on future possibilities are given.

2

Theory

In this chapter the underlying theory of the methods used for this thesis are described. First, the theoretical basis of classical molecular dynamics is explained then following that the necessary background of DFT is given. Finally, a brief description of the machine-learning tool, ROBERT, which was used for aspects of data analysis is given.

2.1. Classical Molecular Dynamics

When modeling at the atomic scale, the most accurate description is that of quantum mechanics. A drawback of this simulation method is the associated high computational costs for large systems. Thus, to model these larger systems an alternative approach must be taken. One such approach is handling the atoms within the given system as classically behaving particles. In many situations, this is possible because the particles behave close to classically at the desired thermodynamic conditions and scale. [31]

In this section, the relevant theory behind classical molecular dynamics (CMD) simulations is described. This encompasses Newton's classical laws of motion, force fields, many-particles systems and how to determine macroscopic properties from them, and how temperature and pressure are controlled in these classical simulations.

2.1.1. Classical mechanics

In 1687 Sir Isaac Newton determined three physical laws which describe the relation between the motion of particles and the forces which act upon them. [32] They are the following:

1. An object maintains its motion until acted upon by an unbalanced, external force.
2. The net force on a given body is equivalent to its mass multiplied by its acceleration (equation 2.1).

$$\vec{F}_i = m_i \vec{a}_i \tag{2.1}$$

3. If two bodies interact with one another, the resulting force upon each body is equal in magnitude but opposite in direction.

These three laws are what formulate the basis of *Newtonian mechanics*. The goal of CMD is to compute the evolution of the positions, speeds, and forces of all atoms over time using these equations.

The mass of each atom is known and the acceleration of an atom is derived from its movement. The question then is how to determine the force acting on a given atom. As the atoms in the system

are correlated through their interactions, collectively referred to the their interatomic potentials, these are used to derive the force which they experience at a given moment in time.[31] It can be shown that force can be calculated through the local gradient of the potential where an atom is positioned (equation 2.2) which, when combined with equation 2.1, results in equation 2.3. Now knowing the force an atom experiences, their motion over time can be determined. With a method to determine an atom's motion, all that remains is determining the potential at the position an atom is located.

$$\vec{F}_i = \nabla_i V(\boldsymbol{r}_i) \tag{2.2}$$

$$m_i \vec{a}_i = \nabla_i V(\boldsymbol{r}_i) \tag{2.3}$$

2.1.2. Force Fields and Motion

A forcefield is a collection of functions which describe all interatomic potentials between the atoms in a simulation. They are parameterized such that the behaviour of the atoms, as time evolves, is as similar to that of bodies handled in a quantum matter as possible.[31] A universal example of one of these interatomic potentials is coulomb interaction (equation 2.4) where k_e is a constant, Z is the nuclear charge of particles i and j, and r_{ij} is the distance between said particles. This equation is universal between all particles, granted they are charged, and is parameterized such that Z is the effective charge of the given particle. Only using this equation leaves out all dispersion and electronic interactions between the atoms. These forces are, in many cases, the primary determinants of the bonding and ultimate structure of a given collection of atoms.

$$\vec{F}(\boldsymbol{r}) = k_e \frac{Z_i \, Z_j}{\boldsymbol{r}_{ij}^2} \tag{2.4}$$

To represent remaining forces there are various semi-empirical equations, which are referred to as force fields. Examples include: the Lennard-Jones potential[33] (equation 2.5) and the Buckingham potential[34] (equation 2.6). The variables ϵ_{ij}, σ_{ij}, A_{ij}, b_{ij}, and C_{ij} are all parameters that can be optimized to yield physically accurate behaviour between atoms i and j. For a CMD simulation these forcefield potentials need to be defined for interaction between all possible combinations of atoms which are found in the simulation. Qualitatively, the two example functions tend towards 0 relatively quickly as the separation of the atoms increases. If the potential between two atoms is close to 0 then it can be neglected. The distance beyond which interactions are neglected is referred to as the cut-off distance and is determined by the one who configures the simulation. If a given atom is beyond the determined cut-off distance from another then its contribution to the potential at "far away" distances is not calculated which allows for computation to be done more efficiently. From this general description structural and thermodynamic properties can be derived however it is not effective at describing reactions or specific changes in bonding, in general[31]. For this, reactive force fields have been developed.

$$V_{LJ}(\boldsymbol{r}_{ij}) = 4\epsilon_{ij} \left[\left(\frac{\sigma_{ij}}{\boldsymbol{r}_{ij}} \right)^{12} - \left(\frac{\sigma_{ij}}{\boldsymbol{r}_{ij}} \right)^{6} \right] \tag{2.5}$$

$$V_{BUCK}(\boldsymbol{r}_{ij}) = A_{ij} e^{-b_{ij} \boldsymbol{r}_{ij}} - \frac{C_{ij}}{\boldsymbol{r}_{ij}^6} \tag{2.6}$$

First presented in 2001, ReaxFF is meant to be an intermediate step between CMD and quantum mechanical calculations when describing reactions.[27] This is achieved through ascribing and recalculating of charge and bond order of each atom at each timestep of the CMD simulation. The total energy of the system (E_{sys}) is calculated as the sum of 9 different energy contributions (equation 2.7).

$$E_{sys} = E_{bond} + E_{over} + E_{under} + E_{val} + E_{pen} + E_{tors} + E_{conj} + E_{vdW} + E_{coul} \tag{2.7}$$

ReaxFF retains terms accounting for coulomb interactions (E_{coul}) and van der Waal's interactions (E_{vdW}), just like normal force fields in CMD. The seven remaining terms are ones accounting for bonding energy (E_{bond}), over/under-coordination (E_{over} & E_{under} respectively), valence angle (E_{val}), a

"penalty" term to recreate the stability of atoms which share a double bond (E_{pen}), torsion angle (E_{tors}), and a correction term for to account for the stability of conjugated system (E_{conj}). All 9 of these terms are dependant on the bond order between two given atoms.

As what is being simulated is a many-body system, an exact analytical form for the trajectory of each atom over time cannot be determined. Hence, numerical integration over time is necessary. A commonly used algorithm for CMD is the velocity-Verlet algorithm[35] (equations 2.8) where $\boldsymbol{r}_i(t+\Delta t)$ is updated position of atom i after time step Δt, $\vec{v}_i(t+\Delta t)$ is the updated velocity, and $\boldsymbol{r}_i(t)$, $\vec{v}_i(t)$, $\vec{a}_i(t)$ are the current position, velocity, and acceleration respectively. Since its introduction it has become a preferred method of time-integration due to its numerical stability and relatively simple calculation procedure.[31]

$$\boldsymbol{r}_i(t+\Delta t) = \boldsymbol{r}_i(t) - \vec{v}_i(t)\Delta t + \frac{1}{2}\vec{a}_i(t)(\Delta t)^2 \tag{2.8a}$$

$$\vec{v}_i(t+\Delta t) = \vec{v}_i(t) + \frac{\vec{a}_i(t) + \vec{a}_i(t+\Delta t)}{2}\Delta t \tag{2.8b}$$

2.1.3. Many-Particle Systems

From statistical thermodynamics it is possible to calculate macroscopic properties from the individual states of the particles within a system.[36] In the context of this study, the macroscopic properties of interest are the temperature of the simulated system and its pressure.

Using the equipartition theorem, is can be proven that the expected kinetic energy ($\langle E_{KIN}\rangle$) of a system can be written as a function directly proportional to its temperature (T).[31] Alternatively, through rewriting this equation, the temperature of a system can be written as a function dependant on the individual velocities of each particle, as stated in equation 2.9. Here k_B is the Boltzmann constant, N is the total number of particles in the system, and m_i and $\vec{v}_i$ are the mass and velocity of particle i which is being simulated. As velocity is necessarily calculated for time integration, T is a readily calculated as well. Knowing the temperature also allows for the calculation of various thermodynamic properties,[36] such as pressure.

$$\langle E_{KIN}\rangle = \frac{3}{2}k_B T \Rightarrow T = \frac{2N}{3\,k_B}\left\langle \sum_{i=1}^{N}\frac{1}{2}m_i\vec{v}_i \right\rangle \tag{2.9}$$

In the simplest sense, pressure (P) is the amount of perpendicular force a system of particles exerts on a surface ($\vec{F}\cdot\hat{n}$) per unit area (A). This is expressed in equation 2.10. Within the context of CMD, it is preferable to derive a value of P from $\langle E_{KIN}\rangle$ (or T) of the system as to save time on calculation. It can be shown that this is possible and that the pressure of a system can in fact possible. P can be calculated from $\langle E_{KIN}\rangle$, the volume (V) of the simulation box, and the dot product between the forces and distances ($\vec{F}_{ij}\cdot\boldsymbol{r}_{ij}$) between every particle.[31] This equation is commonly referred to as the viral equation of state (2.11).

$$P = \frac{\vec{F}\cdot\hat{n}}{A} \tag{2.10}$$

$$P = \frac{2}{3V}\langle E_{KIN}\rangle + \frac{1}{3V}\sum_{ij}^{N}\vec{F}_{ij}\cdot\boldsymbol{r}_{ij} \tag{2.11}$$

2.1.4. Fluctuations in Temperature and Pressure

From statistical thermodynamics, CMD borrows the idea of ensembles. An ensemble, within the context of statistical thermodynamics, is a collection of virtual systems, all of which are possible states for the corresponding macroscopic system. To fix the possible states a system at equilibrium can take on, one must fix one of the three pairs of variables bellow:[36]

1. Number of particles (N) or chemical potential (μ)
2. Volume of the system (V) or pressure (P)
3. Temperature (T) or total energy (E)

Short-hand naming of these possible systems is done by abbreviating the properties which are fix. For example: a system of fixed N, V, and T is referred to as an NVT ensemble. Ensembles within CMD refer to which of the 3 pairs of system properties have been fixed. In reality, for the fixed thermodynamic properties, there is still a chance that these values fluctuate from their specified values.[36]

For CMD, there is compete control over the system. So, if for some reason these fluctuations are unwanted, they can be excluded. Nonetheless, there will be slight variation over time in, for example, T due to numerical error arising from time integration. Due to all derivatives and integrals being done numerically, there are small errors in the updated velocities which the machine calculates. These can be decreased through taking smaller time steps, however this increases the computational time required to reach a given total time and there are still errors regardless. This example is commonly referred to as ”temperature drift”.

To help prevent temperature drift, a thermostat can be applied to the system. A thermostat is also commonly used to control T for simulations which call for the heating or cooling of the system. A direct method of doing this is through applying a scaling factor (λ) to the velocity in equation 2.9. The simplest approach is to directly scale from the newly calculated temperature (T_{new}) to the desired temperature (T_0) (equation 2.12). This approach is now always ideal as it does not allow for the aforementioned fluctuations in temperature a natural system experiences.[31] For this reason the Berendsen thermostat[37] is more commonly employed to calculate the scaling factor (equation 2.13). The derivation of this scaling starts by assuming that the system is coupled to a heat bath of the desired system temperature to which the strength of coupling is determine by a time constant (τ_T). Through this method of scaling the temperature of the system is allowed to fluctuate, like that of natural systems, and how aggressively the system is regulated can be controlled through τ_T. A larger τ_T leads to more aggressive damping of temperature fluctuations.

$$\lambda = \sqrt{\frac{T_{new}}{T_0}} \tag{2.12}$$

$$\lambda = \left[1 + \frac{\Delta t}{\tau_T}\left(\frac{T_0}{T_{new}} - 1\right)\right]^{\frac{1}{2}} \tag{2.13}$$

The same approach may also be taken for P when applying a barostat to regulate the pressure of the system. In natural systems this thermodynamic variable also fluctuates. Furthermore, when the system is heated, the P increases as well. For values of T which can be considered extreme, the forces between particles can cause the simulation to break-down if the simulation box is not scaled in size.[24] To regulate pressure, instead of the velocities being scaled to approach the desired value, the dimensions of the simulation box are multiplied by a scaling constant (μ). To regulate pressure Berendsen et al.[37] also proposed a barostat derived from similar logic to that of their thermostat. The system is coupled to a pressure bath of the desired pressure (P_0) with a given coupling strength (τ_P) (equation 2.14). The variable β is the isothermal compressibility of the material being simulated. Note, in the derivation there is no mention of a specific dimension being scaled, let alone that all dimensions must be scaled equally. Thus, the simulation box can be isotropically (the expansion of a given direction is equal in the positive and negative direction) or anisotropically (the scaling of the positive and negative direction of a dimension is done independently from one another) expanded or contracted and each dimension can be independently scaled from one another. With the simulation box constantly changing dimension, it is possible that a given contraction of the box may cause particles to find themselves outside of it. For

this reason, periodic boundary conditions are required when regulating pressure.

$$\mu = 1 - \frac{\beta\,\Delta t}{\tau_P}(P_0 - P) \tag{2.14}$$

2.1.5. Simulated Annealing

The aim of simulated annealing is to find the global energy minimum on the potential energy surface through a random search.[8] For this CMD is commonly employed and the fact that temperature can be controlled in a simulation is leveraged. Raising the temperature of the system allows greater potential barriers to be over come but, in some situations, could lead to unwanted alterations in the system: changes in bonding, fragmentation. It is then subsequently cooled in a predetermined fashion, a cooling "schedule", in the hope that it will fall into the global energy minimum.[38] If this cooling is slow enough such that the system stays in equilibrium with its so-called surroundings, statistical thermodynamics states that the global minimum will be found with the relative probability 1 and any higher energy state according to equation 2.15. Here $\Delta E_{ref,i}$ is the energy of the state i referenced to that of the energy of global minimum, k_B is the Boltzmann constant.

$$P_i = \exp\frac{-\Delta E_{ref,i}}{k_B T} \tag{2.15}$$

2.2. Density Functional Theory

In reality, atomic systems are not truly classical in their physics. This realization gave birth to quantum mechanics. How to calculate the energy of such non-classical systems was first realized by Erwin Schrödinger in 1926.[39] From first principles, it is calculated from the time-independent, non-relativistic Schrödinger equation (equation 2.16)

$$\hat{\boldsymbol{H}}\Psi_i = E_i\Psi_i \tag{2.16}$$

where E_i is an eigenvalue which corresponds to the energy of the system found when the Hamiltonian operator ($\hat{\boldsymbol{H}}$) acts upon a wave function (Ψ_i) which itself is an eigenfunction of said operator. $\hat{\boldsymbol{H}}$ contains all contributions to both the kinetic and the potential energy within a system. For a molecular system of N electrons, of mass m, and K nuclei, of mass and M and nuclear charge Z, the full Hamiltonian in atomic units is given in equation 2.17a (short-hand in equation 2.17b)

$$\hat{\boldsymbol{H}} = -\sum_{i=1}^{N}\frac{1}{2m_e}\nabla_i^2 - \sum_{g=1}^{K}\frac{1}{2M_g}\nabla_i^2 - \sum_{i=1}^{N}\sum_{g=1}^{K}\frac{Z_g}{\boldsymbol{r}_{ig}} + \sum_{i=1}^{N}\sum_{j>i}^{N}\frac{1}{\boldsymbol{r}_{ij}} + \sum_{g=1}^{N}\sum_{h>g}^{N}\frac{Z_g Z_h}{\boldsymbol{r}_{gh}} \tag{2.17a}$$

$$\hat{\boldsymbol{H}} = \hat{T}_e + \hat{T}_n + \hat{V}_{en} + \hat{V}_{ee} + \hat{V}_{nn} \tag{2.17b}$$

where, from left to right, the kinetic energy of the electrons in the system ($\hat{T}_e$), the kinetic energy of the nuclei in the system ($\hat{T}_n$), attraction between each electron with each nuclei ($\hat{V}_{en}$), repulsion between every electron pair ($\hat{V}_{ee}$) and the repulsion between each nuclei ($\hat{V}_{nn}$).

To simplify equations 2.17 the Born-Oppenheimer approximation can be applied. This states that the nuclei are positioned in the potential created by the electrons of the system.[40] This approximation can be reasonably applied as the nuclei move on a much slower time-scale than the electrons due to being much heavier. This allows for the separation of the total wave function into a product of wave functions. One wave function which describes the electrons and one which describes the nuclei of the system. This separation of wave function also leads to the ability of describing the energy of the system using two different Hamiltonian operators, one for the energy of the electrons (the electronic Hamiltonian) and one for the energy of the nuclei. The electronic Hamiltonian ($\hat{\boldsymbol{H}}_e$) is given by equation 2.18 where $\hat{V}_{ext}$ is an external potential exerted by the nuclei of the system on the electrons.

$$\hat{\boldsymbol{H}}_e = \hat{T}_e + \hat{T}_n + \hat{V}_{ee} + \hat{V}_{ext} \tag{2.18}$$

In principle, finding an analytical solution to the time-independent Schrödinger equation, making use of the electronic Hamiltonian, will give the exact electronic energy of the system. However, this equation cannot be solved for systems containing more than one nucleus and one electron. As such, finding the exact electronic energy of a many-particle system is technically impossible.

2.2.1. The Kohn-Sham Equations

To find an approximate electronic energy for a structure, various methods have been developed. The method of interest for this work is that built upon the theorems proposed by Hohenberg and Kohn.[41] These form the theoretical basis for the equations put forth by Kohn and Sham[42] which themselves are the ground work for density functional theory (DFT).

The first Hohenberg-Kohn theorem proves that $\hat{V}_{ext}(\boldsymbol{r})$ is a unique functional of the electronic density ($\rho(\boldsymbol{r})$). Since $\hat{V}_{ext}(\boldsymbol{r})$ determines $\hat{\boldsymbol{H}}$, and $\hat{\boldsymbol{H}}$ determines the wave function, the ground state of a system is uniquely determined by $\hat{V}_{ext}(\boldsymbol{r})$. A functional is a function that takes another function as its independent variable. The second theorem proves that the variational principle is applicable to this problem. In this context, the essence of the variational principle is that any trial density ($\tilde{\rho}(\boldsymbol{r})$) will have a corresponding energy ($E[\tilde{\rho}]$) that is greater than or equal to the energy associated with the true ground state density and as such the $\tilde{\rho}(\boldsymbol{r})$ with the lowest energy is the one closest $\rho(\boldsymbol{r})$. A consequence of these theorems and their proofs is that DFT, in principle, is an exact method; the ground state electronic energy can be solved for exactly.

