An automatized workflow from molecular dynamic simulation to quantum chemical methods to identify elementary reactions and compute reaction constants

Gunnar Schmitz*, Özlem Yönder†, Bastian Schnieder‡
Rochus Schmid§, Christof Hättig¶

September 27, 2021

Abstract

We present an automatized workflow which, starting from molecular dynamics simulations, identifies reaction events, filters them, and prepares them for accurate quantum chemical calculations using e.g. Density Functional Theory (DFT) or Coupled Cluster methods. The capabilities of the automatized workflow are demonstrated by the example of simulations for the combustion of some polycyclic aromatic hydrocarbons (PAHs). It is shown how key elementary reaction candidates are filtered out of a much larger set of redundant reactions and refined further. The molecular species in question are optimized using DFT and reaction energies, barrier heights, and reaction rates are calculated. The setup is general enough to include at this stage configurational sampling, which can be exploited in the future. Using the introduced machinery, we investigate how the observed reaction types depend on the gas atmosphere used in the molecular dynamics simulation. For the re-optimization on the DFT level, we show how the additional information needed to switch from reactive force-field to electronic structure calculations can be filled in and study how well ReaxFF and DFT agree with each other and shine light on the perspective of using more accurate semi-empirical methods in the MD simulation.

Keywords: Reaction finder, ReaxFF, DFT, automatized workflow, MD simulations. ■

*Corresponding author: gunnar.schmitz@rub.de, Computational Materials Chemistry group, Ruhr-Universität Bochum, D-44780 Bochum, Germany

†Corresponding author: oezlem.yoender@rub.de, Lehrstuhl für Theoretische Chemie, Ruhr-Universität Bochum, D-44780 Bochum, Germany

‡Computational Materials Chemistry group, Ruhr-Universität Bochum, D-44780 Bochum, Germany

§Computational Materials Chemistry group, Ruhr-Universität Bochum, D-44780 Bochum, Germany

¶Lehrstuhl für Theoretische Chemie, Ruhr-Universität Bochum, D-44780 Bochum, Germany



1 INTRODUCTION

Quantum chemical methods are nowadays important tools to model chemical reactions and supply researchers with additional data that is not easily accessible by experimental techniques. Due to the algorithmic progress over the last decades methods like density functional theory (DFT) or even some coupled cluster (CC) methods[1–4] can be routinely applied to large molecules with up to 100 atoms without creating a computational bottleneck. However, in large heterogeneous systems chemical reactions are often very complex in nature and require more than one or two reaction steps. In those cases selecting the elementary reaction steps that are quantum chemically studied only based on chemical or human intuition becomes problematic since the chemical reaction space is too large to be sampled by man power alone. Thus, automatic schemes and tools are necessary to propose possible reactions which can be closer investigated by DFT or CC methods.

There are different routes to such a scheme. One is defining reaction families. The MIT reaction mechanism generator (RMG) is such an example, which has been used to generate kinetic models for the combustion of several fuels such as *n*-butanol[5], diisopropyl ketone[6], and JP-10 jet fuel[7]. RMG uses molecular graphs to convert the reactants to products by the available reactions from its reaction library[8,9]. The thermodynamic properties are then calculated by means of Benson's additivity rules, semi-empirical or ab initio quantum chemistry methods[8,9]. Such kinetic model generators require a pre-knowledge of the type of reactions which would occur. On the other hand, for arbitrary systems, being restricted to reaction families may hinder the inclusion of new chemical reactions valid for that system and lead

to biased results[10]. Other routes, which do not require this pre-knowledge, employ the exploration of potential energy surfaces (PES). These can either be seen as an alternative or complementary to the reaction-family-based approaches. Chemoton is an example of such a reaction network discovery tool developed by Reiher and co-workers, which makes use of conformer generators to explore PESs while minimizing the exhaustivity of the possible reactions by introducing descriptors based on physical quantities such as the electron density[10]. Additional to this explorative scheme, there are other automated first-principle approaches using different aspects or targeting a different scope. Having products and educts fixed, Habershon[11] could for example demonstrate how to automatically identify different reaction pathways by introducing a Hamiltonian, which describes a reaction string connecting reactants and products, and performing molecular dynamics (MD) sampling of this Hamiltonian. The reactants and products in their approach are connected via graph theory. Furthermore, Zimmerman demonstrated in Ref. 12 to bypass the human effort for reaction space search by generating a range of intermediates that represent potential single elementary steps away from a starting structure. A screening of these structures is performed to identify those that are thermodynamically accessible, and then feasible reaction paths to the remaining structures are located.

Another approach to explore PESs is to start from reactive molecular dynamics simulations based on cheap force-field or semi-empirical methods to pre-sample possible relevant reactions. In this work, we propose such a workflow with the special interest in modelling the combustion of polycyclic aromatic hydrocarbons (PAHs) under oxyfuel conditions. Our goal for this initial publication is not to demonstrate it for an exhaustive exploration of the reactions involved during combustion, but to introduce the workflow and to explore its potential for identifying important reaction families within the limitations of the reactive force field, ReaxFF[13,14]. This route, as many others, requires reduction of the number of reaction events that are detected on the MD trajectory. We achieve this by comparisons of molecular graphs for the involved molecular species and thereby identifying unique reactions.