So, from the first theorem of Hohenberg and Kohn we know that a given $\rho(\boldsymbol{r})$ maps to a given $\Psi(\boldsymbol{r})$ and the second theorem gives a guideline on how to choose the best approximation for $\rho(\boldsymbol{r})$. However, the theorems give no method of, one, how to actually map the density to a wave function and, two, how to vary $\tilde{\rho}(\boldsymbol{r})$ to get closer to $\rho(\boldsymbol{r})$. Further more, if $\rho(\boldsymbol{r})$ were to be converted into a wave function then there would be no simplification in the solving of this problem since the final step would be solving the Schrödinger equation.[8] These problems where addressed in 1964 through the proposal of the Kohn-Sham self-consistent field approach by Kohn and Sham.[42] This approach involves taking a fictitious system of non-interacting electrons which has the same $\rho(\boldsymbol{r})$ as the real system where the electrons do interact. Since $\rho(\boldsymbol{r})$ is identical for both the real and fictitious system the position and atomic number of the nuclei in both systems must be identical as well. Since we are assuming a system of non-interacting particles, we can write the Hamiltonian of the full system as a sum of Hamiltonians for each individual electron. This also has the beneficial consequence that the Hamiltonian of the system has eigenfunctions which are Slater determinants of each one-electron Hamiltonian and eigenvalues which are sums of the individual eigenvalues of the one-electron Hamiltonians.[43] Another beneficial consequence of this method is that it reduces the dimensionality of the problem from 4 degrees of freedom per electron, the cartesian coordinates and spin required to describe them, to exactly 3 dimensions, the three cartesian coordinates required to describe every point in the electron density.

Mathematically, the Kohn-Sham energy functional is split into the following respective components

$$E[\rho(\boldsymbol{r})] = T_{ni}[\rho(\boldsymbol{r})] + V_{ne}[\rho(\boldsymbol{r})] + V_{ee}[\rho(\boldsymbol{r})] + \Delta T[\rho(\boldsymbol{r})] + \Delta Vee[\rho(\boldsymbol{r})] \tag{2.19}$$

where the terms on the right-hand side of the equation, from left to right, are: the kinetic energy of the non-interacting electrons, the nuclear-electron attraction, the classical electron-electron repulsion, a correction term on the kinetic energy of the electrons to account for their interactions, and a correction term that account for any non-classical behaviour of the electrons. As stated earlier, DFT is an exact method and if all 5 of the terms in equation 2.19 can be found - exactly - for a given $\rho(\boldsymbol{r})$ then the exact electronic ground state energy of that density is known. The energies associated with the first three terms can be found exactly. The difficulty lies in the final two terms, which are generally grouped together and commonly denoted as $E_{XC}[\rho(\boldsymbol{r})]$, which are referred to as the exchange-correlation functional which gives the exchange-correlation energy (E_{XC}).

There are various $E_{XC}[\rho(\boldsymbol{r})]$ which can be split into groups depending on their general degree of mathematical complexity for computing E_{XC}. These are summarized through Jacob's ladder (figure 2.1), proposed by Perdew[44]. For the purposes of this report, a "generalized gradient approximation" functional will be used. This calculates E_{XC} through taking both the electron density at a specific point and its local gradient ($\nabla\rho(\boldsymbol{r})$). This method provides a compromise between computational accuracy and efficiency.

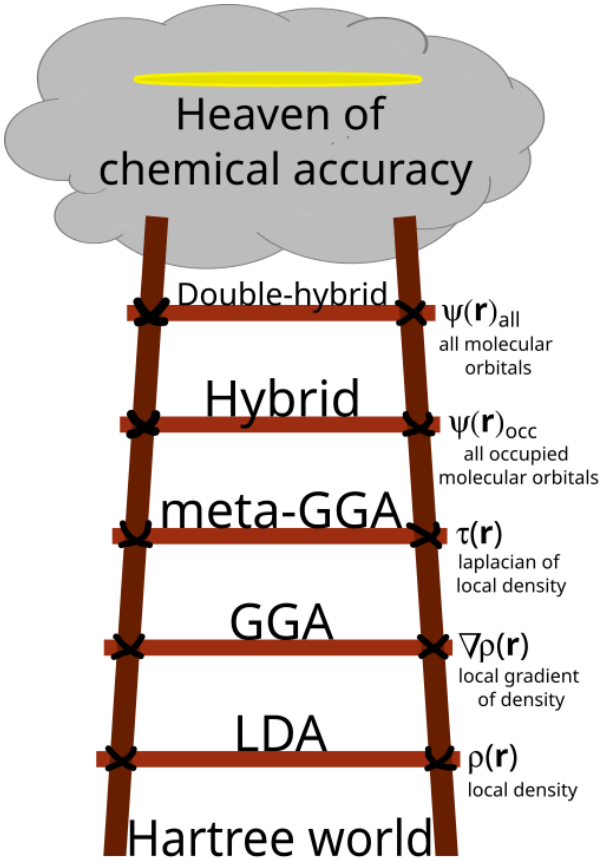


Figure 2.1: Illustration of Jacob's ladder

2.2.2. Basis Sets

To calculate the energy associated with a given $\tilde{\rho}(\boldsymbol{r})$ using a chosen $E_{XC}[\rho(\boldsymbol{r})]$ a mathematical description of the density must be made. This problem consists of approximating the molecular orbitals (MOs) of each electron within the system. To carry out this approximation a set of pre-defined mathematical functions is chosen. This set of functions is referred to as the basis set. The choice of these functions is freely up to the one who constructs the basis set but one popular choice is a basis set consisting of Gaussian type orbitals (GTO)[8] as they are efficient to integrate. The ability of a program to construct a realistic $\rho(\boldsymbol{r})$ is dependant on the choice of basis-set. At the same time, a larger basis set will lead to longer computation times as, while an increase in basis functions can improve accuracy, every electron for which the increased basis set in defined adds one additional coefficient to vary from the variational principle.

One way of decreasing the number of MOs to calculate is make use of pseudopotentials. Using the Si atom as an example, the core electrons - those in filled orbitals - are not particularly affected by the electrons in the valence shell (figure 2.2). If the contributions of the core electrons stays almost constant, then it is reasonable to only vary valence electrons and replace all core electrons with relatively accurate

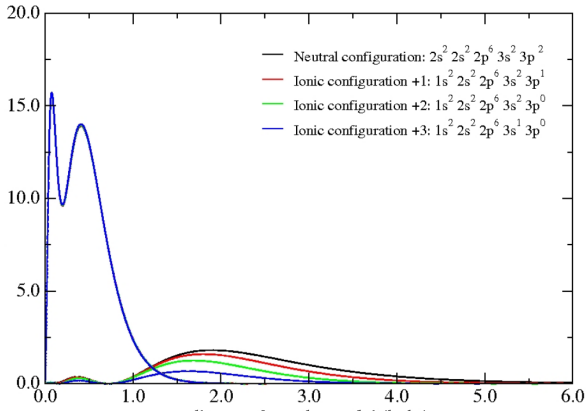


Figure 2.2: Radial profile of the electron density of the Si atom after 0 ionizations (black), 1 ionization (red), 2 ionizations (green), and 3 ionizations (blue). Adapted from Garcia[4]

analytical functions.[8]

Within the software package CP2K[45], the Gaussian plane-wave (GPW) method is used when dealing with periodic systems. This entails describing the valence electrons using Gaussian functions which, along with the Pseudopotentials, are converted into a plane-wave description.[46] This conversion into plane-waves is done because, one, they are inherently periodic which allows them to describe periodic systems and, two, they have beneficial mathematical properties which allow for efficient calculations to be done for said periodic systems.[46]

2.2.3. Dispersion Correction

With DFT being based on electron density, it does not fully account for dispersion forces. These forces are a result of induced dipole moments within the system. For larger systems this causes a deviation in the accuracy of the calculations. Examples of methods to correct for dispersion are the D3[47] and D3BJ[48] correction schemes of Grimme. The D3 method entails adding a correction term to the energy calculated through the Kohn-Sham self-consistency method which encompasses all disperse interactions (equation 2.20) where E_{DFT-D3} is the corrected electronic energy after dispersion interactions are taken into account, E_{KS-DFT} is the electronic energy as calculated through Kohn-Sham DFT, and E_{disp} is the energy correction for dispersion interactions.

$$E_{DFT-D3} = E_{KS-DFT} - E_{disp} \tag{2.20}$$

E_{disp} is calculated through equation 2.21a where $E^{(2)}$ is the two-body correction term and $E^{(3)}$ is the three-body correction term. The two-body correction term is calculated through equation 2.21b where s_n is an empirically determined scaling dependant on the n taken, C_n^{AB} is the nth order dispersion constant determined through time-dependant-DFT between atoms A and B, r_{AB}^n is the distance between atoms A and B, and $f_{d,n}(r_{AB})$ is a damping function. $f_{d,n}(r_{AB})$ is necessary so that the correction gives repulsive values when the atoms get too close together; without $f_{d,n}(r_{AB})$ equation 2.21b would give larger and larger attractive values as the atoms get closer and closer together. The correction is truncated after terms corresponding to n = 8 because the inclusion of terms of larger n led to numerical instability. $E^{(3)}$ is calculated between all triplets of atoms. E^{ABC} is calculated using C_9^{ABC}, which is the negative of the geometric average between all C_6 terms between each possible pair of atoms A, B, and C and θ_i which is equal to the internal angle of the triangle created those atoms.

$$E_{disp} = E^{(2)} + E^{(3)} \tag{2.21a}$$

$$E^{(2)} = \sum_{AB} \sum_{n=6,8} s_n \frac{C_n^{AB}}{r_{AB}^n} f_{d,n}(r_{AB}) \tag{2.21b}$$

$$E^{(3)} = \sum_{ABC} f_{d,(3)}(\bar{r}_{ABD}) E^{ABC} \tag{2.21c}$$

$$E^{ABC} = \frac{C_9^{ABC} \left(3\cos(\theta_a)\cos(\theta_b)\cos(\theta_c) + 1\right)}{(r_{AB} r_{BC} r_{CA})^3} \tag{2.21d}$$

D3BJ builds upon D3 through altering the damping function used to that of Becke and Johnson.[48]

2.2.4. Semi-empirical DFT

With CMD being physically inaccurate and DFT costing a significant amount of resources, using semi-empirical DFT is a quicker method of simulating large systems while still retaining non-classical phenomena. This approach replaces costly two-electron integrals with either approximations fitted to experimental data or approximate analytical functions.[49] These simplifications make it considerably faster than true DFT. For this work, focus is put on self-consistent charge density-functional tight binding (SCC-DFTB).[50]

Built upon the work of Porezag et al.[51], SCC-DFTB aims to bring self-consistency into the methodology of DFTB. DFTB in its purest implementation takes a given density, determined through whatever method is deemed fitting, and optimizes the Kohn-Sham orbitals to fit said density.[8] From these orbitals a new density is not calculated and as such is not self-consistent. However, adding a step when the partials atomic charge is calculated based on the Kohn-Sham orbitals reintroduces a self-consistency requirement in the optimization process. SCC-DFTB has been found to compare favorably to the MP2 method and the true DFT functional B3LYP for specific biological systems.[8]. This goes to show that the assumptions and approximations made when using semi-empirical methods can provide relatively accurate predictions as to what the electronic energy is.

2.3. ROBERT

ROBERT[52] is a collection of automated machine learning workflows aimed at simplifying the task of applying machine learning to cheminformatics problems. It is capable of automated curation of data, model selection, prediction based on the given model, and verification of the results. Verification is done through various methods such as y-shuffle and y-mean tests, and k-fold cross-validation.[53] ROBERT is a particularly useful tool when working with descriptors which may not necessarily have linear relations to the target property.

3

Methods

In this chapter the computational details are given regarding the computational methods employed for the generation of data used for this thesis. Following that the definition used to calculate the strain of a given model is given and the method used for cleaving bulk $SiO_2$to generate surfaces is described. Lastly, a general overview of the algorithm used to saturate models is given.

3.1. Computational Details

3.1.1. Classical Molecular Dynamics

Classical molecular dynamics simulation were carried out using the Large scale Atomic/Molecular Massively Parallel Simulator (LAMMPS) software package.[54] The Buckingham (cut-off 5.5 Å) and Lennard-Jones (cutoff 1.2 Å for Si-O and 1.6 Å for O-O) potentials were chosen as force fields with parameters identical to that of Nguyen and Laird[24] (table 3.1). Charges for coulomb interactions (cut-ff 8 Å) can also be found in Table 3.1. Throughout all simulations periodic boundary conditions in x, y, and, when stated, z directions where enforced. All simulations had fixed xy cross-sectional dimensions of 21.5 Å. A Berendsen thermostat (damping constant of 1 ps) and Berendsen barostat (damping constant 1 ps, modulus 360000 atm) where also employed to control the temperature and pressure. These above-mentioned details will be referred to as the NP_ZAT ensemble. A time step of 0.5 fs was used along with the velocity-Verlet integration scheme. Ewald summation[55] (desired relative error of 10^{-4}) was used to account long-range periodic coulomb forces.

Table 3.1: Buckingham, Lennard-Jones forcefield parameters and assigned charges of atoms

	Buckingham parameters			Charge	Lennard-Jones parameters	
i-j	$A_{ij}(eV)$	$b_{ij}(\text{ Å}^{-1})$	$C_{ij}(eV\ \text{Å}^{-6})$	$q_i(e)$	$\epsilon_{ij}(eV)$	$\sigma_{ij}(\text{ Å}^{-1})$
Si-Si	0	0	0	2.4	0	0
O-O	1388.7730	2.76000	175.0000	-1.2	2.0	1.2
Si-O	18,003.7572	4.87318	133.5381	-	2.6	1.6

To generate the amorphous bulks, a 3 by 3 by *N* unit cells of β-cristobalite[11] was generated. *N* being the number of repeating units in the z-direction. This perfect crystal was heated to 8000 K then subsequently cooled to 298 K at a rate of 1 K/ps. This structure was saved to be reused as the staring

point for if more distorted crystals were desired. The saved structure was then heated to 4000 K within 200 ps then, again, cooled at a rate of 1 K/ps. In total 750 models were generated. 600 were of *N* = 3, 60 were of *N* = 1 and 2, and 30 for *N* = 4.

To equilibrate cleaved bulk, simulations were done at 298 K for 2 ns under an NVT ensemble where the z-directions of the unit cell was expanded in each direction by 20 Å. The Change in ensemble from NP_ZAT to NVT was done to prevent contraction of the simulation box leading to periodic interactions in the z-direction. The damping constant of the Berendsen barostat was removed however the modulus was kept as is.

To functionalized the surfaces using LAMMPS, ReaxFF was employed. ReaxFF parametrization from the work of Yeon and van Duin[30] was taken. The simulation was set up as follows: a block of 200 water molecules was placed above and bellow the cleaved surface with at least 2.5 Å between the interfaces then allowed to simulated for 1ns. These blocks of water were generated using Packmol.[56]

3.1.2. DFT Geometry Optimizations

All DFT calculations were carried out using the CP2K software package[45] using periodic boundary conditions identical to the CMD simulations with the z-direction expanded to 80 Å as to minimize periodic interactions in that direction. The GPW method of calculation was used with the PBE functional and D3BJ dispersion correction[48]. GTH-pseudo potentials[57–59] and the DZVP basis-set[60] was used for Si, O, and H and the cutoff for plane waves was 300 Ry for dry surfaces and 450 Ry for saturated ones. SCF calculations were converged to 10^{-4} Eh.

3.1.3. Semiempirical DFT Geometry Optimizations

Semiempirical calculations were done using SCC-DFTB[50] method as implemented in the CP2K[45] software package. The semiempirical parameters, optimized for aluminosilicate nanotubes of imogolite, used were taken from the work of Guimarães et al.[61] An Ewald summation[55] was used to calculate long-range coulomb forces and remaining details were identical to that of DFT optimizations.

3.2. Cleaving of Amorphous Bulk

To cleave the amorphous silica bulks, inspiration was taken from the method of Nguyen and Laird.[24] Surfaces were cleaved using random functions generated using the following formula, a stochastic Fourier series

$$z(x,y) = \sum_{m=1}^{m_{max}} \sum_{n=1}^{n_{max}} b_{mn} \sin\left(\frac{m\pi x}{L_x}\right) \sin\left(\frac{n\pi y}{L_y}\right) \tag{3.1}$$

where m and n are the number of Fourier modes for the given function, b_{mn} is the Fourier coefficient corresponding to a given m with n, and L_x and L_y are the x and y dimensions of the simulation box. The maximum values of m and n, m_{max} and n_{max}, were chosen such that the minimum half-wavelength a given maximum mode is approximately 1.8 Å. This resulted in 12 Fourier modes in both the x- and y-direction.

The Fourier coefficients for a given m and n where randomly sampled from the following normal distribution

$$P(b_{mn}) = \sqrt{\frac{\alpha(m^2+n^2)}{2\pi}} exp\left[-\frac{\alpha(m^2+n^2)}{2} b_{mn}^2\right] \tag{3.2}$$

where α is introduced as an empirical roughness parameter which can be freely chosen. As the value of α increases the cleaving functions generally become more planar as the standard deviation of the distribution becomes smaller. A normal distribution is sampled because there is not yet physical evidence to suggest a more suitable.[24] For this work values of 0.01, 0.02, 0.03, 0.04, 0.05, and 0.1 were chosen and equally represented within each different *N*.

3.3. Algorithmic Saturation of Amorphous Surfaces

The algorithm coded was genetic in genetic in nature in the sense that it uses the ”optimal” structures previously generated as the basis for new ones. The procedure coded to saturate surfaces can be put broadly into the following four steps:

1. Determine all over/under-coordinated atoms.
2. Saturate until a full generation has been made.
3. Calculate the single-point energies of all structures and chose the parents for the following generation.
4. Output most stable structures after the whole list of over/under-coordinated atoms is exhausted.

These steps are made such that the most stable structures are carried onto the step generation and ultimately get approximately the most stable structure possible. This also lays the frame-work for further saturation when breaking Si-O bonds of correctly-coordinated Si’s and O’s. A visual overview of the algorithm can be found in Figure 3.1

Step one, in more detail, determines the coordination of a given atom as the number of atoms within a 2.0 Å sphere are it. This cutoff is derived form the work of Wells et al.[62] and is approximately 62% of the sum of the van der Waal’s radii of the two atoms. All Si and O which are found to have more or less than 4 other atoms or 2 other atoms within this sphere, respectively, are added to the list of over/under-coordinated atoms.