For the automated framework our ideas are similar to those presented by Döntgen et al. and Kröger et al. in Refs. 15–17 which inspired the current work. We avoid, however, to analyse trajectories in batches and keep track of the configurations of the molecular

species throughout our workflow and avoid the reduction of chemical structure information to SMILES strings. Like in the work of the latter authors, we use ReaxFF for the MD simulation due to its computational efficiency. However, for our specific interest ReaxFF has some weaknesses like a certain inertness of O_2. This makes it mandatory to perform the MD at rather high temperatures and under high pressure to trigger reaction events. This approach is acceptable since the MD simulation in this setup is considered as a sampling technique in the reaction space rather than a simulation approach for the combustion process. The actual reactions are then closer investigated by subsequent electronic structure calculations. This approach is not new. It is used by ChemTrayzer[15–17] and is, for example, also pursued by Wang et al.[18] in their chemical nano reactor, where they additionally apply periodically a force acting on all atoms directed to the center of the simulation box. An other way to drive reactions in MD simulations to certain products is the combination with coordinate driving techniques as for example applied in Ref. 19. Although not applied to MD simulations there are other ways to increase the rate of reaction events are to add artificial attractive interatomic forces[20–22]. Concerning MD simulations and reaction sampling there emerge new strategies, which employ machine learning along side as outlined in Ref. 23. All these approaches allow identifying possible reaction events, but this opens two new challenges which we want to address in this work. First, the number of reaction events can become quite large, and it is important to filter only those which are of interest for the main research objective. Second, the molecular structures extracted from the MD trajectories are due to the high temperature usually quite deformed and far from ground state equilibria, which makes subsequent optimization at, for example, the DFT level, difficult, especially for the transition states. Here we present an automated scheme to handle those cases and comment on challenging aspects. The species can be (re-)optimized on various levels of theory (LoT) to obtain reaction pathways and accurate reaction energies and activation barriers.

The paper is organized as follows: First, we outline the basic theory of our automatized workflow. This is followed by a detailed overview of the computational settings used in this study. Afterwards, the results are presented and discussed. For this purpose we first validate our workflow against a known set of elementary reactions and then closer invesitigate the reaction event screening. On a qualitative level it will be studied how the atmosphere applied

in the simulation affects the observed reaction types and, more importantly, if common reaction types important for combustion can be identified. In the last part, we present some exemplary DFT results and discuss some challenges that arise from using the reactive force field ReaxFF as precursor method for the MD simulations.

2 COMPUTATIONAL METHODS

2.1 The workflow

Our integrated workflow as illustrated in Fig. 1 starts with MD simulations with the trajectories stored in HDF5 files[24]. The first central part of the workflow, which connects the MD simulations with the subsequent quantum chemical calculations, is the program *findR* (vide infra) which analyses the trajectories frame by frame and stores the identified reaction events in a relational SQL database. The reaction events found are then filtered to arrive at a non-redundant set of reactions. As similarity measure we use different approaches based on graph theory, but the setup is general enough to use also other kinds of similarity measures. During this process it can also be ensured that not only the identified reactions are non-redundant, but that also the molecular species are stored in a non-redundant fashion. As a next step, a reaction network can be created which connects the different reaction events for visual analysis and to ensure that no reactions were missed. In the Supporting Information we give two examples for this visual analysis as HTML file generated by our program. Furthermore, and key to our purpose, the found reactions can be investigated closer with (higher-level) quantum chemical methods. The educts, products, and transition states are optimized, which allows calculating barriers and reaction rates.

2.2 Database layout

Before going into detail how the reaction database is populated and which algorithms are in play to avoid redundancies and to automate the quantum chemical calculations, we want to give a simplified overview of the database layout. A simplified schematic overview of the tables in the database is depicted in Fig. 2. Arrows indicate which tables relate to each