Next, an under-coordinated Si and O are chosen together at random and they are saturated with OH and H, respectively. The indexing number of the atoms chosen are then appended to a list to keep track of which atoms have already been saturated. This list is specific to the structure and stays with it for entire process. Once pairs under-coordinated atoms are no longer available, over coordinated atoms are chosen. These atoms are not directly saturated but the bond between them and another atom is broken and the newly under-coordinated atom is saturated. For over-coordinated Si’s, the O farthest away from it, within the 2.0 Å tolerance, is chosen and moved away such that it is at least 2.5 Å away from the Si it was previously bound to and more than 2 Å away from any other atoms. These numbers were determined empirically through visual testing. This O is then saturated with a H. For over-coordinated O’s, they are moved away from the farthest Si until just outside of the 2.0 Å cut-off radius. The Si which it was moved away from is subsequently saturated with a OH group. This process of saturation is done until a predetermined number, a ”generation”, of structures are created with one water molecule added to them compared to their parents.

Once a generation is filled, the single-point energies of each structure is calculated using the GFNFF[63] method as implemented by xTB.[64] This force field method was chosen as it was easily implementable. The structures with the lowest energies are then chosen as the parents of the next generation, the structures which the next generation is based upon. This creation and energy calculation of structures is repeated until all atoms within the initial list of over- and under-coordinated atoms have been exhausted at which point the algorithm outputs the structures with the lowest energies.

For the saturation of surfaces, only the sizes of 2- and 3- unit cells thick were chosen as they find a balance between being representative models and computational efficiency.

3.4. Determination of Strain

The strain of a surface (ΔE_{strain}) was determined using equation 3.3 where $E_{surface}$ is the electronic energy of the surface, $E_{\alpha-quartz}$ and E_{H_2O} and the electronic energies of α-quartz and water, respectively, as calculated by a given method. m is the number of α-quartz units cells required to build model and n is the number of water molecules use to saturate the model, if applicable. α-quartz was chosen

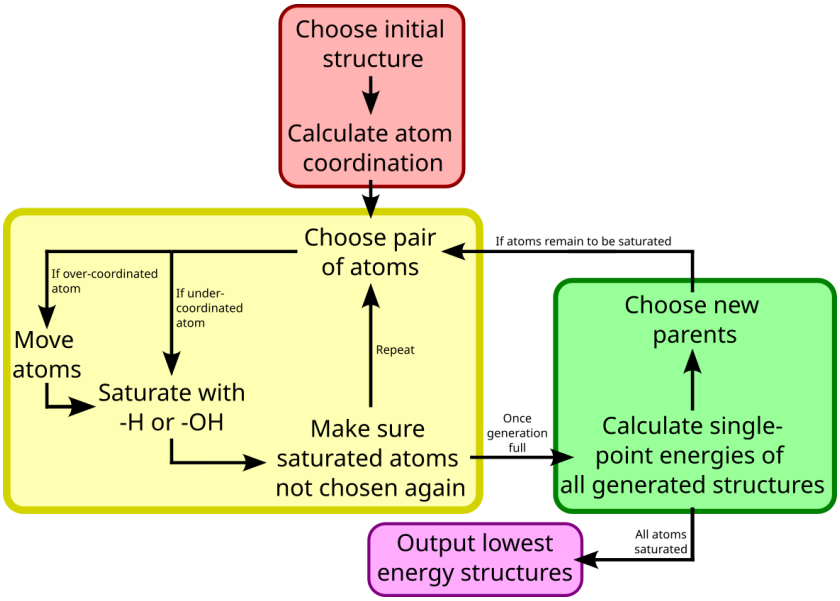


Figure 3.1: Visualization of algorithm for saturating the surface.

as it is the most stable polymorph of SiO_2.

$$\Delta E_{strain} = E_{surface} - n\, E_{\alpha-quartz} - m\, E_{H_2O} \tag{3.3}$$

Strain is an accumulated property. The more atoms and bonds a model has, the more strain a model can build-up. In order to compare models of differing *N*, the strain must be normalized. For this work that was done through taking the strain per Si-O bond ($\Delta E_{per\ Si-O}$). On a conceptual level, this was the easiest way to reason about differences of total strain between models and the magnitude of these differences.

4

Results and Discussion

In this chapter, the results of the gathered data are presented and discussed. The surface roughness and expected values for topological features are compared between varying values of α. Following this, the strain of dry and saturated surfaces is presented. Descriptors are then explored to predict these values of strain. Lastly, the results of screening of semi-empiric DFT are presented.

4.1. Creation of surfaces

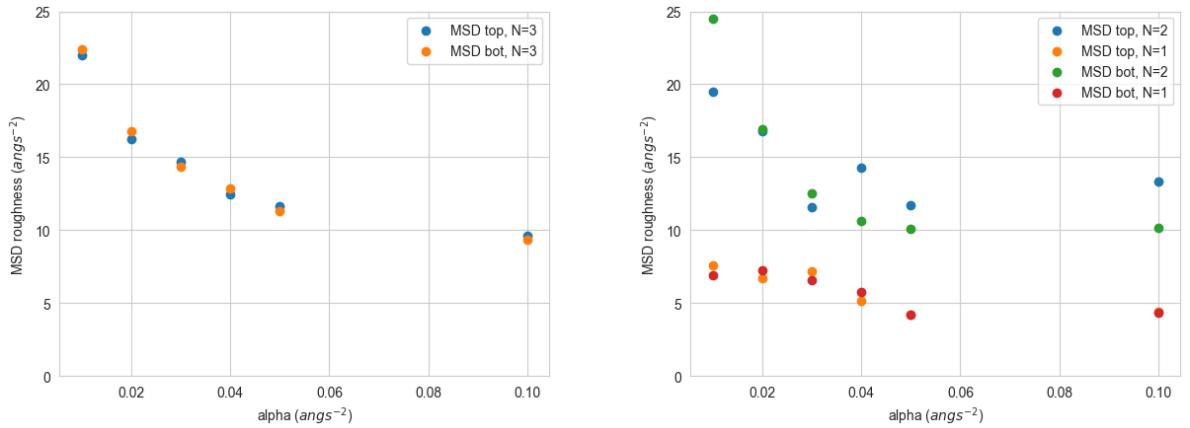
4.1.1. Roughness of Generated Models

Bulk amorphous SiO_2 was made through distorting β-cristobalite crystals tiled three times in the z-direction (*N* = 3) through simulated annealing. These bulks were turned into models for the surface of amorphous SiO_2 through cleaving using a randomly generated stochastic Fourier series through which roughness can be controlled through an empirical parameter (α). As this method is random, multiple models were made to assess what expectation values for given α might be and how well the roughness, as determined through the mean-square deviation (MSD) of the surface is controlled through it. This is measured through equation 4.1 where n is the number of atoms on the surface, $\bar{z}$ is the average z-coordinate of the surface atoms, and $z_i(x, y)$ is the z-coordinate of a given atom i on the surface of the model. Figure 4.1a shows the average MSD for the varying values of α for the top (blue dots) and bottom (Orange dots) surfaces of the generated models. For the models of *N* = 3, the trend of greater α leading to lower average MSD is observed. This decrease looks to follow a logarithmic decrease, exactly as Nguyen and Laird as observed.[24] This means that the implemented method functions as intended. α can be used to directly influence the expected surface roughness when generating models.

$$MSD = \frac{1}{n}\sum_{i=1}^{n}(z_i(x, y) - \bar{z})^2 \tag{4.1}$$

As models of difference sizes were also of interest, more were made starting with different *N* using the same values if α as before. In Figure 4.1b the same trend of increasing alpha leading to decreasing MSD is seen. Furthermore, the same logarithmic trend is observed for *N* = 2 and the resulting values of MSD are similar to that for *N* = 3 but it does not seem to follow it as closely. This could be due to there being one-tenth the models of *N* = 2, and *N* = 1 for that matter, and as such average values maybe not result in exactly the expected trend. Models generated of size 1 have average MSDs which deviate comparably far from the other two values of *N* and have a much less significant decrease in MSD across the range of α. This could be due to how the cleaving function ranges between -10 Å and 10 Å in the z-direction. Models of *N* = 1 unit cell are approximately 7 Å in height and as such a

relatively greater number of atoms are displaced than that of larger models. In other words, models of N = 1 unit-cell are too small to effectively be made using this method as they are cleaved to too great an extent. This could also be argued for models of N = 2 unit-cell however Figure 4.1b provide evidence to the contrary; they still follow the expected trend for MSD roughness.



(a) **(b)**

Figure 4.1: Mean-square displacement roughness compared to alpha of varying initial bulk thicknesses for a) models of N = 3 b) models of N = 1, 2

Analytically, Nguyen and Laird showed that the expected MSD roughness of the cleaving function tends towards 0 as α tends towards infinity.[24] For the surfaces of N = 2 and 3 the MSD looks to plateau around 10 Å^2 as α increases. Amorphous SiO_2 exist as discrete atoms which are usually spaced out by at least 1.6 Å so there will always be some intrinsic surface roughness and deviation from expected MSD. Furthermore, during the re-equilibration after cleaving, the surfaces will reconstruct as to decrease the number of dangling atoms which were induced and this should also decrease the roughness of the model. Regardless, a plateauing at an MSD of 10 Å is arguably high. This could be explained through the cleaving function cutting a nano-pore-like structure into the model which is not closed during re-equilibration. With the probing method used to identify surface atoms, atoms in this hole would be defined as part of the surface.

4.1.2. Topology of Generated Models

Surfaces can not only be characterized through their roughness but also their topology, that is: bond lengths, bond angles, rings, defects. For this reason, these properties were extracted from the models and averaged for given N and α to analyse the effect of these variables. Any differences in average values between differing N is likely because of the relative number of displaced atoms being greater for smaller models and as such more atom recombining during re-equilibration. Furthermore, as observed earlier in the MSD, models of sizes other than 3 likely have too few models to meaningfully make conclusions based on their average values. For these reasons, only analysis done on models of N = 3 will be discussed. Graphs and tables for the remaining values of N can be found in appendix B.

Starting with the hypothesis that lower α will lead to more defects within the model as there are more atoms displaced and dangling bonds induced when cleaving with rougher functions. In this context, a defect is defined as any over/under-coordinated atom remaining in the model after re-equilibration. For Si atoms these are 3- or 5-coordinated atoms (^{3}Si, ^{5}Si) and for O this is 1- or 3-coordinated atoms (^{1}O, ^{3}O). The coordination of a given atom was determined through counting the number of atoms within a 2 Å radius. The average percentage of each defect and the total number of defect Si and O are given in Table 4.1. Between the two types of defects, over-coordination is found to occur more frequently for both atoms. For the total number of defect atoms there is no strong correlation with α. For the most

part is it decreasing however to say that defects are directly related α is a stretch. There is also not a large decrease in the total number of defect atoms as there is a range of 2 Si and 3.5 O between the maximum and minimum of each. These two facts lead to the conclusion that α did not have a strong influence on the rate of defects forming when creating the models.

Table 4.1: Average percentage of defects for surfaces of *N* = 3

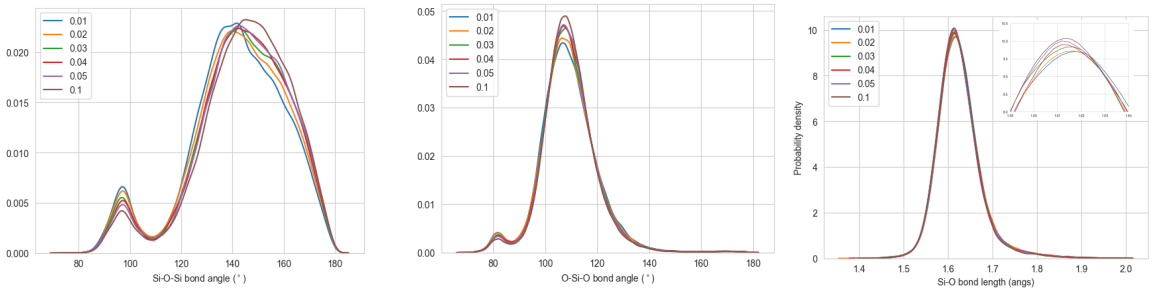
α (Å^{-2})	% ^{3}Si	% ^{5}Si	number Si defects	% ^{1}O	% ^{3}O	number O defects
0.01	1.47	2.55	8.7	2.17	2.71	21.1
0.02	1.50	2.92	9.5	2.21	2.91	22.1
0.03	1.53	2.56	8.8	2.25	2.78	21.7
0.04	1.25	2.69	8.5	1.98	2.70	20.2
0.05	1.29	2.44	8.1	2.08	2.66	20.5
0.1	1.38	2.08	7.5	1.98	2.33	18.6

Not only the number of defects but also their positions relative to other notable surfaces features is interesting. It has previously been suggested that ^{5}Si are generally found in 3-membered rings (3-MR) and occasionally 2-membered rings (2-MR).[65] The reason for this has been attributed to ^{5}Si having a lower energy of formation when part of a strained ring compared to when they are not.[66] The generated models agree as 60% of ^{5}Si are found in 3-MR which increases to 73% when taking those in 2-MR into account (Table B.1). ^{3}Si were found to have higher energies of formation, especially when within strained 2- or 3-MR. Through the same analysis it appears that 7% of ^{3}Si are found participating in 2-MR while 60% are found in rings larger than 4 (Table B.1). These higher energies of formation for ^{3}Si also explain why they are less frequently found in models than ^{5}Si.

Greater fractions of 2- and 3-MR should be identifiable from the distribution of bond angles for a given model.[24] Thus the reasoning that comparing the average distribution of bond angles across α's should give a qualitative indication as to the average number of rings found in models. The distributions of Si-O-Si bond angles for the different values of α is shown in Figure 4.2a and that for O-Si-O in Figure 4.2b. The difference over α is more pronounced for the Si-O-Si with a clear decrease in the secondary peak and a shift in the primary peak to the right. The decrease in the secondary peak has been attributed to an increased fraction of 2-MR and the described shift corresponds to an increased fraction of 3-MR and 4-MR.[67] This agrees with Figure 4.3 which shows the average number of 2-MR and 3-MR decreasing as α increases. As such, the distribution of bond angles is reasonably affected through the rings present in the model and according to the distribution of bond-angles, α has a systematic effect on strained rings present in models.

Beginning with the hypothesis that 3-MR are responsible for increased surface roughness, Figures 4.1 and 4.3 are compared. From Figure 4.3 the amount of 3-MR is seen decreasing in a fashion similar to that of the average MSD with α. It also has a similar effect on the number of 2-MR. The decrease in rings for lower values of α can also be argued to be due to the cleaving itself as lower values will induce less unsaturated atoms on the surface which can recombine and form strained rings during re-equilibration.[67] Thus, the conclusion that one of the mechanisms through which α controls surface roughness is through promoting the formation of strained rings during the re-equilibration after cleaving.

The final topological property of the models which was analyzed was the lengths of Si-O bonds and their relation to α. These probability distributions are plotted in Figure 4.2c. The bond lengths of the generated models look to be distributed normally with mean values close to 1.625 Å. Changes in α



(a) (b) (c)

Figure 4.2: Probability distributions for a) Si-O-Si bond angles (°), b) O-Si-O bond angles (°), c) all Si-O bond lengths (Å), across all models of N = 3

are seen to slightly effect the distribution of bond lengths. Primarily, a very small shift in the maximum of the distribution towards a longer bond length and a small decrease in its probability as α increases. This also leads to a slight decrease in the average bond length for as α decreases. The total change in average bond length from α = 0.01 to 0.1 is a decrease of 0.026 Å. As such, from this data, it is concluded that the roughness of cleaving function used it make the models has little, but still noticeable, impact on the bond lengths of the resulting model.

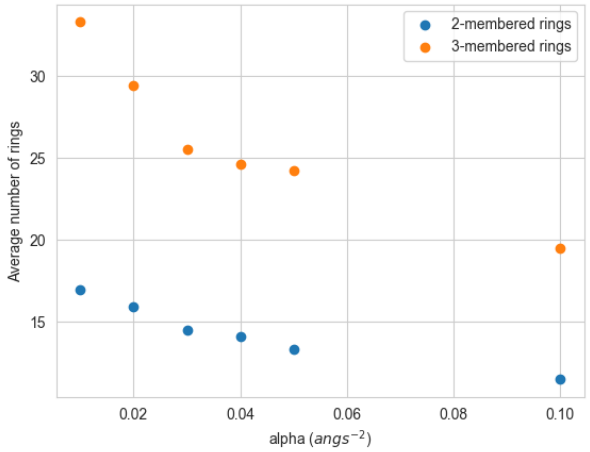


Figure 4.3: Number of two-membered (blue) and three-membered (orange) rings averaged across surfaces of given α

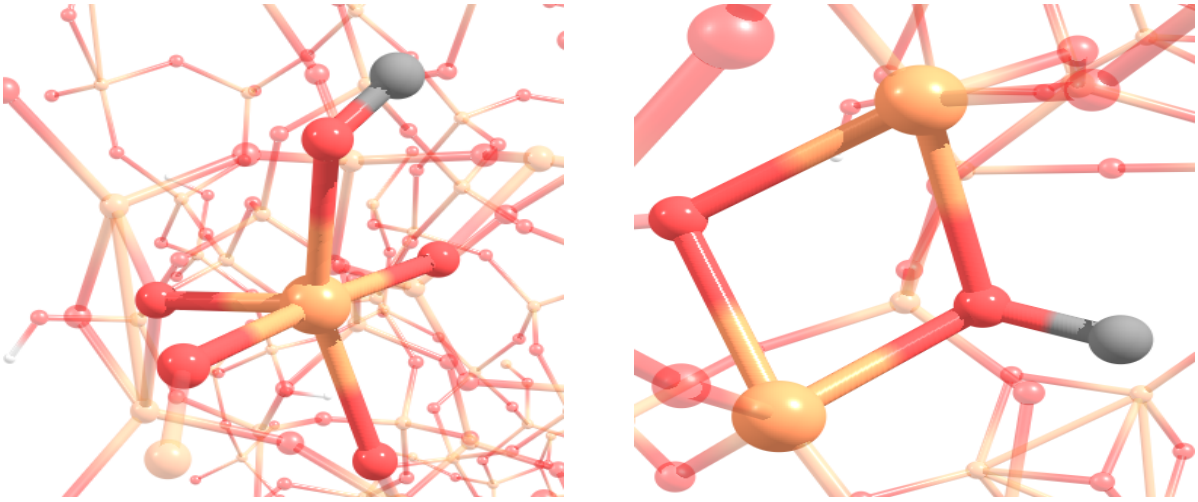
4.1.3. Saturation of Generated Models

For the saturation of models, it was considered ideal to still be able to use CMD. This was seen as an improvement over manually saturating the surfaces and would remove the human-bias when introducing silanol groups through breaking bonds. For this ReaxFF was employed. After running the simulation and removing all molecules not bound to the SiO_2 model, it was found that the majority of the time the resulting surface was positively charged. The hydroxide ions necessary to keep the system charge neutral were generally found to be in the water bulk. Comparing to the results of Nguyen and Laird it was found that their models had the same issue. This is not physically reasonable as naturally occurring solids are not charged on the macroscopic scale. For this reason, the surfaces were saturated algorithmically, as described in section 3.3.

Due to its empiric nature, no meaningful topological analysis can be done on the surfaces immediately after they have been saturated using the algorithm. Moreover, comparing the topologies of structures which are a result of DFT to those of CMD is also not reasonable as they are very different

methods in how they influence the structure of the model. As such, the results of the DFT optimizations only qualitatively analyzed.

Due to the algorithm being designed to try and bring all atoms to their preferred coordination, after DFT optimization, it was expected that over/under-coordinated atoms would no longer be present in the model with perhaps the exception of one or atom. Moving forward with this expectation, the coordination of each atom within the new structure was recalculated and analyzed. The resulting structures generated through the algorithm had almost 0 remaining under-coordinated atoms after geometry optimization. However, the number of ^{5}Si is maintained. The average across all models decreases only by 0.4 after saturation, from 5.5 to 5.1. The occurrence of ^{6}Si within models was also observed for the first time. ^{6}Si are not inherently non-physical as they are naturally occurring within the mineral stishovite.[68] What is of greater interesting is the excess of ^{5}Si within the models. Qualitatively, a fair amount of the ^{5}Si were found to have a -OH group bound to them (Figure 4.4a), some even sharing this group between with another Si atom as seen in Figure 4.4b. This cannot be ruled as inherently non-physical due to the unknown general structure of amorphous SiO_2. However, this occurs relatively frequently with some structures containing this shared-hydroxyl motif in excess of 5 times. It is possible that this is a consequence of the DFT optimization, not the method of the algorithm. For geometry optimization the system is simulated at exactly 0 K so it has no more dynamic movement. As such, different results might be achieved through using simulated annealing in place of geometry optimization and this over-coordination need not be due to the method of saturation. Regardless, the algorithm is effective at removing the energetically less favorable under-coordinated atoms while over-coordinated seem more difficult remove through saturating the model. Considering their relative stability of over-coordinated atom, it might not even be necessary to go beyond removing under-coordinated atoms for a physically reasonable model.



(a) (b)

Figure 4.4: structures of Si (orange), O (red) and H (grey). ^{5}Si found in saturated structures where a) the silanol is attached to the ^{5}Si and b) the silanol is shared between Si atoms.

4.2. Strain

4.2.1. Strain of Dry Surfaces

Much like for surface roughness, a relationship between the strain defined in equation 3.3 and the roughness parameter, α, was also expected. This stems from the realization that α influences the number of strained rings within a model. From Figure 4.5, strain is observed to be inversely proportional

to α. Along with that, the decrease in the mean values of strain seems to decrease similarly to the number of 3-MR and 2-MR. The range of strain for each α spans about 3000 to 4000 kJ/mol and there is a relatively large degree of overlap in these ranges when comparing between α. This brings into mind that if a given strain, or surface topology, is desired then multiple models should be made as the method. This is exemplified through the highest strained surface of α = 0.1 is fairly close to that for α = 0.02.

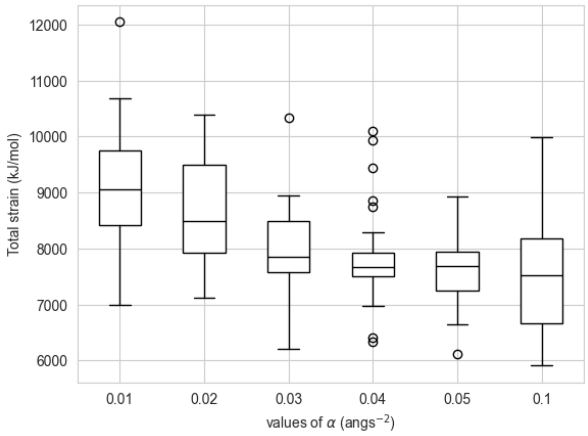


Figure 4.5: Total strain of models *N* = 3 units for given values of α

From the observation that strain follows similar trends to roughness, it was hypothesized that model of *N* = 2 would have similar strain to those of *N* = 3 and *N* = 1 would have less. However, since the models of different sizes they were compared on the basis of their strain per Si-O bond ($\Delta E_{per\ Si-O}$) as strain is an accumulated in every bond and every angles; the more there are the more opportunities to accumulate. In figure 4.6 the different values of $\Delta E_{per\ Si-O}$ shown for the values of *N*. What is initially noticeable is that the average strain inversely proportional to model size, contrary to expectations. The average values of $\Delta E_{per\ Si-O}$ decrease from 16.5 to 12 to around 9 kJ/mol from *N* = 1 to 2 to 3 and 4, respectively. The overlap in the calculated $\Delta E_{per\ Si-O}$ is initially the lower quartile of *N* = 1 and the upper quartile of *N* = 2. For larger *N* the overlap in ranges increases with higher *N* overlapping more with each other. The inverse relationship between *N* and $\Delta E_{per\ Si-O}$ was postulated to be due to the difference in the proportion of atoms displaced during cleaving. As the range of the cleaving function frequently encompasses the full thickness of bulks, especially for *N* = 1 and *N* = 2, smaller models have less atoms which are left unaffected by the cleaving. With ”unaffected” it is meant that an atom is not shifted or given dangling bonds. This greater number of unaffected atoms means that there are less which will have to recombine during re-equilibration and more which will stay in close to the same state as to before cleaving.

With the hypothesis that the higher $\Delta E_{per\ Si-O}$ of smaller models is due to less of the initial amorphous bulk being unaffected by cleaving, an estimate for $\Delta E_{per\ Si-O}$ of amorphous SiO_2 bulk was calculated. As the system being modelled is amorphous, there is no exact way of determining its intrinsic average amount of $\Delta E_{per\ Si-O}$ as there is no long-range periodic order to the structure. We cannot be sure of how the bulk looks like. However, as larger and larger models are considered calculations of $\Delta E_{per\ Si-O}$ should converge to an average value, grated the models are physically reasonable. Under this assumption, the bulk model generated to create the *N* = 3 models was optimized using DFT and it's $\Delta E_{per\ Si-O}$, when compared to the reference states given in equation 3.3, is 0.21 kJ/mol. When comparing this number to the full dataset, it is considerable smaller than the lowest $\Delta E_{per\ Si-O}$ of 5.5 kJ/mol and two orders of magnitude smaller than the largest value. From this it can be concluded that the major contribution to the $\Delta E_{per\ Si-O}$ of the models is the fact that they are no long amorphous bulks and have been given strained features from cleaving.

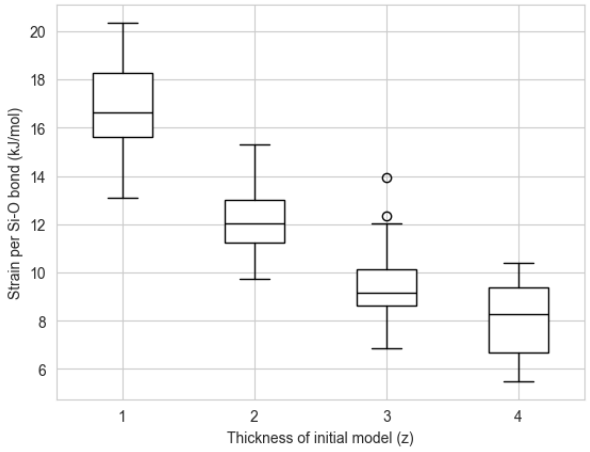


Figure 4.6: Range of $\Delta E_{per\ Si-O}$ for models of sizes *N* = 1, 2, 3, 4 unit cells thick

4.2.2. Strain of Saturated Surfaces

The algorithm made is empirical in the manner which it saturates models and the method of choosing which structures are carried forward for the next generation is the single-point energy as calculated through GFN-FF. For these reasons it may not always effectively create and choose models. In an attempt to assess the performance of the algorithm, a structure of *N* = 3 was chosen at random and the lowest energy structure in each generation had its geometry optimized. After each successive saturation the strain of the surfaces looks to decrease linearly (figure 4.7). On average, each addition of H_2O decreases the $\Delta E_{per\ Si-O}$ by 0.45 kJ/mol. There are deviations from this trend. Namely, after adding 5 H_2O molecules and between 8 and 11 H_2O molecules. Here the strain either increases compared to the previous structure or decreases at double the average amount.

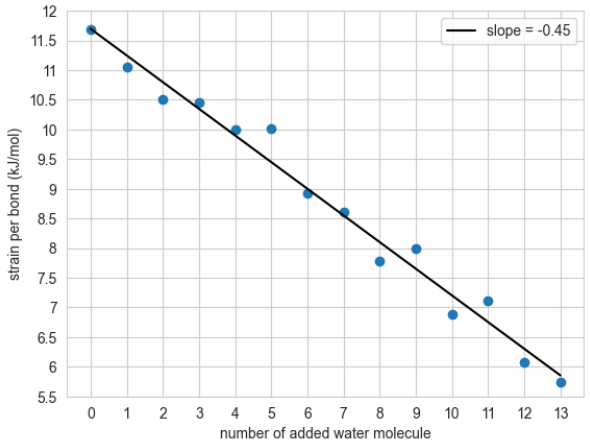
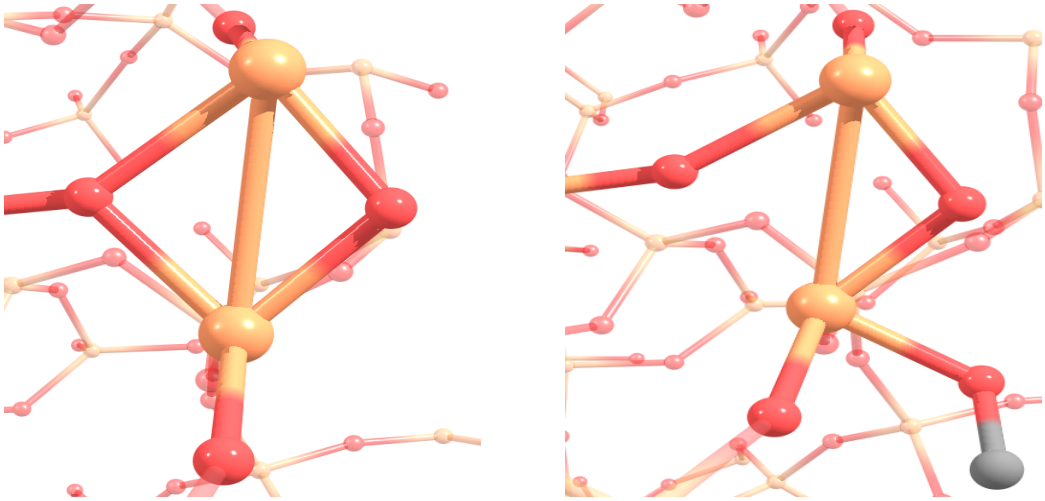


Figure 4.7: $\Delta E_{per\ Si-O}$ of selected structure as the surface is saturated by the algorithm

These increases in energy seem to happen when a 2-MR is broken through the displacement of ^{3}O, as shown in figure 4.8. This results in a remaining bond angle which is typical for the 2-MR but otherwise smaller than a standard a Si-O-Si angle. This would cause the surrounding to relax during the DFT optimization as there is clearly some force driving the two Si of the 2-MR apart. As there is a greater amount of potential energy the GFN-FF single-point calculation resulting in a higher energy. For this reason, it is likely better practice to make use of geometry optimizations when generating structures in place of single-point energy calculations so that the structure is allowed to reach a relaxed state.



(a) (b)

Figure 4.8: 2-MR of a) generation 10, before saturation, and b) generation 11, after saturation. Si (orange), O (red), H (grey)

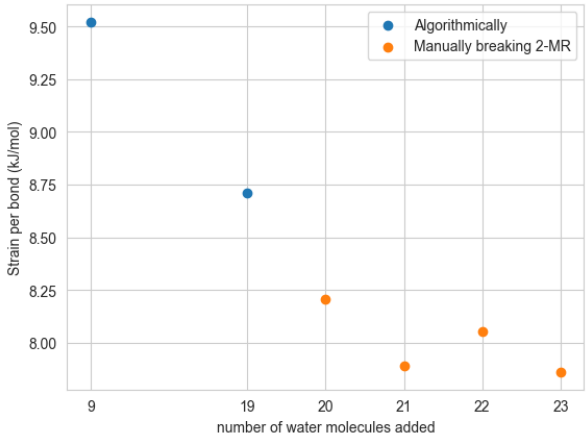


Figure 4.9: $\Delta E_{per\ Si-O}$ after adding 9 and 19 water molecules algorithmically (blue) and manually breaking two-membered rings (orange)

With the algorithm only taking atom coordination into account, it is possible that a lower energy state can easily be achieved through functionalizing strained features such as 2-MR. Thus, the structure of $N = 2$ with the highest strain was chosen and further saturated through randomly adding water across bonds at the bottom of the surface. This was done through slightly extending the capabilities of the already existing algorithm. Ultimately, 10 more H_2O molecules were added on top of the 9 already used to saturate the surface. What was noticed is that the resulting structure decreased $\Delta E_{per\ Si-O}$ by approximately 0.75 kJ/mol after adding all the remaining H_2O (figure 4.9). This is much less than saturating over/under-coordinated atoms. Upon visually inspecting the resulting surface it was found that 2-MR were generally left untouched and there 4 were remaining in the model, all relatively spaced away from each other.

2-MR are causes of higher local strain and they are most likely to be some of the first features that react with water.[30] Thus, the remaining 4 rings were functionalized. One ring was broken at a time and the structure was optimized using DFT. After adding water to the first two 2-MR, the $\Delta E_{per\ Si-O}$ is seen to decrease by a further 0.75 kJ/mol and after the removal of the two remaining rings $\Delta E_{per\ Si-O}$ stays essentially the same at around 7.50 kJ/mol. This result highlights a few things: two possible improvements to the algorithm and not all 2-MR are equally strained. One of the improvements is that

geometry optimization should be used, again. It is rather likely that the 2-MR were left untouched as the remaining structure after functionalizing the Si-O bond of a 2-MR leaves behind a unnaturally strained structure. The other improvement is that, if geometry optimizations are not implemented, the algorithm should be made to prioritize strained rings or they should be manually broken at the minimum. Finally, when breaking bonds, the amount of strain is also affected by the local environment of said bond. If breaking a 2-MR were to always lead to a consistent decrease in $\Delta E_{per\ Si-O}$ they, this would have been seen in Figure 4.9. Instead, there is an increase in $\Delta E_{per\ Si-O}$ after breaking the third ring followed by a decrease which results in a similar value of $\Delta E_{per\ Si-O}$ as before the increase in energy. There is no immediate reason as to why this would be unique to 2-MR and as such the postulation that the local amount of strain it important when considering the strain of a model after a saturation or functionalization is done.

As saturated surfaces are likely the ones which are better representations of the systems which are the interest of study, their strain is arguably of more concern than that of dry surfaces. Predicting the strain of a saturated, or functionalized, model from the strain of the corresponding dry model would be ideal as this would decrease the computational resources required to know if a model of desired strain and roughness is achieved. To see if this prediction of strain from the dry model is possible if the surface is saturated using the algorithm the $\Delta E_{per\ Si-O}$ of saturated surfaces is compared to that of the dry surfaces to see if there is a correlation. This can be seen in Figure 4.10 for surfaces of N = 2 and 3. Here it can be seen that surfaces of higher strain when dry lead to surfaces of higher strain after saturation. The algorithm only considers under/over-coordinated atoms when saturating the surface. As such, bonds within strained features are not broken unless they contain an over-coordinated atom. Thus, any energy in these features is not released and still present after saturation. This makes it logical that there is an approximate relationship between the strain of the dry and saturated surfaces. It also makes it possible to have an estimate of what the strain of a saturated surface will before further processing it.

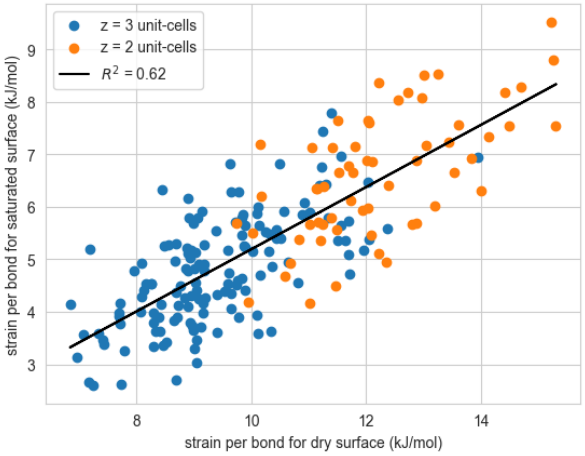


Figure 4.10: $\Delta E_{per\ Si-O}$ of saturated surfaces plotted against strain of corresponding dry surface for N = 2 (orange) and 3 (blue) unit cells

4.2.3. Descriptors for the Strain of Dry Surfaces

When bonds are explicitly defined in CMD their associated energies are frequently described through the harmonic oscillator. Likewise for bond angles. This leads to not calculating an actual energy of a system but a predicted energy penalty when comparing to the equilibrium state of each bond and angle, essentially the strain the system is under. Taking inspiration from this, linear descriptors for $\Delta E_{per\ Si-O}$ in form of the topological features discussed earlier were applied to attempt to predict it. This includes the coordination of atoms, average bond length and average Si-O-Si bond angle. All linear relations

can be found in appendix C. Of the Si and O atoms, the latter seem to generally give a better fit, R^2 = 0.11 and 0.24 for ^{1}O and ^{3}O respectively (Figures C.1, C.2), than that of Si, R^2 = 0.05 and 0.27 for ^{3}Si and ^{5}Si respectively (Figures C.3, C.4). As a whole, none of those linear fits are great. However, there are twice as many O atoms as there are Si atoms in the model so there is also a greater range in values of defect O. Average bond length has an R^2 of 0.11 and has a trend of longer average bond lengths leading to higher $\Delta E_{per\ Si-O}$. The R^2 is lower than expected considering that this is one of the sources of additional energy within the molecular modeling. However, the trend of the fit does make physical sense. Models with longer bond lengths are likely, on average, farther from their equilibrium bond length than those with shorter. The best individual descriptor is the average Si-O-Si bond angle with a R^2 of 0.7 and the fit shows that $\Delta E_{per\ Si-O}$ is inversely proportional to average bond angle. This makes sense as a larger secondary peak would lead to a lower average and larger number of strained rings leads to a larger secondary peak. Using average Si-O-Si bond angles alone to predict strain would likely lead to significant deviations. This is argued from the points at the lower bond angles. Here they deviate from the trend line rather significantly with the regression being 5 kJ/mol off of the true $\Delta E_{per\ Si-O}$. This number alone is not significantly larger however it is per bond. Even for the smaller of models at *N* = 1 this is nearly 1500 kJ/mol away from the true value. So, there is arguably no one descriptor that can predict strain on it own.