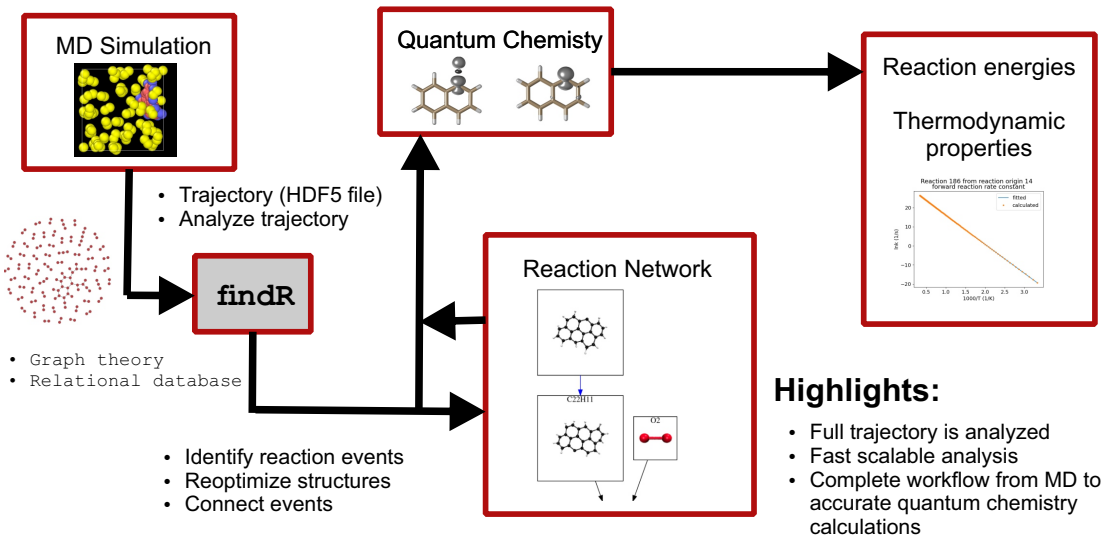


Figure 1: Illustration of the computational workflow employing *findR* for analyzing MD trajectories and the subsequent steps involving quantum chemistry calculations for determining reaction rates and thermodynamic parameters.

other and the arrow direction indicates the direction of the reference. The central table is `reac2spec` which links an abstract `species` table to the `reactions` table. The `species` table is general for a molecular species e.g. water and several optimized species (`opt_species`) can point to it. It does not contain structural information, but is just a global label for a species and the idea is to keep this table non-redundant. Geometries can be retrieved for various LoT from the `opt_species` table, where they are stored as `mfpx` file, which is a self-developed file format containing the coordinates and the connectivity. However, the `species` table contains an abstract object which is the basis for comparisons. Currently, we only use molecular graphs (vide infra), but other similarity measures can be implemented and stored in the database. The `reactions` table stores all reactions and allows for self-referencing in the case a studied reaction can spawn new reactions e.g. due to configurational sampling or different possible spin cases in the involved educts and products. We refer to the `species`, `reac2spec`, `reactions`, `LoT` and `opt_species` table as reaction space and the `md_species` and `revent` table as MD space. The latter is extracted from the MD simulation and the tables belonging to the reaction space are obtained from post-processing steps. The first post-processing step is the filtering step described in section 2.5 which reduces the reaction

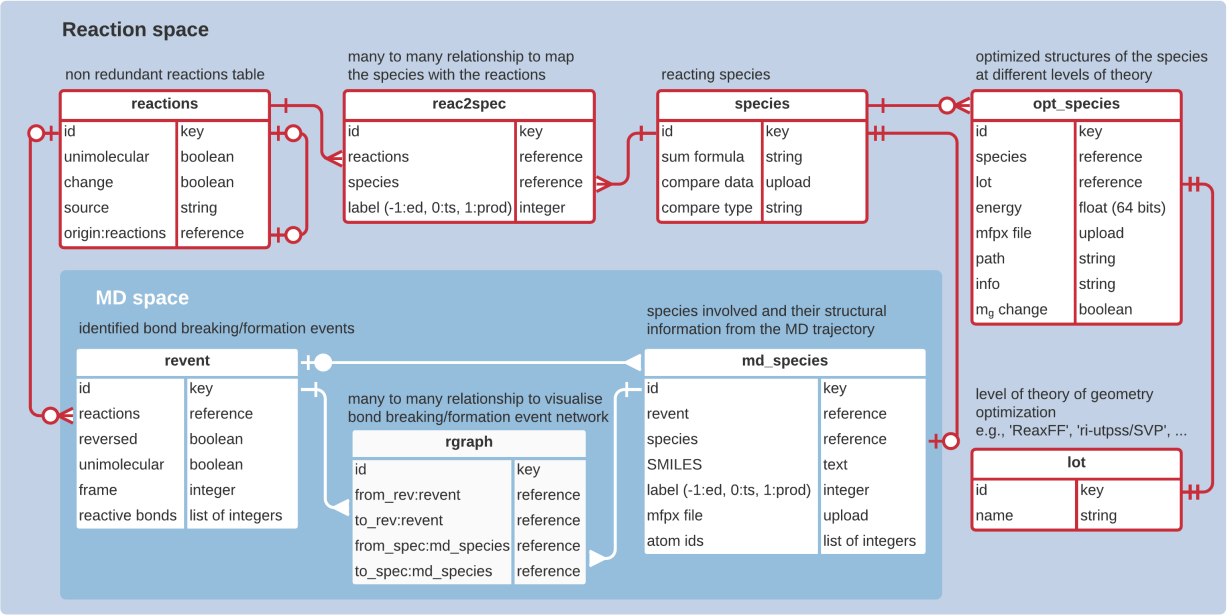


Figure 2: Simplified layout of the used database to store reactions and molecular species.

events stored in `revent` to a non-redundant set.

2.3 Molecular graph representation

The basis of all comparisons of molecular species in the trajectories is graph theory. A so-called molecular graph m_g is constructed for each molecular species, where the vertices are given by the atom types and the bonds are represented using edges. Figure 3 demonstrates how a molecular structure is converted into a graph. The connectivity for the graph construction can be obtained by several means. In the current work, it is based on the ReaxFF bond order, but other schemes could be applied as well. Just looking at the visualization of the graph, it looks like a 2-dimensional projection of the molecular structure, but the graph representation has some advantages. There are efficient available algorithms[25,26] to identify if two graphs are isomorphic. Graph isomorphism ensures the same connectivity, but does not rely on the atom order or the relative orientation of functional groups. This makes it a very robust comparison measure. A drawback is that for many reactions the configuration of the involved molecular species matters. However, a configurational sampling can always be added additionally after the initial filtering. From this step onward, new reactions can be

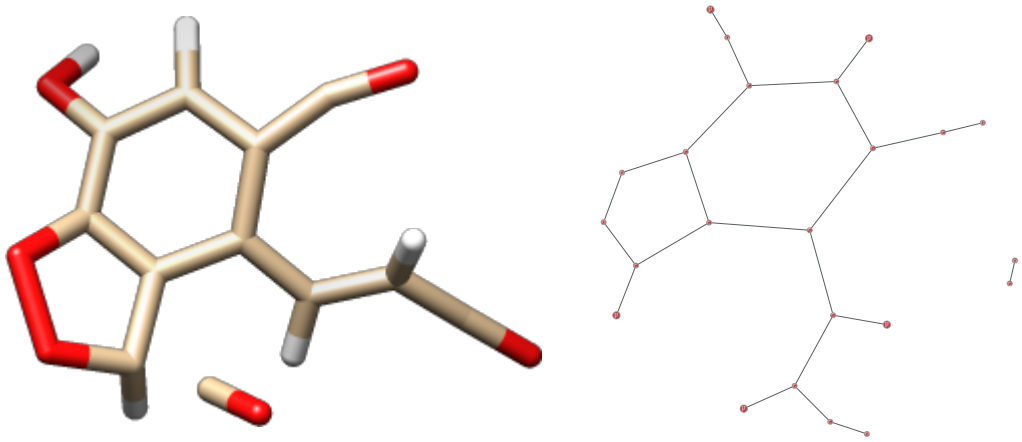


Figure 3: Illustration of how a molecular structure is mapped onto a graph, where the vertices are the atom types and the bonds are incorporated as edges. The depicted case shows a reactive complex involving two molecular species.

added anytime in the reaction database and reference to the reaction where they originate from.

2.4 Identifying reaction events

To identify reaction events, the frames in the MD trajectory are compared with each other on the basis of the molecular species. Since we are interested in the combustion of polycyclic aromatic hydrocarbons (PAHs) and not in the (known) gas phase chemistry of small molecules, not all species are tracked, but only those that fulfil certain criteria, for our examples only those which have at least 6 atoms and are composed of at least 2 carbon atoms. Starting from the first frame it is checked from frame to frame if the graph for the tracked species changed. For efficiency reasons the trajectory is not searched frame by frame, but the frame index is incremented by a user-defined step size (default 200). If a change in a molecular system is detected, the search direction is reversed using a smaller step size (default 10) which will be reduced down to 1 to locate the frames, where the reaction event took place. There are two levels of comparison: At level 1 it will be checked if the species are composed of the same atoms. If the composition does not change, it will be checked on level 2 if the bonding changed. As mentioned above, bonds are determined based on the ReaxFF

bond order, where per default a bond is classified by having a bond order larger than 0.5. Thus, the changes on level 2 classify unimolecular reactions while changes on level 1 indicate a reaction with two or more reaction partners. If a reaction event was found, three frames representing the educt(s), product(s), and the transition state, are stored for later analysis. During the frame comparison always two frames are compared with each other. If a reaction event was located, the frame in which the bond order of the reactive bond is closer to 0.5 is chosen as the transition state frame and this frame is stored as frame for the reaction event. The frames for educt(s) and product(s) are chosen relatively (-1,+1) to the transition state frame. The stored frames, however, do not contain minimum energy structures, which makes it later necessary to optimize the species for which different LoTs can be used. Following the outlined recipes all reaction events in the MD are registered, but the number of reaction events can easily become quite large which makes it mandatory to use filtering mechanisms, like those described in the next section.

2.5 Filtering non-redundant molecular species and reaction events

One strength of the *findR* scheme is that the whole trajectory is analysed and that one is not in danger of overlooking reaction steps. However, this strength also creates some challenges. Many reaction events occur during an MD simulation and some, if not most, of them are redundant or oscillatory motions where bond orders dropped only for a few frames below the threshold. Therefore, it is pivotal to develop a scheme to filter only the unique reactions. Along the same lines it is pivotal to store only a non-redundant set of molecular species. We will first describe how to filter the molecular species. For this purpose the full molecular graph is used, but additionally its stereo centers are considered in order to differentiate stereoisomers. For the detection of the latter, we use the algorithm implemented in OpenBabel[27] to identify chiral centers. The structures in the database can later be optimized at various levels. As outlined in Fig. 2 our database is flexible enough to allow for different comparison measures besides the molecular graph. However, if the used comparison type differs, two molecular species are automatically regarded as dissimilar.