Considering there are likely multiple sources of strain within a model, one single descriptor that can accurately predict $\Delta E_{per\ Si-O}$ would imply that there is one major source of it and the remaining contributions are either minimal or get canceled out by one another. As such, a multivariate model would be more fitting. Conceptually, applying a multivariate linear regression in this context can be though of as summing total contributions and deductions from the descriptor to total $\Delta E_{per\ Si-O}$. For this O atoms were chosen as they have a greater range of values which would hopefully allow for better fitting. The average bond length and average bond angle were taken as well. Combining all of these descriptors into one model results in the fit outlined in Table 4.2. The R^2 of the model is 0.925 with it claiming that greater ^{1}O and average bond length leads to greater $\Delta E_{per\ Si-O}$ with the remaining descriptors leading to lower values. All of them have significance within the model below that of the conventional standard p = 0.05. The coefficient associated with the average bond length is three orders of magnitude larger than that of the others. Due to the small range in values average this descriptor can take, small changes imply relatively significant changes in strain. Thus, the conclusion earlier that the difference in average bond lengths was likely incorrect.

Table 4.2: Coefficients, t-values, significance of t-values for multivariate linear regression describing $\Delta E_{per\ Si-O}$ of all dry surfaces involving given independent variables. Original statsmodel reports can be found in figure D.1

R^2 = 0.925

independent variables	coef	t	P|t|
intercept	-302	-7.75	0.00
^{1}O	0.457	14.0	0.00
^{2}O	-0.0181	-14.5	0.00
^{3}O	-0.147	-3.35	0.00
Avg bond length	226	9.34	0.00
Avg Si-O-Si angle	-0.350	-11.3	0.00

To visualize the model, the $\Delta E_{per\ Si-O}$ predicted by this model against that calculated through DFT was plotted. This is seen in Figure 4.11. What is apparent from the figure is that when the model predicts $\Delta E_{per\ Si-O}$ greater than 15 kJ/mol the accuracy diminishes. These data points primarily correspond to models of *N* = 1. This could be due to higher $\Delta E_{per\ Si-O}$ being less represented within the data set, as they are only frequently reached by models of *N* = 1, and as such it is fit to better predict values below this value. As these smaller models are not large enough to be effectively made through the used methodology, the fact that their $\Delta E_{per\ Si-O}$ is not as accurately predicted is not considered an issue. Thus, the conclusion that linear regressions seem to be able to effectively predict $\Delta E_{per\ Si-O}$ when combining multiple topological features of a given model of amorphous SiO_2.

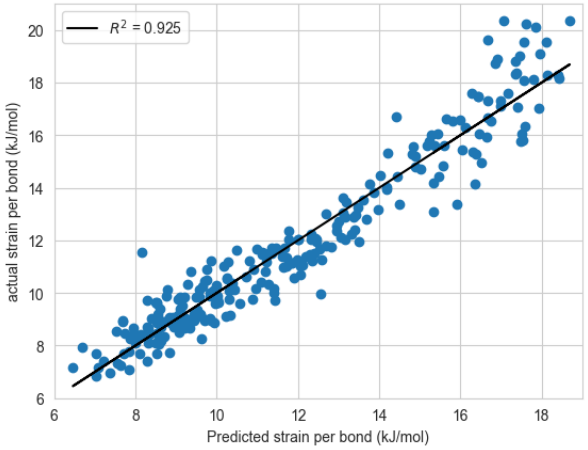


Figure 4.11: Predicted $\Delta E_{per\ Si-O}$ from the multivariate linear regression model against the calculated from DFT $\Delta E_{per\ Si-O}$

Even with linear regressions working effectively, there is no reason for them to necessarily be the most accurate description of the relationship between the descriptors and $\Delta E_{per\ Si-O}$. One example is the average bond length. With chemical bonds being accurately described by the harmonic oscillator, it can be reasonable that a quadratic relationship between it a strain is more effective when combined with the rest. Furthermore, there is no immediately obvious physical logic as to why coordination should be linearly related with $\Delta E_{per\ Si-O}$. ROBERT, an automated machine learning protocol, was applied to the data set for this reason. Doing this, a model of resulting in an R^2 = 0.94 was found using a multivariate linear regression model (Figure D.2). This at least confirms that the model that was initially proposed is satisfactory. Implicitly, it also suggests that all descriptors are best fit to the data linearly. For bond lengths it can be argued that the given range of the values in the dataset is so small that they are essentially linear anyway; given a small enough range any function can be accurately approximated as linear. The data set is unbalanced as there are more models with *N* = 3 unit cells than *N* = 1 and *N* = 2 combined. ROBERT can over-fit to this majority in the dataset. This is seen in the outliers as they consist primarily of models of *N* = 1, noticeably also of higher values of α of 0.04 and 0.05. This is the same as could be seen in Figure 4.11.

With the current multivariate linear regression model appearing to be the most effective way to reasonably predict $\Delta E_{per\ Si-O}$, the values corresponding to the specific values were shuffled. The purpose of this is to see whether they are meaningful to the description the data, together; how important the specific combination of values is to the prediction. If the R^2 significantly drops significantly then the value of every descriptor matters to the model. Doing this shuffling does cause the R^2 to reach a value close to 0 (figure D.3). With this it is concluded that the multivariate linear regression model is definitely using topological descriptors of a surface to predict it's $\Delta E_{per\ Si-O}$.

What is worthwhile to discuss is the descriptor of ^{2}O. As seen in figure 4.12, the number of ^{2}O are very heavily clustered just bellow the values 432, 288, and 144. These are the number of O atoms that each size of model contains. While the R^2 is 0.82 that value itself is meaningless due ^{2}O, in essence almost, being a categorical variable and alone having a very clear indication of $\Delta E_{per\ Si-O}$. However, from Figure 4.6 it is very apparent that there is a fairly broad range in the possible values of $\Delta E_{per\ Si-O}$ for a given *N*. Consequently, including ^{2}O may give a better fit however it is arguably unreasonable to use it for this data set due to the rigid model sizes. This reasoning is partially supported when applying the same shuffling mentioned previously and the value of R^2 staying above 0.8 (D.4). Before definitively taking ^{2}O as a descriptor more intermediate model sizes would need to be added to the set of data.

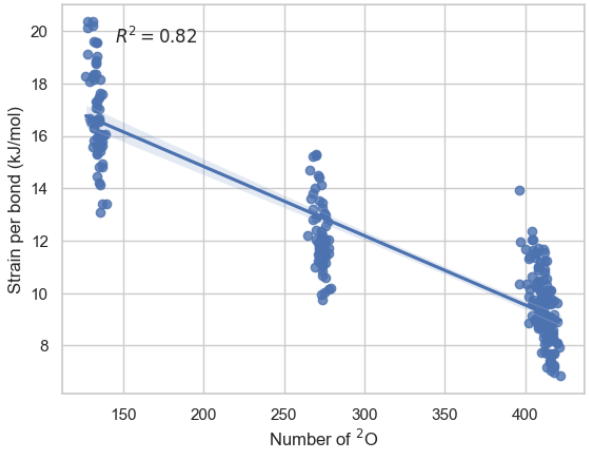


Figure 4.12: Linear trend between the number of average bond angle and the strain per bond for the given model.

As ^{2}O may be leading to an artificial improvement in the model, it was removed to see what the impact on the model would be. Doing this resulted in a model with an R^2 of 0.86. The trends in the coefficients stay the same (figure D.5) and the resulting R^2 is still considered more than acceptable. This categorization, in combination with the $\Delta E_{per\ Si-O}$ barely overlapping between different *N* (Figure 4.6), leads to the argument that the data set should not be analyzed together as it has until now but separately and that the resulting R^2 is not the most accurate measure of performance.

Because of the observation that ^{2}O is rather categorical, the different *N* were analyzed separately in the same fashion to see if the previous results could still be replicated. As a result, the new multivariate linear regression had values of R^2 between 0.74 and 0.8 (table 4.3) and shuffling descriptors again gave R^2 close to 0 (figures D.10, D.11, D.12). This is still an acceptable explained variance however the significance of each variable in each fit is not always above 0.05 and the values of the coefficients are, intuitively, not always physical. For model of 1 and 2 unit cells thick there are no issues, the same descriptors can be used as they are. However, for surfaces of 3 unit cells thick the significance of three of the five descriptors rises to above 0.5. Removing any of the three descriptors based on coordination of O atoms fixes this, and regardless of choice, the R^2 comes out to 0.77. The coefficients of the new regressions do not all share the same trends. There are negative coefficients for the descriptor of ^{1}O for surface of 1 and 2 thickness and a positive coefficient for ^{2}O for surfaces of 3 thickness. This leads to the conclusion that the method is not always strictly physical. Intuitively, ^{1}O would not be considered a species lower in energy that that of an O in α-quartz due to its charge. However, this does not necessarily mean that it is incapable of predicting strain and does not detract from the proposal that topological features can be used to predict the strain of a given structure.

Table 4.3: Coefficients for multivariate linear regressions describing $\Delta E_{per\ Si-O}$ of dry surfaces of specified thickness involving given independent variables. Original statsmodel reports can be found in appendix D

		coeff					
N	R^2	intercept	^{1}O	^{2}O	^{3}O	Avg bond length	Avg Si-O-Si angle
1	0.74	-0.0426	-1.14	-2.30	-2.70	235	-0.280
2	0.807	-0.0161	-1.03	-1.68	-1.93	336	-0.387
3	0.773	-321	0.108[1]	-0.270[1]	-0.404[1]	299	-0.280
3[2]	0.773	-494	0.510	0.113	-	298	-0.282

[1] Independent variable has p-value above 0.5
[2] Descriptor ^{3}O removed from regression

4.2.4. Screening Semi-empirical DFT

Another method of predicting the strain would be to use cheaper calculations to determine electronic energy such as DFTB. Before doing this is must be known if the given method and parametrization corresponds to DFT data. In figures 4.13 the correlation between the strain as calculated by DFTB and DFT has an R^2 of 0.91 for surfaces of size N = 2 and N = 3 unit cells. DFTB seems to systematically under-estimate the strain of a surface. However, the goal is not to get the exact value of strain but rather to be able to predict the relative strain between two or more surfaces. Considering these calculations are comparably cheap, they could be used in place of descriptors.

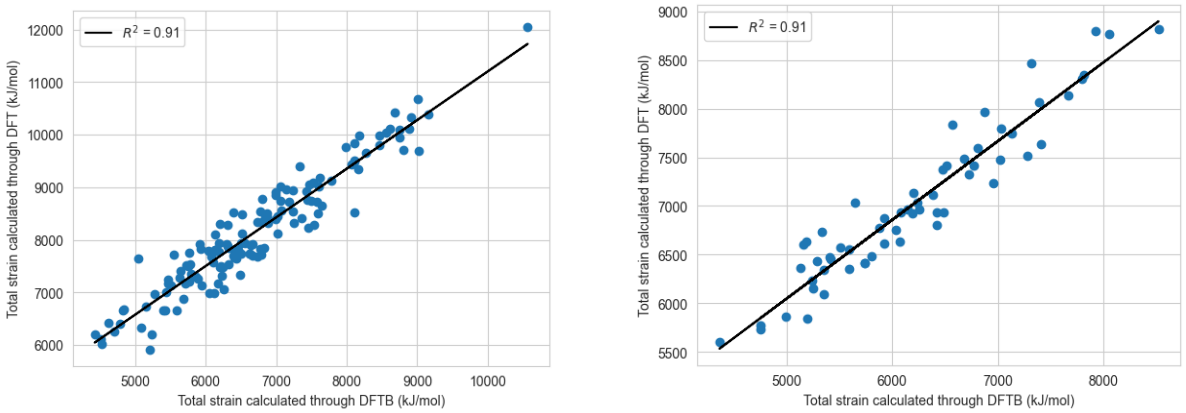


Figure 4.13: Correlation between DFTB calculated electronic energy and DFT calculated electronic energy for dry surfaces of a) 3 units cells and b) 2 unit cells thick

This approach can also be applied to saturated surfaces. Again, DFBT seems to systematically under-estimate the strain of saturated surfaces as well (figures 4.14. The correlation is also not as great as that of the dry surfaces but nonetheless adequate at R^2 of 0.83 for both. With semi-empirical DFT providing an adequate degree of accuracy this makes it possible to apply computationally more expensive tasks, such as simulated annealing using non-force field method to calculate force, when creating surfaces.

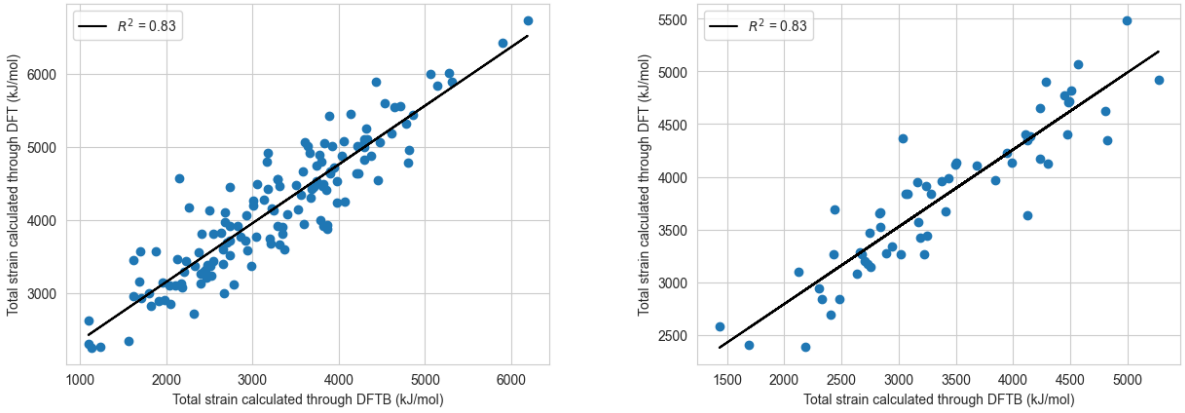


Figure 4.14: Correlation between DFTB calculated electronic energy and DFT calculated electronic energy for saturated surfaces of a) 3 units-cells and b) 2 unit cells thick

5

Conclusion

5.1. Conclusion

The goal of this study was to create and characterize models for the surface of amorphous SiO_2. Surfaces were generated using classical molecular dynamics to distort crystals of β-cristobalite through simulated annealing. These distorted crystals were turned into surfaces by cleaving using a random stochastic Fourier expansion, for which surface roughness can be controlled through the empiric parameter α, and re-equilibrating the system. This was done for models of varying thickness in the z-direction (*N*). The resulting surface roughness and topology of the models was analyzed and compared to previous studies to rationalize results and confirm the method was implemented properly. To introduce silanol groups, models were saturated algorithmically as using force fields led to non-physical, charged models. The strain of the dry and saturated surfaces was compared and the effect of varying *N* on strain was also analyzed. Using the topological characteristics and features of the surfaces, descriptors were then explored using multivariate linear regression and automated machine-learning programs to find relations which could be used to predict the strain of a given model. To see if these descriptor models were the best method of predicting strain, the accuracy of DFTB was also determined.

From the generated models it was concluded that the proposed method of cleaving does work and is generally applicable various sizes of models. They were observed to follow a logarithmic decrease in mean-squared displacement, an indicator of surface roughness, as α got larger. However, below a *N* of 1 the cleaving function displaces too many atoms for there to be any meaningful difference in surface roughness after re-equilibration. This was characterized with the aforementioned logarithmic decrease no longer being as apparent. Using lower values of α led to models with more strained features, 2-MR and 3-MR. But, the amount of defect atoms was not greatly effected as they decrease by 3.5 and 2, for O and Si respectively. Average bond length was found to vary almost negligibly with α, a range of 0.026 Å, but was shown to have a significant impact when describing strain. In general, lower values of α led to higher values of strain however the range of possible values for a given roughness was approximately 3000 to 4000 kJ/mol leading to large overlaps between the values of α. From this the conclusion that multiple surfaces should be generated when looking to create a model of specific strain and surface topology. Surface size was also found to greatly impact the amount of strain within a model when applying the method outlined in this work with the average strain per Si-O bond decreasing from 16.5 to 12 to 9 kJ/mol going from *N* = 1 to 2 to 3 or 4.

Saturating defect atoms led to strains which generally correlate with that of the dry surfaces, R^2 of 0.62. The total number of under-coordinated atoms decreased to essentially 0 however there were still 5.1 over-coordinated Si being present in a model, on average. This was explained through over-

coordinated Si being energetically more stable than their under-coordinated counter parts and the conclusion that it might not even be necessary to algorithmically remove over-coordinated atoms. Further saturation of the surfaces showed that it can continue to decrease strain but simply adding more water will not necessarily lead to a lower strain. Rather, what is saturated is important. Manually saturating 2 2-MR decreased strain per bond by 0.75 kJ/mol. Thus was concluded that, indeed, breaking strained features such as 2-MR is crucial when wishing to decrease strain within a model.

It was shown that strain can be predicted purely from the topology of a model. Using ^{1}O, ^{2}O, ^{3}O, average bond length, and average bond angle as descriptors resulted in an R^2 of 0.925, p-values all well below 0.05, when all data was analyzed together. It was concluded analyzing all models of all values of *N* together might be a wrong choice due to the descriptor of ^{2}O almost being categorical. However, even removing this descriptor, the linear model maintained a relatively high R^2 of 0.86. Analyzing each value of *N* separately led to R^2 between 0.74 and 0.8. Regardless of how the data set was analyzed, shuffling the values of the descriptors between the given values of strain led to values of R^2 close to 0. To check for non-linear relations between the descriptors and strain, machine-learning was applied to the data set. The best model was still one of multi-variate linear regression. This concludes that linear models are an effective way of describing the relationship between said descriptors and a model's strain.