To differentiate a reaction event r from its unique reaction class, we use in the following a bar ($\bar{r}$). The list of reactions $\mathbf{r}$ contains all observed and possible redundant reactions while

$\bar{\mathbf{r}}$ is a set of unique reaction events. Algorithm 1 outlines the filtering that is applied. After identifying all elementary reactions $\mathbf{r}$ we loop over them and compare the molecular species $\mathbf{m}^{(r)}$ i.e. educts, products and the transition state participating in the found reaction r with the analogue molecular species $\mathbf{m}^{(\bar{r})}$ in the set of already known unique reaction events $\bar{\mathbf{r}}$. The comparison is made on the basis of the molecular graphs constructed as outlined in sections 2.3; however, a few fast checks based on for example the sum formula are implemented to bypass graph comparisons if the atomic compositions differ.

During the comparison it is tracked if the molecular species belong to the educt or product side to classify the reaction r as a forward or backward reaction. If there is no match in $\bar{\mathbf{r}}$ the reaction r is added as a new unique reaction. The first reaction of its kind is always classified as a forward reaction.

For the comparison we employ two schemes. In the first one, the comparison is made using the same molecular graph as used in the initial analysis to identify reaction events, but in the second scheme we only consider the molecular graph based on certain atom types. Thus, a reaction event is only found if the connectivity of such an atom-type changed. For a C-graph this means that we only consider C-C bonding. Reactions like H abstraction or the addition of O_2 are not regarded as reactions. For our intended application in char combustion a sane choice is to restrict the atom types to carbon (if the model system does not contain potential hetero-atoms, e.g., nitrogen or sulfur). This essentially gives two layers of filtering. We start from an initial list of all reactions r, but then filter only the unique reactions $\bar{r}$ based on the molecular graph. However, also this shortened list still contain reactions in which the molecular C-graph does not change. If we also consider the change in the C-graph, we arrive at the final stage of filtering $\bar{r}^{(C)}$.

The number of reaction events $n_{\bar{r}}^{(C)}$ is a measure for the decomposition of the carbon structure. In Section 3.3 we investigate how this number depends on conditions like the atmosphere. During the analysis, we need to keep track of the reactive species for a reaction. They will be stored in lists $\mathcal{L}_{m,t}$, where m refers to the molecular graph and t to the type (educt, product, or transition state).

Algorithm 1 Filtering unique reactions from redundant list

Read in all already found unique reactions $\bar{\mathbf{r}}$

Loop over all reactions in $\mathbf{r}$

for $r = r_1, ..., r_n$ **do**

Get set of molecular species $\mathbf{m}^{(r)}$ in r

for $m^{(r)} \in \mathbf{m}^{(r)}$ **do**

Identify type t as educt, product or transition state

Create molecular graph $m_g^{(r)}$

Store $m_g^{(r)}$ and t in list $\mathcal{L}_{m,t}$

end for

for $\bar{r} = \bar{r}_1, ..., \bar{r}_m$ **do**

for $m^{(\bar{r})} \in \mathbf{m}^{(\bar{r})}$ **do**

Identify type τ as educt, product or transition state

Create molecular graph $m_g^{(\bar{r})}$

Store $m_g^{(\bar{r})}$ and τ in list $\mathcal{L}_{\bar{m},\tau}$

end for

if $\mathcal{L}_{m,t} = \mathcal{L}_{\bar{m},\tau}$ **then**

$\rightarrow r \in \bar{\mathbf{r}}$

Exit Loop

end if

end for

if $r \notin \bar{\mathbf{r}}$ **then**

Add r to $\bar{\mathbf{r}}$

end if

end for

2.6 Refining reaction and activation energies using density functional theory

A second central part of the automated workflow is a new self-developed module which automates the quantum chemistry calculations with the TURBOMOLE program package[28]. More specifically, this module is used to get initial equilibrium and transition state structures from the database and refine them to obtain accurate thermodynamic and kinetic data. In the current work, the ReaxFF optimized structures are used as initial structures and only the non-redundant (unique) reactions are considered. For the refinement of the molecular structures, density functional theory (DFT) is used.

In a first step, the ReaxFF optimized initial structures of the educts, products, and transition states are refined using a computationally efficient low-cost DFT functional (avoiding exact exchange) and a small basis set, TPSS-D3/SVP in this study. The data for reactions refined at the DFT level contains inherently different information than available from the ReaxFF pre-optimization. Therefore, each reaction at the DFT level is registered as a new reaction to the database together with the information about the reaction origin, i.e., the reaction ID of the reference MD ReaxFF bond breaking/formation event as registered to the reactions table. Together with each new reaction, the successfully located transition state structures and the equilibrium structures connected to them or for the practically barrierless reactions just the equilibrium structures of the educts and products are registered in the optimized species table (`opt_species`) of the database. Afterwards, the data about these species can be retrieved from the database, re-optimized using a more accurate functional and a larger basis set, e.g., TPSSh-D3/TZVP, and be registered in the optimized species table with the new level of theory. The general workflow of the refinement of reaction and activation energies using an inexpensive DFT functional and small basis set is shown in algorithm 3.

Since the MD with ReaxFF provides only molecular geometries and energies, the quantum chemical electronic structure calculations need additional information to specify the electronic wave function. These are the charges (or number of electrons) and the z component of the electronic spin, which sets a lower bound for the electron spin and the number

of unpaired electrons. Therefore, to determine the occupation of the molecular orbitals, we have used Fermi smearing as implemented in TURBOMOLE[29] with unrestricted DFT orbitals as indicated in algorithm 2. In the current study, it is assumed that all molecular species are charge neutral. For the equilibrium structures, an initial occupation for Fermi smearing is assigned based on the number of electrons in the system: For those with even number of electrons as $(n_\alpha\text{-}n_\beta) = 2$, $m_s^{start} = 1$, and for those with odd number of electrons as $(n_\alpha\text{-}n_\beta) = 1$, $m_s^{start} = 1/2$, where n_α and n_β are the number of electrons with, respectively, α and β spin and $m_s = (n_\alpha - n_\beta)/2$ is the secondary spin quantum number. The initial orbitals are generated using an Extended Hückel Theory (EHT) guess. After the Fermi smearing the fractional occupation numbers are rounded to the closest integer value, resulting in a new value m_s^{Fermi} for the secondary spin quantum number. To assure that the electronic state with the lowest energy is found, the DFT energy is then also evaluated with the secondary spin quantum number increased and, if possible, also decreased by 1 from the value found by the Fermi smearing ($m_s^{Fermi} \pm 1$). The spin orbital occupation that resulted in the lowest energy (m_s^{Emin}) is then kept for the subsequent calculations.

Algorithm 2 Determine orbital occupation of equilibrium structures

if odd number of electrons **then**

$m_s^{start} = 1/2$

else if even number of electrons **then**

$m_s^{start} = 1$

end if

Create start orbitals from extended Hückel Theory $\rightarrow m_s^{start}$

Minimize the Kohn-Sham energy and determine the orbital occupations with Fermi smearing $\rightarrow m_s^{Fermi}$

Turn off Fermi smearing

Minimize the Kohn-Sham energy of the species by modifying occupations $m_s^{Fermi} \pm 1$

Choose the one with lowest energy $\{m_s^{Fermi} - 1, m_s^{Fermi}, m_s^{Fermi} + 1\} \rightarrow m_s^{minE}$

Optimize the geometry of the species with $m_s^{minE} \rightarrow x^{QM,ref}$

Skip reaction

if $\mathcal{m}_g^{(\mathrm{ReaxFF},ref)} = \mathcal{m}_g^{(QM,ref)}$ **then**

return $m_s^{final} = m_s^{minE}$

else

Optimize the geometry of the species with $m_s^{minE} + 1 \rightarrow x^{QM,ref}$

if $\mathcal{m}_g^{(\mathrm{ReaxFF},ref)} = \mathcal{m}_g^{(QM,ref)}$ **then**

return $m_s^{final} = m_s^{minE} + 1$

else

Skip reaction

end if

end if

Then a geometry optimization is performed. To avoid that geometry optimizations for equilibrium structures get trapped on saddle points because a start structure had a higher symmetry than the searched-for minimum on the PES, a random noise with a range of $\pm 0.05\,\text{Å}$ is added to the coordinates of the ReaxFF optimized structures. For the transition state structures, the noise is not applied to the atoms which belong to the reacting bonds (see below). If the molecular graph changes, additional geometry optimizations are done

with the original start geometry with $m_s^{Emin} + 1$. If the graph still changes, the reaction is skipped because this means that the particular reaction has no counterpart at the DFT level. Additionally, reactions where the initial educt or product structures dissociate into two or more molecules during the geometry optimization are skipped, because then it is not trivial to automatically track the actual reaction. In the future we will address this issue, but this current limitation is not a severe one, because actual termolecular reactions are rather rare for an elementary reaction step. In other words, in the outlined workflow we aim to identify elementary reaction steps, which are classified by the reactants reacting in a single reaction step and with a single transition state.

To ensure that the occupations of the spin orbitals of the educts, the transition state, and the products of a reaction are compatible also in non-unimolecular cases, we define for the educts and for the products sets of possible m_s values, m_s^{educts} and $m_s^{products}$, respectively, that can be reached by coupling of the individual spins, as shown in algorithm 3. If the two sets contain common m_s values which would result in a spin-conserving reaction, then only these values are considered. In the workflow for the location of the transition state on the DFT PES we have then the possibility to either refine the ReaxFF pre-optimized transition state, i.e. the structure which is taken from the MD trajectory where the bond order was 0.5 and then geometry optimized by constraining the bonds involved in the reaction using ReaxFF ($x^{\mathrm{ReaxFF},BO=0.5}$), with a geometry optimization for saddle points or, if this fails, to use a double ended reaction pathway optimization (*vide infra*). If the sets of possible values for the secondary spin quantum numbers of the educts and products do not contain a common element ($\mathbf{m}_s^{educts} \cap \mathbf{m}_s^{products} = \{\}$), only the transition state optimization workflow based on the ReaxFF start structure is carried out, trying in this case all possible values ($\mathbf{m}_s^{educts} \cup \mathbf{m}_s^{products}$). The educts and products of the reaction on the DFT PES are then determined starting from the DFT transition state as described below. For elementary reactions which involve more than two molecules, the generation of a reactive complex (*vide infra*) is a challenging problem. Such elementary reactions are, however, very rare. ~~This happens at most for 1-3 reactions in the analysed trajectory.~~ Therefore, these cases are not automated and the program skips these reactions in this initial proof of concept study.