5.2. Outlook

Based on the results and conclusion of this research the following suggestions are thought to be possibilities for further research:

- Stemming from industrial applications, amorphous aluminosilicates are also popular catalyst supports. While this method has been developed for amorphous SiO_2 there is nothing in inherently restrictive in the model of surfaces it may be able to generate. This would entail using force fields which are able to describe the structure of aluminosilicates or developing a method of converting the currently generated models. This also poses a question of how to describe strain in the sense that there is no immediate simple unit cell of an aluminosilicate to which the energy of the surface can be compared. Undertaking this would lead to a method which is able to different classes of models for which the method may be more generalizable to amorphous structures.
- The current set of data is relatively rigid in the size of models it contains. Expanding this to not strictly conform to multiples of the initial unit cell would allow for more analysis which could lead to a model, like the one proposed, which is more generally able to predict the strain of a surface.
- The current method of describing strain is relatively simplistic. While it works as proof of concept, it may not be readily applicable to models of all sizes. To remedy this, a more complex model could be developed perhaps based on the categorization of Si atoms. Instead of simply taking the coordination of the Si atom more in-depth analysis could be done categorize them, individually, according to how strained they are and perhaps to the strain of their local environment. Doing this could also lead the way to describing the local strain of a given area of a model instead of purely the global strain as this work does now.
- General improvements to the algorithm used for saturation could also be made. The ones which are most readily apparent is improving the capability of the model when breaking bonds. Specifically, have it focus on strained features instead of randomly selecting atoms on the surface. Furthermore, the improvement of how it determines the energies of the models it generates could be done either through the annealing of the structures or simply using geometry optimizations over single-point energy calculations when evaluating the energy of a child.

6

Acknowledgements

I would like to acknowledge the use of computational resources of the DelftBlue supercomputer, provided by Delft High Performance Computing Centre (https://www.tudelft.nl/dhpc) and thank SURF (www.surf.nl) for the support in using the National Supercomputer Snellius.

This project was performed as a final thesis at the Inorganic systems engineering group at TU Delft to obtain a degree of Bachelor of Science. I would like to state by thanking the group as a whole and everyone who was there during my time with the group.

I would like to specifically my daily supervisor: Dr. Alexander ”Sasha” Kolganov. For if no other reason, he was willing to put up with my with bi-weekly crisis when data wasn't making sense or literature wasn't producing the results it claims it should. In addition, for asking me to come and do another research project with the group after LO2 - not that it went terribly - its just nice to have not had to search too hard for a project that would interest me. Finally, for letting me take this project in whatever direction seemed logical or interesting to me and being supportive of my use of CPU-hours. I would also like to thank MSc. Adarsh Kalikadien. He may not have been my daily supervisor but he was always willing to help whenever I found myself in too deep with data science and helping me make sense of the machine-learning I was doing (and sometimes not doing 100% correctly). Finally, I would also like to thank Prof. Dr. Evgeny Pidko for having me and for his enthusiasm during the project. Furthermore, I would like to thank him for the meetings that were had, even if I did get lost during them a couple of times, and for the insightful feedback that he provided on all facets of the project.

I would like to finish by thanking the fellow bachelor students I shared an office with. David, Joyce, Nina, it was fun all stressing together near the end of all this and when it was our week to present at the group meeting.

Bibliography

[1] M. Caricato, “Cluster model simulations of metal-doped amorphous silicates for heterogeneous catalysis,” *The Journal of Physical Chemistry C*, vol. 125, no. 50, pp. 27509–27519, 2021.

[2] B. R. Goldsmith, B. Peters, J. K. Johnson, B. C. Gates, and S. L. Scott, “Beyond ordered materials: Understanding catalytic sites on amorphous solids,” *ACS Catalysis*, vol. 7, no. 11, pp. 7543–7557, 2017.

[3] A. Comas-Vives, “Amorphous sio2 surface models: energetics of the dehydroxylation process, strain, ab initio atomistic thermodynamics and ir spectroscopic signatures,” *Phys. Chem. Chem. Phys.*, vol. 18, pp. 7475–7482, 2016.

[4] A. García, “The pseudopotential concept - siesta web page,” Oct 2023.

[5] E. Scrocco and J. Tomasi, “Electronic molecular structure, reactivity and intermolecular forces: An euristic interpretation by means of electrostatic molecular potentials,” *Advances in Quantum Chemistry*, vol. 11, pp. 115–193, 1978.

[6] S. R. Gadre, C. H. Suresh, and N. Mohan, “Electrostatic potential topology for probing molecular structure, bonding and reactivity,” *Molecules*, vol. 26, no. 11, 2021.

[7] D. E. Needham, I. C. Wei, and P. G. Seybold, “Molecular modeling of the physical properties of alkanes,” *Journal of the American Chemical Society*, vol. 110, no. 13, pp. 4186–4194, 1988.

[8] C. J. Cramer, *Essentials of computational chemistry: Theories and Models*. Wiley, 2 ed., 2013.

[9] H. Hauptman, “The direct methods of x-ray crystallography,” *Science*, vol. 233, no. 4760, pp. 178–183, 1986.

[10] K. Persson, “Materials data on sio2 (sg:152) by materials project,” 7 2014. An optional note.

[11] K. Persson, “Materials data on sio2 (sg:227) by materials project,” 7 2014. An optional note.

[12] T. Armbruster and M. E. Gunter, “Crystal Structures of Natural Zeolites,” *Reviews in Mineralogy and Geochemistry*, vol. 45, pp. 1–67, 01 2001.

[13] M. D. Lavender, “The importance of silica to the modern world,” *Indoor and Built Environment*, vol. 8, no. 2, pp. 89–93, 1999.

[14] P. G. Jeelani, P. Mulay, R. Venkat, and C. Ramalingam, “Multifaceted application of silica nanoparticles. a review,” *Silicon*, vol. 12, p. 1337–1354, Jul 2019.

[15] S. Soled, “Silica-supported catalysts get a new breath of life,” *Science*, vol. 350, no. 6265, pp. 1171–1172, 2015.

[16] Q. Sun, N. Wang, Q. Xu, and J. Yu, “Nanopore-supported metal nanocatalysts for efficient hydrogen generation from liquid-phase chemical hydrogen storage materials,” *Advanced Materials*, vol. 32, no. 44, p. 2001818, 2020.

[17] N. Martín and F. G. Cirujano, “Organic synthesis of high added value molecules with mof catalysts,” *Org. Biomol. Chem.*, vol. 18, pp. 8058–8073, 2020.

[18] Y. Bouhoute, D. Grekov, K. C. Szeto, N. Merle, A. De Mallmann, F. Lefebvre, G. Raffa, I. Del Rosal, L. Maron, R. M. Gauvin, L. Delevoye, and M. Taoufik, “Accessing realistic models for the wo3–sio2 industrial catalyst through the design of organometallic precursors,” *ACS Catalysis*, vol. 6, no. 1, pp. 1–18, 2016.

[19] C. S. Ewing, A. Bagusetty, E. G. Patriarca, D. S. Lambrecht, G. Veser, and J. K. Johnson, “Impact of support interactions for single-atom molybdenum catalysts on amorphous silica,” *Industrial & Engineering Chemistry Research*, vol. 55, no. 48, pp. 12350–12357, 2016.

[20] F. Núñez-Zarur, X. Solans-Monfort, and A. Restrepo, “Mechanistic insights into alkane metathesis catalyzed by silica-supported tantalum hydrides: A dft study,” *Inorganic Chemistry*, vol. 56, no. 17, pp. 10458–10473, 2017. PMID: 28809544.

[21] O. Staples, M. S. Ferrandon, G. P. Laurent, U. Kanbur, A. J. Kropf, M. R. Gau, P. J. Carroll, K. McCullough, D. Sorsche, F. A. Perras, M. Delferro, D. M. Kaphan, and D. J. Mindiola, “Silica supported organometallic iri complexes enable efficient catalytic methane borylation,” *Journal of the American Chemical Society*, vol. 145, no. 14, pp. 7992–8000, 2023. PMID: 36995316.

[22] B. Peters and S. L. Scott, “Single atom catalysts on amorphous supports: A quenched disorder perspective,” *The Journal of Chemical Physics*, vol. 142, p. 104708, 03 2015.

[23] F. Tielens, M. Gierada, J. Handzlik, and M. Calatayud, “Characterization of amorphous silica based catalysts using dft computational methods,” *Catalysis Today*, vol. 354, pp. 3–18, 2020. SI: Fascinating catalysis.

[24] N. P. Nguyen and B. B. Laird, “Generation of amorphous silica surfaces with controlled roughness,” *The Journal of Physical Chemistry A*, vol. 127, no. 46, pp. 9831–9841, 2023. PMID: 37938899.

[25] S. H. Lee, R. J. Stewart, H. Park, S. Goyal, V. Botu, H. Kim, K. Min, E. Cho, A. R. Rammohan, and J. C. Mauro, “Effect of nanoscale roughness on adhesion between glassy silica and polyimides: A molecular dynamics study,” *The Journal of Physical Chemistry C*, vol. 121, no. 44, pp. 24648–24656, 2017.

[26] P. N. Wimalasiri, N. P. Nguyen, H. S. Senanayake, B. B. Laird, and W. H. Thompson, “Amorphous silica slab models with variable surface roughness and silanol density for use in simulations of dynamics and catalysis,” *The Journal of Physical Chemistry C*, vol. 125, no. 42, pp. 23418–23434, 2021.

[27] A. C. T. van Duin, S. Dasgupta, F. Lorant, and W. A. Goddard, “Reaxff: A reactive force field for hydrocarbons,” *The Journal of Physical Chemistry A*, vol. 105, no. 41, pp. 9396–9409, 2001.

[28] C.-L. Kuo, S. Lee, and G. S. Hwang, “Strain-induced formation of surface defects in amorphous silica: A theoretical prediction,” *Phys. Rev. Lett.*, vol. 100, p. 076104, Feb 2008.

[29] T. A. Michalske and B. C. Bunker, “Slow fracture model based on strained silicate structures,” *Journal of Applied Physics*, vol. 56, pp. 2686–2693, 11 1984.

[30] J. Yeon and A. van Duin, “Reaxff molecular dynamics simulations of hydroxylation kinetics for amorphous and nano-silica structure, and its relations with atomic strain energy,” *The Journal of Physical Chemistry C*, vol. 120, 12 2015.

[31] R. Santamaria, *Molecular dynamics*. Springer, 1 ed., 2023.

[32] I. Newton, I. Cohen, A. Whitman, and J. Budenz, *The Principia: Mathematical Principles of Natural Philosophy*. University of California Press, 1999.

[33] J. E. Jones and S. Chapman, “On the determination of molecular fields.—i. from the variation of the viscosity of a gas with temperature,” *Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character*, vol. 106, no. 738, pp. 441–462, 1924.

[34] R. A. Buckingham and J. E. Lennard-Jones, “The classical equation of state of gaseous helium, neon and argon,” *Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences*, vol. 168, no. 933, pp. 264–283, 1938.

[35] W. C. Swope, H. C. Andersen, P. H. Berens, and K. R. Wilson, “A computer simulation method for the calculation of equilibrium constants for the formation of physical clusters of molecules: Application to small water clusters,” *The Journal of Chemical Physics*, vol. 76, pp. 637–649, 01 1982.

[36] D. A. McQuarrie, *Statisical mechanics*. University Science Books, 2 ed., 2000.

[37] H. J. C. Berendsen, J. P. M. Postma, W. F. van Gunsteren, A. DiNola, and J. R. Haak, “Molecular dynamics with coupling to an external bath,” *The Journal of Chemical Physics*, vol. 81, pp. 3684–3690, 10 1984.

[38] Y. Nourani and B. Andresen, “A comparison of simulated annealing cooling strategies,” *Journal of Physics A: Mathematical and General*, vol. 31, p. 8373–8385, Oct 1998.

[39] E. Schrödinger, “Quantisierung als eigenwertproblem,” *Annalen der Physik*, vol. 384, no. 4, pp. 361–376, 1926.

[40] M. Born and R. Oppenheimer, “Zur quantentheorie der molekeln,” *Annalen der Physik*, vol. 389, no. 20, pp. 457–484, 1927.

[41] P. Hohenberg and W. Kohn, “Inhomogeneous electron gas,” *Phys. Rev.*, vol. 136, pp. B864–B871, Nov 1964.

[42] W. Kohn and L. J. Sham, “Self-consistent equations including exchange and correlation effects,” *Phys. Rev.*, vol. 140, pp. A1133–A1138, Nov 1965.

[43] N. E. Henriksen and F. Y. Hansen, *Theories of Molecular Reaction Dynamics: The Microscopic Foundation of Chemical Kinetics*. Oxford University Press, 2 ed., 2019.

[44] J. P. Perdew and K. Schmidt, “Jacob’s ladder of density functional approximations for the exchange-correlation energy,” *AIP Conference Proceedings*, vol. 577, pp. 1–20, 07 2001.

[45] T. D. Kühne, M. Iannuzzi, M. Del Ben, V. V. Rybkin, P. Seewald, F. Stein, T. Laino, R. Z. Khaliullin, O. Schütt, F. Schiffmann, D. Golze, J. Wilhelm, S. Chulkov, M. H. Bani-Hashemian, V. Weber, U. Borštnik, M. Taillefumier, A. S. Jakobovits, A. Lazzaro, H. Pabst, T. Müller, R. Schade, M. Guidon, S. Andermatt, N. Holmberg, G. K. Schenter, A. Hehn, A. Bussy, F. Belleflamme, G. Tabacchi, A. Glöß, M. Lass, I. Bethune, C. J. Mundy, C. Plessl, M. Watkins, J. VandeVondele, M. Krack, and J. Hutter, “CP2K: An electronic structure and molecular dynamics software package - Quickstep: Efficient and accurate electronic structure calculations,” *The Journal of Chemical Physics*, vol. 152, p. 194103, 05 2020.

[46] J. VandeVondele, M. Krack, F. Mohamed, M. Parrinello, T. Chassaing, and J. Hutter, “Quickstep: Fast and accurate density functional calculations using a mixed gaussian and plane waves approach,” *Computer Physics Communications*, vol. 167, no. 2, pp. 103–128, 2005.

[47] S. Grimme, J. Antony, S. Ehrlich, and H. Krieg, “A consistent and accurate ab initio parametrization of density functional dispersion correction (DFT-D) for the 94 elements H-Pu,” *The Journal of Chemical Physics*, vol. 132, p. 154104, 04 2010.

[48] S. Grimme, S. Ehrlich, and L. Goerigk, “Effect of the damping function in dispersion corrected density functional theory,” *Journal of Computational Chemistry*, vol. 32, no. 7, pp. 1456–1465, 2011.

[49] M. P. Johansson, V. R. I. Kaila, and D. Sundholm, *Ab Initio, Density Functional Theory, and Semi-Empirical Calculations*, pp. 3–27. Totowa, NJ: Humana Press, 2013.

[50] M. Elstner, D. Porezag, G. Jungnickel, J. Elsner, M. Haugk, T. Frauenheim, S. Suhai, and G. Seifert, “Self-consistent-charge density-functional tight-binding method for simulations of complex materials properties,” *Phys. Rev. B*, vol. 58, pp. 7260–7268, Sep 1998.

[51] D. Porezag, T. Frauenheim, T. Köhler, G. Seifert, and R. Kaschner, “Construction of tight-binding-like potentials on the basis of density-functional theory: Application to carbon,” *Phys. Rev. B*, vol. 51, pp. 12947–12957, May 1995.

[52] D. Dalmau and J. V. Alegre Requena, "Robert: Bridging the gap between machine learning and chemistry," *ChemRxiv*, 2023. This content is a preprint and has not been peer-reviewed.

[53] ROBERT Project, "ROBERT Documentation." `https://robert.readthedocs.io/en/latest/`, 2024. Accessed: 2024-06-12.

[54] A. P. Thompson, H. M. Aktulga, R. Berger, D. S. Bolintineanu, W. M. Brown, P. S. Crozier, P. J. in 't Veld, A. Kohlmeyer, S. G. Moore, T. D. Nguyen, R. Shan, M. J. Stevens, J. Tranchida, C. Trott, and S. J. Plimpton, "LAMMPS - a flexible simulation tool for particle-based materials modeling at the atomic, meso, and continuum scales," *Comp. Phys. Comm.*, vol. 271, p. 108171, 2022.

[55] P. P. Ewald, "Die berechnung optischer und elektrostatischer gitterpotentiale," *Annalen der Physik*, vol. 369, no. 3, pp. 253–287, 1921.

[56] L. Martínez, R. Andrade, E. G. Birgin, and J. M. Martínez, "Packmol: A package for building initial configurations for molecular dynamics simulations," *Journal of Computational Chemistry*, vol. 30, no. 13, pp. 2157–2164, 2009.

[57] J. P. Perdew, K. Burke, and M. Ernzerhof, "Generalized gradient approximation made simple," *Phys. Rev. Lett.*, vol. 77, pp. 3865–3868, Oct 1996.

[58] S. Goedecker, M. Teter, and J. Hutter, "Separable dual-space gaussian pseudopotentials," *Phys. Rev. B*, vol. 54, pp. 1703–1710, Jul 1996.

[59] C. Hartwigsen, S. Goedecker, and J. Hutter, "Relativistic separable dual-space gaussian pseudopotentials from h to rn," *Phys. Rev. B*, vol. 58, pp. 3641–3662, Aug 1998.

[60] J. VandeVondele and J. Hutter, "Gaussian basis sets for accurate calculations on molecular systems in gas and condensed phases," *The Journal of Chemical Physics*, vol. 127, p. 114105, 09 2007.

[61] L. Guimarães, A. N. Enyashin, J. Frenzel, T. Heine, H. A. Duarte, and G. Seifert, "Imogolite nanotubes: Stability, electronic, and mechanical properties," *ACS Nano*, vol. 1, no. 4, pp. 362–368, 2007. PMID: 19206688.

[62] S. Wells, S. Menor, B. Hespenheide, and M. Thorpe, "Constrained geometric simulation of diffusive motion in proteins," *Physical biology*, vol. 2, pp. S127–36, 12 2005.

[63] S. Spicher and S. Grimme, "Robust atomistic modeling of materials, organometallic, and biochemical systems," *Angewandte Chemie International Edition*, vol. 59, no. 36, pp. 15665–15673, 2020.

[64] C. Bannwarth, E. Caldeweyher, S. Ehlert, A. Hansen, P. Pracht, J. Seibert, S. Spicher, and S. Grimme, "Extended tight-binding quantum chemistry methods," *WIREs Computational Molecular Science*, vol. 11, no. 2, p. e1493, 2021.

[65] X. Yuan and A. Cormack, "Local structures of md-modeled vitreous silica and sodium silicate glasses," *Journal of Non-Crystalline Solids*, vol. 283, no. 1, pp. 69–87, 2001.

[66] P. Dagenais, L. Lewis, and S. Roorda, "Dominant structural defects in amorphous silicon," *Journal of Physics: Condensed Matter*, vol. 27, p. 345004, Sep 2015. Epub 2015 Aug 3.

[67] J. Du and A. N. Cormack, "Molecular dynamics simulation of the structure and hydroxylation of silica glass surfaces," *Journal of the American Ceramic Society*, vol. 88, no. 9, pp. 2532–2539, 2005.

[68] J. Finster, "Sio2 in 6:3 (stishovite) and 4:2 co-ordination—characterization by core level spectroscopy (xps/xaes)," *Surface and Interface Analysis*, vol. 12, no. 5, pp. 309–314, 1988.

A

Declaration of use of AI

For this Thesis AI was used as a replacement for documentation of python modules. No text within this thesis, nor figure shown, nor code written was generated using AI.

B

Remaining Topological Information

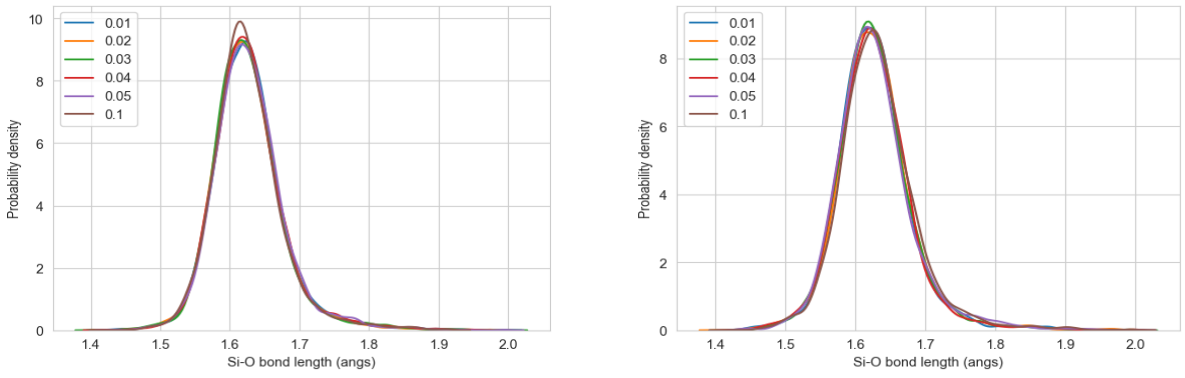
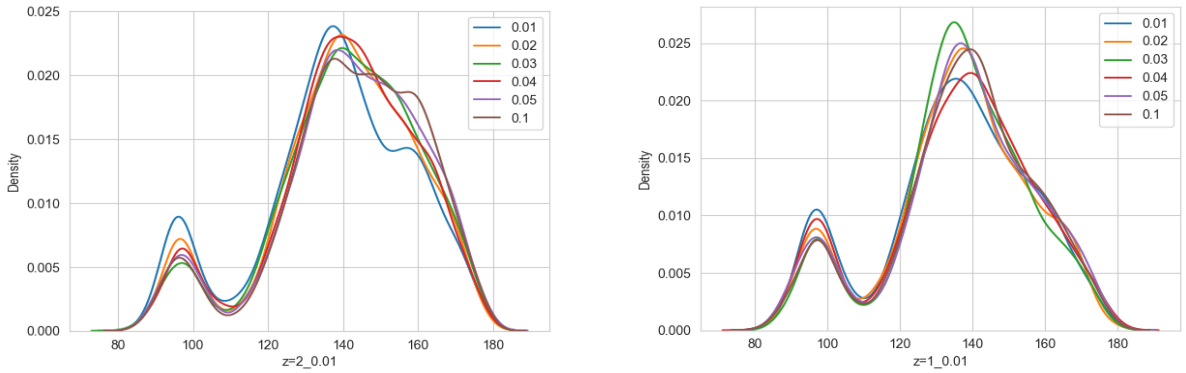
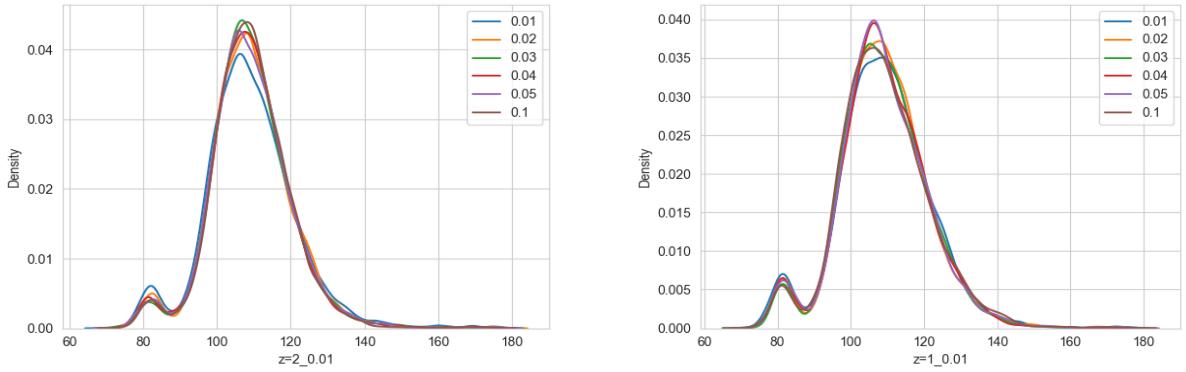


Figure B.1: Probability distribution of bond lengths of a) $N = 2$ and b) $N = 1$



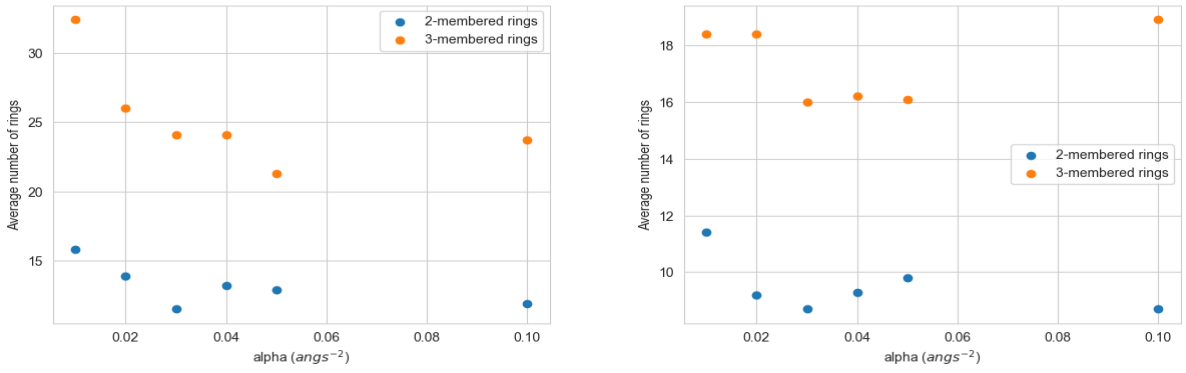
(a) (b)

Figure B.2: Probability distribution of Si-O-Si bond angles of a) N = 2 and b) N = 1



(a) (b)

Figure B.3: Probability distribution of O-Si-O bond angles of a) N = 2 and b) N = 1



(a) (b)

Figure B.4: Average number of rings for a) N = 2 and b) N = 1

Table B.1: Average number of ^{3}Si, ^{4}Si, and ^{5}Si participating in 2-MR, 3-MR, both types of rings and in none of the previously named rings. And, the probabilty of finding specific Si in a given type of ring

Atom type	average number	percentage
^{3}Si lone	1.81	0.84
^{4}Si lone	128	59
^{5}Si lone	1.46	0.68
^{3}Si in 2-MR	0.223	0.104
^{4}Si in 2-MR	15.9	7.4
^{5}Si in 2-MR	0.676	0.313
^{3}Si in 3-MR	0.993	0.46
^{4}Si in 3-MR	54.5	25
^{5}Si in 3-MR	1.71	0.79
^{3}Si in both	0.00668	0.0031
^{4}Si in both	8.76	4.1
^{5}Si both	1.63	0.76

conditional probability	percentage chance
P(3-MR|^{5}Si)	61
P(2-MR|^{3}Si)	7.5

C

Single Variable Linear Regressions

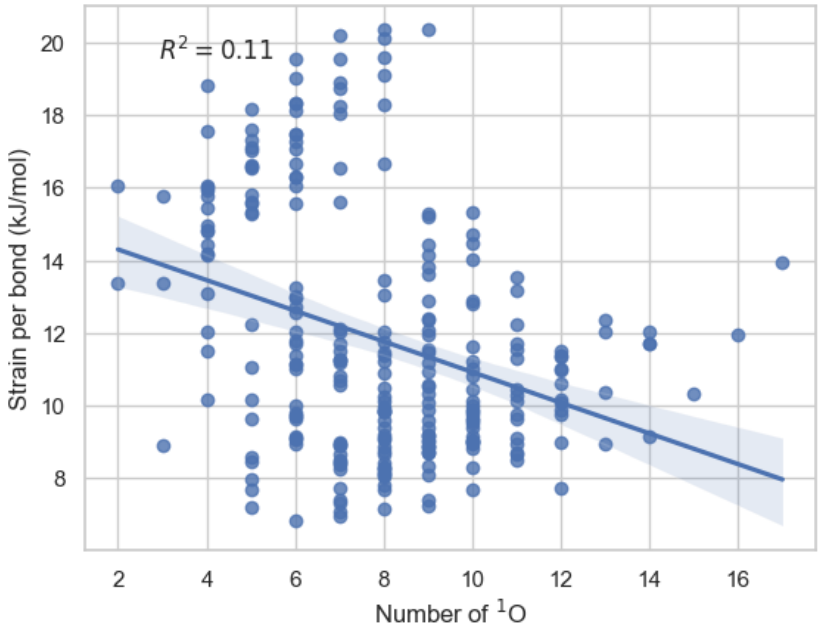


Figure C.1: Linear trend between the number of 1-coordinate O atoms and the strain per bond for the given model.

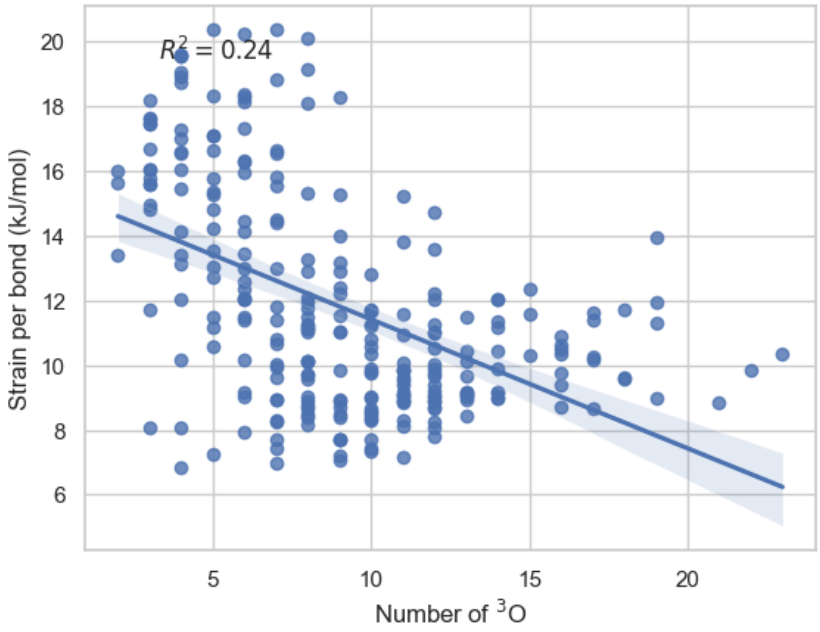


Figure C.2: Linear trend between the number of 3-coordinate O atoms and the strain per bond for the given model.

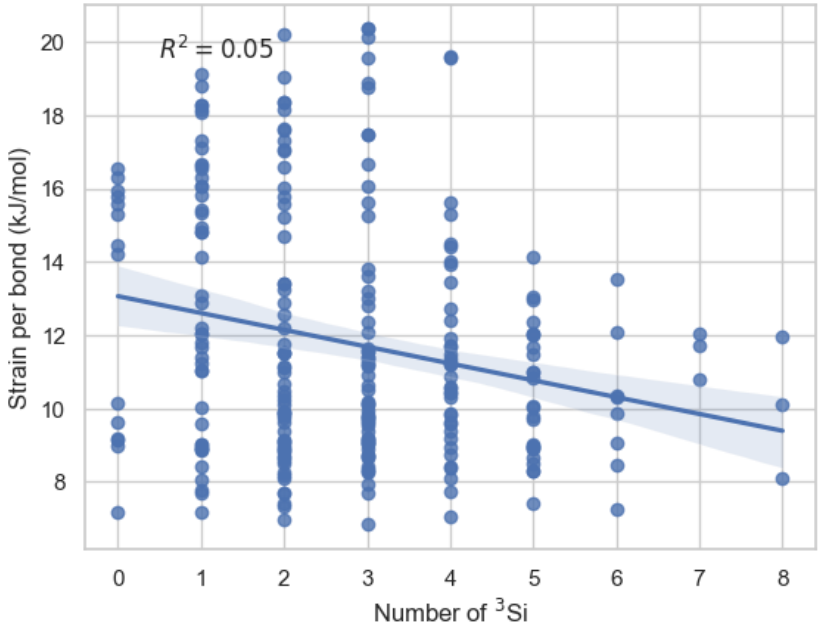


Figure C.3: Linear trend between the number of 3-coordinate Si atoms and the strain per bond for the given model.

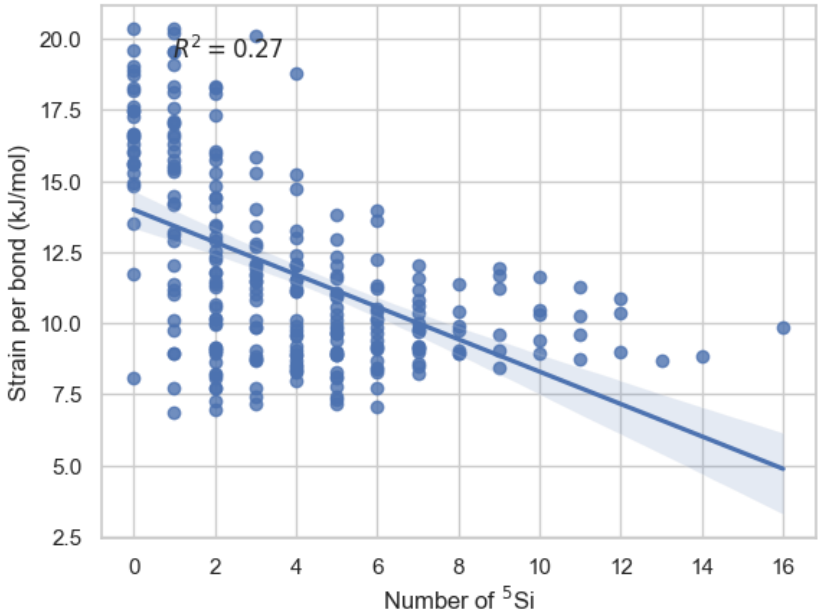


Figure C.4: Linear trend between the number of 5-coordinate Si atoms and the strain per bond for the given model.

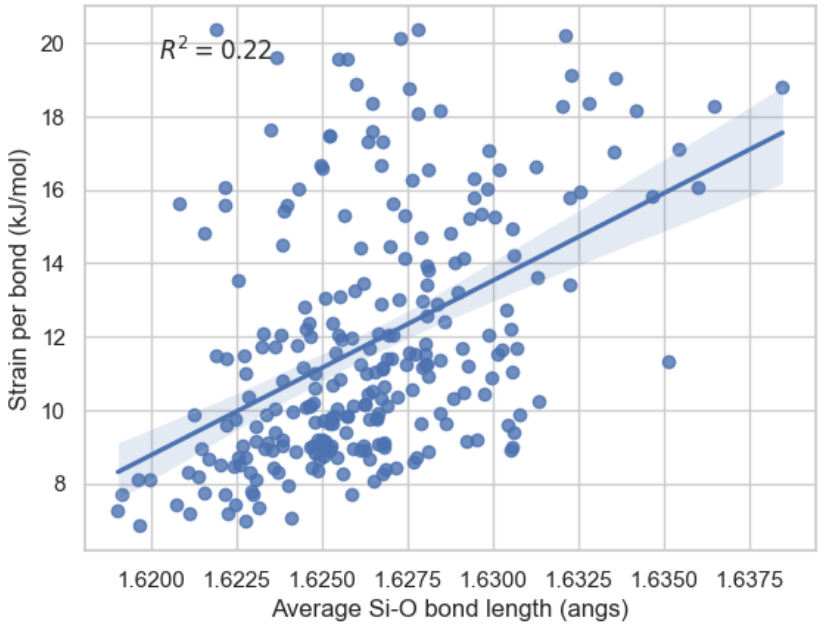


Figure C.5: Linear trend between the number of average bond length and the strain per bond for the given model.

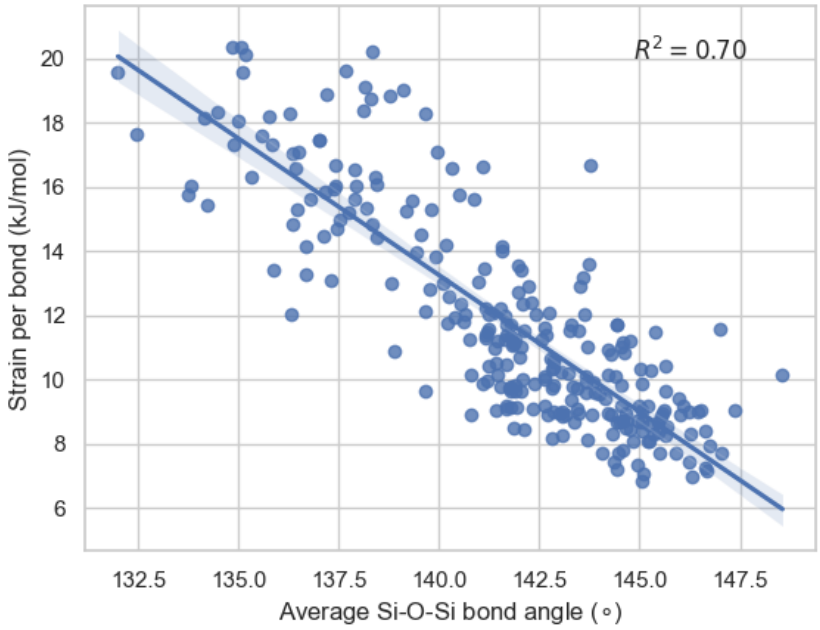


Figure C.6: Linear trend between the number of average bond angle and the strain per bond for the given model.