Algorithm 3 Automation of density functional theory calculations

Read in all unique reactions $\bar{\mathbf{r}}$ from the database

for $r = \bar{r}_1, ..., \bar{r}_n$ **do**

Read the ReaxFF geometry optimized reference structures $\mathbf{x}^{\text{ReaxFF}}$

Add random noise $\delta x_{max} = 0.1\,\text{Å}$ to $x^{\text{ReaxFF}} \in \mathbf{x}^{\text{ReaxFF}}$

for $x^{\text{ReaxFF}} \in (\mathbf{x}^{\text{ReaxFF},educts} \cap \mathbf{x}^{\text{ReaxFF},products})$ **do**

Determine the orbital occupations and optimize the geometries:

$x^{\text{ReaxFF}} \rightarrow x^{QM,ref}, m_s^{QM,ref}$ (Algo. 2)

end for

if $n_{educts} \leq 2$ **and** $n_{products} \leq 2$ **then**

Couple the spins $\begin{cases} \mathbf{m}_s^{educts} = \left\{(m_s^{ed_1} + m_s^{ed_2}), (m_s^{ed_1} + m_s^{ed_2}) - 1, ..., |m_s^{ed_1} - m_s^{ed_2}|\right\} \\ \mathbf{m}_s^{products} = \left\{(m_s^{prod_1} + m_s^{prod_2}), (m_s^{prod_1} + m_s^{prod_2}) - 1, ..., |m_s^{prod_1} - m_s^{prod_2}|\right\} \end{cases}$

if $\mathbf{m}_s^{educts} \cap \mathbf{m}_s^{products} = \{\}$ **then**

for $m_s \in (\mathbf{m}_s^{educts} \cup \mathbf{m}_s^{products})$ **do**

Pre-optimize the TS structure $x^{\text{ReaxFF},BO=0.5} \rightarrow x^{\dagger,QM,ref,m_s}$

Optimize TS structure and IRC end points $x^{\dagger,QM,ref,m_s} \rightarrow x^{\dagger,m_s}, x^{IRC,m_s}$ (Algo. 4)

end for

else

for $m_s \in (\mathbf{m}_{s,educts} \cap \mathbf{m}_{s,products})$ **do**

Pre-optimize the TS structure $x^{\text{ReaxFF},BO=0.5} \rightarrow x^{\dagger,QM,ref,m_s}$

Optimize TS structure and IRC end points $x^{\dagger,QM,ref,m_s} \rightarrow x^{\dagger,m_s}, x^{IRC,m_s}$ (Algo. 4)

if $m_g^{(IRC)} \neq m_g^{(QM,ref)}$ **then**

Calculate approximate reaction path using woelfling

Select the peaks of the energy profile as TS start structures $x^{\dagger,QM,ref,m_s}$

Optimize TS structure and IRC end points $x^{\dagger,QM,ref,m_s} \rightarrow x^{\dagger,m_s}, x^{IRC,m_s}$ (Algo. 4)

end if

end for

end if

end if

end for

The module uses two alternative algorithms to locate a transition state. The first one is partially adopted from ChemTraYzer[15–17] and creates a start structure for the transition state optimization by constrained minimization of the ReaxFF pre-optimized 'transition state' structure ($x^{\mathrm{ReaxFF},BO=0.5} \rightarrow x^{\dagger,QM,ref}$). For this, the internal coordinates of the atoms which belong to the "reacting" bonds as identified by *findR*, i.e., the bonds for which the bond order changes (see Section 2.4), are fixed during this initial energy minimization. This resulting structure will be referred to as DFT pre-optimized TS structure. The Hessian of this structure is calculated using the *aoforce* module of TURBOMOLE[30,31]. Then, the constraints on the structure are removed for the optimization of the TS structure, i.e. the saddle point, with the trust-region image minimization (TRIM) algorithm[32] as implemented in the *statpt* program of TURBOMOLE. Then, the Hessian is calculated for the optimized saddle point. If this confirms that a saddle point is found successfully, i.e., that the geometry optimization converged to a stationary point with only one imaginary frequency ($N_{\nu_{im}}(x^{\dagger}) = 1$), an intrinsic reaction coordinate (IRC)[33] calculation is used to test if the saddle point connects the expected educts and products ($x^{QM,ref}$). It is checked if there are multiple molecules at the optimized end points, which should be the case for non-unimolecular reactions, and if this is the case, the molecules are separated using the topology of the molecular graph[34]. ~~The separated geometries are optimized following the same procedure as described in algorithm 2 to determine the occupation numbers for the α and β spin orbitals of the individual fragments.~~ The molecular orbitals of the separated fragments are generated by localization of spin orbitals of the optimized IRC end points. Then, the geometries of the separate fragments are optimized. Afterwards, the molecular graphs of the optimized IRC end points (m_g^{IRC}) are compared to those of the reference educts and products ($m_g^{QM,ref}$). It should be noted that the educts and products are not necessarily the same as those detected during the *findR* step. If the geometry-optimized IRC end points (x^{IRC}) differ from the expected educts and products (m_g^{IRC}), the equilibrium structures found from the IRC end points (x^{IRC}) are assigned instead of the former as its educts and products.