D

Full statistical models Reports

Dep. Variable:	ref_E_dry_dft_per_bond	**R-squared:**	0.925
Model:	OLS	**Adj. R-squared:**	0.924
Method:	Least Squares	**F-statistic:**	632.7
Date:	Tue, 11 Jun 2024	**Prob (F-statistic):**	1.83e-141
Time:	15:18:01	**Log-Likelihood:**	-345.89
No. Observations:	261	**AIC:**	703.8
Df Residuals:	255	**BIC:**	725.2
Df Model:	5		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-302.1403	39.010	-7.745	0.000	-378.964	-225.317
O_one_coord	0.4565	0.033	13.983	0.000	0.392	0.521
O_two_coord	-0.0181	0.001	-14.525	0.000	-0.021	-0.016
O_three_coord	-0.1465	0.027	-5.352	0.000	-0.200	-0.093
AVG_BL_ALL	225.5846	24.157	9.338	0.000	178.012	273.158
avg_angle	-0.3496	0.031	-11.291	0.000	-0.411	-0.289

Omnibus:	18.249	**Durbin-Watson:**	1.655
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	28.606
Skew:	0.444	**Prob(JB):**	6.14e-07
Kurtosis:	4.358	**Cond. No.**	2.91e+05

Figure D.1: Original report from stats-model for the strain per bond of all surfaces as function of ^{1}O (O_one_coord), ^{2}O (O_two_coord), ^{3}O (O_three_coord), average Si-O bond length (AVG_BL_ALL), and average Si-O-Si bond length (avg_angle).

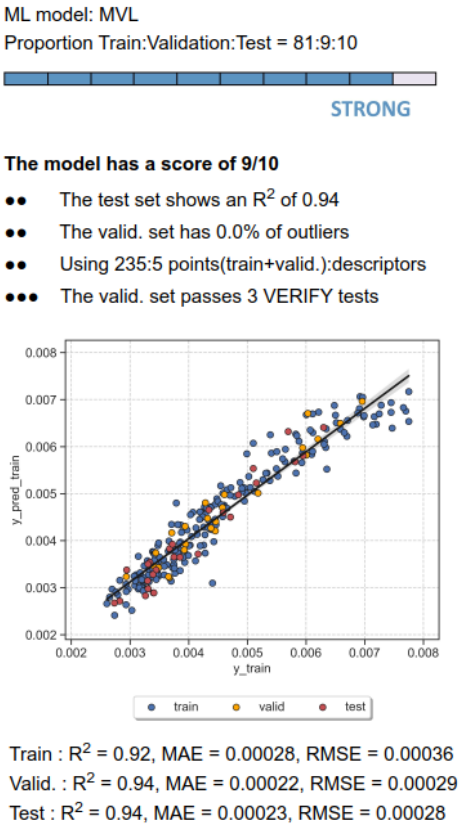


Figure D.2: Performance of ROBERT proposed model.

Dep. Variable:	ref_E_dry_dft_per_bond	**R-squared:**	0.015
Model:	OLS	**Adj. R-squared:**	-0.004
Method:	Least Squares	**F-statistic:**	0.7953
Date:	Thu, 13 Jun 2024	**Prob (F-statistic):**	0.554
Time:	12:23:47	**Log-Likelihood:**	1372.3
No. Observations:	261	**AIC:**	-2733.
Df Residuals:	255	**BIC:**	-2711.
Df Model:	5		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
Intercept	-0.0242	0.054	-0.448	0.654	-0.131	0.082
O_one_coord	-6.24e-05	4.52e-05	-1.381	0.168	-0.000	2.66e-05
O_two_coord	7.437e-07	1.73e-06	0.430	0.667	-2.66e-06	4.15e-06
O_three_coord	-2.167e-06	3.79e-05	-0.057	0.954	-7.67e-05	7.24e-05
AVG_BL_ALL	0.0159	0.033	0.474	0.636	-0.050	0.082
avg_angle	2.241e-05	4.28e-05	0.523	0.601	-6.2e-05	0.000

Omnibus:	23.686	**Durbin-Watson:**	0.306
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	26.840
Skew:	0.759	**Prob(JB):**	1.48e-06
Kurtosis:	2.593	**Cond. No.**	2.91e+05

Figure D.3: Original report from stats-model for the strain per bond of all surfaces as function of ^{1}O (O_one_coord), ^{2}O (O_two_coord), ^{3}O (O_three_coord), average Si-O bond length (AVG_BL_ALL), and average Si-O-Si bond length (avg_angle) after shuffling all values except for ^{2}O (O_three_coord)

Dep. Variable:	ref_E_dry_dft_per_bond	**R-squared:**	0.819
Model:	OLS	**Adj. R-squared:**	0.815
Method:	Least Squares	**F-statistic:**	230.6
Date:	Thu, 13 Jun 2024	**Prob (F-statistic):**	1.97e-92
Time:	16:06:13	**Log-Likelihood:**	1593.2
No. Observations:	261	**AIC:**	-3174.
Df Residuals:	255	**BIC:**	-3153.
Df Model:	5		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
Intercept	-0.0025	0.021	-0.118	0.906	-0.044	0.039
O_one_coord	-3.744e-06	1.93e-05	-0.194	0.846	-4.17e-05	3.42e-05
O_two_coord	-9.952e-06	3.06e-07	-32.483	0.000	-1.06e-05	-9.35e-06
O_three_coord	4.399e-06	1.2e-05	0.367	0.714	-1.92e-05	2.8e-05
AVG_BL_ALL	0.0046	0.013	0.362	0.718	-0.020	0.029
avg_angle	1.914e-05	1.26e-05	1.522	0.129	-5.62e-06	4.39e-05

Omnibus:	4.463	**Durbin-Watson:**	1.457
Prob(Omnibus):	0.107	**Jarque-Bera (JB):**	4.487
Skew:	0.319	**Prob(JB):**	0.106
Kurtosis:	2.922	**Cond. No.**	2.63e+05

Figure D.4: Original report from stats-model for the strain per bond of all surfaces as function of ^{1}O (O_one_coord), ^{2}O (O_two_coord), ^{3}O (O_three_coord), average Si-O bond length (AVG_BL_ALL), and average Si-O-Si bond length (avg_angle) after shuffling values of descriptors except .

Dep. Variable:	ref_E_dry_dft_per_bond	**R-squared:**	0.864
Model:	OLS	**Adj. R-squared:**	0.862
Method:	Least Squares	**F-statistic:**	405.5
Date:	Tue, 11 Jun 2024	**Prob (F-statistic):**	1.85e-109
Time:	15:18:43	**Log-Likelihood:**	-424.56
No. Observations:	261	**AIC:**	859.1
Df Residuals:	256	**BIC:**	876.9
Df Model:	4		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
Intercept	-533.9591	48.025	-11.118	0.000	-628.533	-439.386
O_one_coord	0.5220	0.044	11.967	0.000	0.436	0.608
O_three_coord	-0.4149	0.027	-15.238	0.000	-0.469	-0.361
AVG_BL_ALL	394.7746	28.552	13.826	0.000	338.547	451.002
avg_angle	-0.6823	0.028	-24.275	0.000	-0.738	-0.627

Omnibus:	24.400	**Durbin-Watson:**	1.587
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	35.062
Skew:	0.613	**Prob(JB):**	2.43e-08
Kurtosis:	4.312	**Cond. No.**	1.03e+05

Figure D.5: Original report from stats-model for the strain per bond of all surfaces as function of ^{1}O (O_one_coord) ^{3}O (O_three_coord), average Si-O bond length (AVG_BL_ALL), and average Si-O-Si bond length (avg_angle).

Dep. Variable:	ref_E_dry_dft_per_bond	**R-squared:**	0.741
Model:	OLS	**Adj. R-squared:**	0.722
Method:	Least Squares	**F-statistic:**	39.37
Date:	Thu, 13 Jun 2024	**Prob (F-statistic):**	1.55e-15
Time:	12:37:30	**Log-Likelihood:**	-80.218
No. Observations:	60	**AIC:**	170.4
Df Residuals:	55	**BIC:**	180.9
Df Model:	4		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-0.0426	0.009	-4.715	0.000	-0.061	-0.024
O_one_coord	-1.1363	0.391	-2.904	0.005	-1.920	-0.352
O_two_coord	-2.3028	0.432	-5.336	0.000	-3.168	-1.438
O_three_coord	-2.6963	0.494	-5.457	0.000	-3.686	-1.706
AVG_BL_ALL	234.8756	39.750	5.909	0.000	155.214	314.537
avg_angle	-0.2804	0.060	-4.655	0.000	-0.401	-0.160

Omnibus:	0.260	**Durbin-Watson:**	1.878
Prob(Omnibus):	0.878	**Jarque-Bera (JB):**	0.310
Skew:	0.147	**Prob(JB):**	0.856
Kurtosis:	2.806	**Cond. No.**	6.61e+17

Figure D.6: Original report from stats-model for the strain per bond of N = 1 surfaces as function of ^{1}O (O_one_coord), ^{2}O (O_two_coord), ^{3}O (O_three_coord), average Si-O bond length (AVG_BL_ALL), and average Si-O-Si bond length (avg_angle).

Dep. Variable:	ref_E_dry_dft_per_bond	**R-squared:**	0.807
Model:	OLS	**Adj. R-squared:**	0.792
Method:	Least Squares	**F-statistic:**	57.32
Date:	Thu, 13 Jun 2024	**Prob (F-statistic):**	5.58e-19
Time:	12:37:08	**Log-Likelihood:**	-54.286
No. Observations:	60	**AIC:**	118.6
Df Residuals:	55	**BIC:**	129.0
Df Model:	4		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-0.0161	0.003	-6.050	0.000	-0.021	-0.011
O_one_coord	-1.0272	0.231	-4.437	0.000	-1.491	-0.563
O_two_coord	-1.6769	0.254	-6.608	0.000	-2.186	-1.168
O_three_coord	-1.9341	0.288	-6.721	0.000	-2.511	-1.357
AVG_BL_ALL	336.2778	44.765	7.512	0.000	246.568	425.988
avg_angle	-0.3871	0.043	-8.948	0.000	-0.474	-0.300

Omnibus:	3.828	**Durbin-Watson:**	1.898
Prob(Omnibus):	0.147	**Jarque-Bera (JB):**	2.989
Skew:	-0.369	**Prob(JB):**	0.224
Kurtosis:	3.807	**Cond. No.**	8.05e+17

Figure D.7: Original report from stats-model for the strain per bond of N = 2 surfaces as function of ^{1}O (O_one_coord), ^{2}O (O_two_coord), ^{3}O (O_three_coord), average Si-O bond length (AVG_BL_ALL), and average Si-O-Si bond length (avg_angle).

Dep. Variable:	ref_E_dry_dft_per_bond	**R-squared:**	0.773
Model:	OLS	**Adj. R-squared:**	0.765
Method:	Least Squares	**F-statistic:**	92.10
Date:	Thu, 13 Jun 2024	**Prob (F-statistic):**	9.13e-42
Time:	12:36:16	**Log-Likelihood:**	-130.18
No. Observations:	141	**AIC:**	272.4
Df Residuals:	135	**BIC:**	290.1
Df Model:	5		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-321.0645	274.420	-1.170	0.244	-863.784	221.655
O_one_coord	0.1075	0.628	0.171	0.864	-1.135	1.350
O_two_coord	-0.2704	0.629	-0.430	0.668	-1.514	0.973
O_three_coord	-0.4039	0.629	-0.643	0.522	-1.647	0.839
AVG_BL_ALL	298.7354	27.708	10.782	0.000	243.938	353.533
avg_angle	-0.2796	0.033	-8.491	0.000	-0.345	-0.214

Omnibus:	30.217	**Durbin-Watson:**	1.936
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	85.628
Skew:	0.790	**Prob(JB):**	2.55e-19
Kurtosis:	6.475	**Cond. No.**	2.28e+06

Figure D.8: Original report from stats-model for the strain per bond of N = 3 surfaces as function of ^{1}O (O_one_coord), ^{2}O (O_two_coord), ^{3}O (O_three_coord), average Si-O bond length (AVG_BL_ALL), and average Si-O-Si bond length (avg_angle).

Dep. Variable:	ref_E_dry_dft_per_bond	**R-squared:**	0.773
Model:	OLS	**Adj. R-squared:**	0.766
Method:	Least Squares	**F-statistic:**	115.5
Date:	Thu, 13 Jun 2024	**Prob (F-statistic):**	9.76e-43
Time:	12:40:04	**Log-Likelihood:**	-130.40
No. Observations:	141	**AIC:**	270.8
Df Residuals:	136	**BIC:**	285.5
Df Model:	4		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-494.1649	52.388	-9.433	0.000	-597.766	-390.564
O_one_coord	0.5103	0.044	11.561	0.000	0.423	0.598
O_two_coord	0.1334	0.024	5.545	0.000	0.086	0.181
AVG_BL_ALL	298.1017	27.630	10.789	0.000	243.461	352.742
avg_angle	-0.2815	0.033	-8.603	0.000	-0.346	-0.217

Omnibus:	29.875	**Durbin-Watson:**	1.956
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	84.547
Skew:	0.780	**Prob(JB):**	4.37e-19
Kurtosis:	6.458	**Cond. No.**	4.93e+05

Figure D.9: Original report from stats-model for the strain per bond of N = 3 surfaces as function of ^{1}O (O_one_coord), ^{2}O (O_two_coord), average Si-O bond length (AVG_BL_ALL), and average Si-O-Si bond length (avg_angle).

Dep. Variable:	ref_E_dry_dft_per_bond	**R-squared:**	0.051
Model:	OLS	**Adj. R-squared:**	-0.018
Method:	Least Squares	**F-statistic:**	0.7372
Date:	Thu, 13 Jun 2024	**Prob (F-statistic):**	0.571
Time:	12:44:03	**Log-Likelihood:**	-119.19
No. Observations:	60	**AIC:**	248.4
Df Residuals:	55	**BIC:**	258.9
Df Model:	4		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-0.0086	0.017	-0.497	0.621	-0.043	0.026
O_one_coord	-0.1375	0.749	-0.183	0.855	-1.639	1.364
O_two_coord	-0.4049	0.826	-0.490	0.626	-2.061	1.251
O_three_coord	-0.6953	0.946	-0.735	0.466	-2.591	1.201
AVG_BL_ALL	55.1644	76.114	0.725	0.472	-97.371	207.700
avg_angle	-0.1072	0.115	-0.930	0.357	-0.338	0.124

Omnibus:	2.981	**Durbin-Watson:**	1.787
Prob(Omnibus):	0.225	**Jarque-Bera (JB):**	1.860
Skew:	0.192	**Prob(JB):**	0.395
Kurtosis:	2.228	**Cond. No.**	6.92e+17

Figure D.10: Original report from stats-model for the strain per bond of *N* = 1 surfaces as function of ^{1}O (O_one_coord), ^{2}O (O_two_coord), ^{3}O (O_three_coord), average Si-O bond length (AVG_BL_ALL), and average Si-O-Si bond length (avg_angle) after shuffling values of descriptors

Dep. Variable:	ref_E_dry_dft_per_bond	**R-squared:**	0.046
Model:	OLS	**Adj. R-squared:**	-0.023
Method:	Least Squares	**F-statistic:**	0.6616
Date:	Thu, 13 Jun 2024	**Prob (F-statistic):**	0.621
Time:	12:43:38	**Log-Likelihood:**	-102.16
No. Observations:	60	**AIC:**	214.3
Df Residuals:	55	**BIC:**	224.8
Df Model:	4		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-0.0028	0.006	-0.469	0.641	-0.015	0.009
O_one_coord	-0.1551	0.514	-0.302	0.764	-1.185	0.875
O_two_coord	-0.2340	0.564	-0.415	0.680	-1.363	0.895
O_three_coord	-0.4092	0.639	-0.640	0.525	-1.690	0.872
AVG_BL_ALL	56.3883	99.410	0.567	0.573	-142.834	255.610
avg_angle	-0.0804	0.096	-0.836	0.407	-0.273	0.112

Omnibus:	1.760	**Durbin-Watson:**	1.499
Prob(Omnibus):	0.415	**Jarque-Bera (JB):**	1.743
Skew:	0.379	**Prob(JB):**	0.418
Kurtosis:	2.650	**Cond. No.**	1.32e+18

Figure D.11: Original report from stats-model for the strain per bond of *N* = 2 surfaces as function of ^{1}O (O_one_coord), ^{2}O (O_two_coord), ^{3}O (O_three_coord), average Si-O bond length (AVG_BL_ALL), and average Si-O-Si bond length (avg_angle) after shuffling values of descriptors

Dep. Variable:	ref_E_dry_dft_per_bond	**R-squared:**	0.062
Model:	OLS	**Adj. R-squared:**	0.027
Method:	Least Squares	**F-statistic:**	1.773
Date:	Thu, 13 Jun 2024	**Prob (F-statistic):**	0.122
Time:	12:42:55	**Log-Likelihood:**	-230.33
No. Observations:	141	**AIC:**	472.7
Df Residuals:	135	**BIC:**	490.3
Df Model:	5		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
Intercept	-434.3331	558.312	-0.778	0.438	-1538.503	669.83
O_one_coord	0.9949	1.278	0.778	0.438	-1.533	3.523
O_two_coord	0.9035	1.279	0.706	0.481	-1.626	3.433
O_three_coord	0.8517	1.279	0.666	0.507	-1.677	3.381
AVG_BL_ALL	45.7601	56.372	0.812	0.418	-65.726	157.24
avg_angle	-0.1472	0.067	-2.197	0.030	-0.280	-0.015

Omnibus:	8.787	**Durbin-Watson:**	1.326
Prob(Omnibus):	0.012	**Jarque-Bera (JB):**	8.617
Skew:	0.549	**Prob(JB):**	0.0135
Kurtosis:	3.510	**Cond. No.**	2.28e+06

Figure D.12: Original report from stats-model for the strain per bond of N = 3 surfaces as function of ^{1}O (O_one_coord), ^{2}O (O_two_coord), ^{3}O (O_three_coord), average Si-O bond length (AVG_BL_ALL), and average Si-O-Si bond length (avg_angle) after shuffling values of descriptors