If the stationary point found by the saddle point optimization has no imaginary frequency or if it is connected to other educts or products than expected, an algorithm to locate the right transition state starting with a double-ended reaction path optimization between the

DFT-optimized educt and product structures ($x^{QM,ref}$) is employed in those cases where this is possible (see above). For this purpose we use the *woelfling* module[35] of TURBOMOLE. Afterwards, the maxima on the PES along the reaction pathways are chosen as new start structures for transition state optimization.

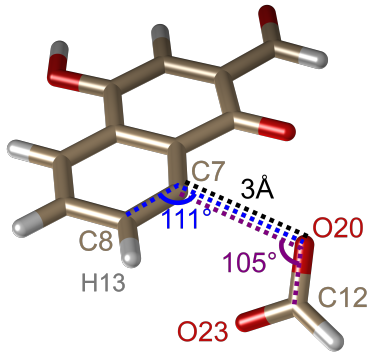


Figure 4: Sketch of the Z-matrix coordinates that determine the position and orientation of the formate radical relative to the PAH structure. The three torsion angles are not shown.

For bimolecular reactions this requires the combination of the structures of two reactants, stored separately in the database, to a reactive complex to get the start or end structures for the reaction path optimization. This is achieved by combining the Z-matrices of the two reacting species by making use of the Python module chemcoord[36]. In order to position the second molecule with respect to the first one, the missing internal coordinates related to the first three atoms of the second molecule have to be defined. Figure 4 shows an example of a reaction where a formate radical is released by breaking a C-O bond. Accordingly, atom O20 (cmp. Fig. 4) is set to be the reactive atom in the formate species and located at a distance of 3 Å from atom C7 in the remaining PAH fragment. For all internal coordinates of the separate fragments the values are taken from the respective optimized geometries. (The Z-matrices of the two products, the ReaxFF pre-optimized transition state structure, and the product complex are given in the Supporting Information.) The relative orientation of the fragments is determined as follows: For atom O20, the bond angle ($\angle$(O20C7C8)) and the dihedral angle (τ(O20C7C8H13)) are set to the values in the ReaxFF pre-optimized transition state structure. Similarly, for C12 the angle ($\angle$(C12O20C8)) and the dihedral (τ(C12O20C8H13)), and for O23 the dihedral (τ(O23O20C12H13)) are set to the values

in the ReaxFF pre-optimized transition state structure. After the structure of the reactive complex has been generated, the orbitals and occupations of the DFT pre-optimized TS structure are used to start the reaction path optimization.

For unimolecular reactions, before carrying out the reaction path optimization, the educt and product structures obtained at the DFT level ($x^{QM,ref}$) are compared via a similarity recognition program described in Ref. 37 which uses atomic matching and rotational superposition. If the educt and product are similar, which can occur due to oscillatory motion at high temperatures in the MD trajectory (see above), the reaction path optimization is skipped. This almost never happens for bimolecular reactions as the reacting species are optimized separately.

Algorithm 4 Determine transition state (TS) and connect it with educts and products

Calculate the Hessian matrix of the TS start structure $\rightarrow \dfrac{\partial^2 E}{\partial (x^{\dagger,QM,ref})^2}$
if $N_{\nu_{im}}(x^{\dagger,QM,ref}) \geq 1$ **then**
  Trust region image minimization $x^{\dagger,QM,ref} \rightarrow x^{\dagger}$
  if Geometry optimization is converged **and** $N_{\nu_{im}}(x^{\dagger}) == 1$ **then**
    Optimize the end points from IRC $\rightarrow x^{IRC}$
    Calculate their Hessian matrices $\rightarrow \dfrac{\partial^2 E}{\partial (x^{IRC})^2}$
    if $N_{\nu_{im}}(x^{IRC}) == 0$ **then**
      Compare $m_g^{(IRC)}$ with $m_g^{(QM,ref)}$
      Register new reaction and optimized species $(x^{\dagger}, x^{IRC})$ to the database
    end if
  end if
end if

2.7 Calculation of reaction energies and rate constants

After the initial refinement of the reactions with the computationally non-demanding density functional, the optimization of the geometries and the calculation of the Hessian matrices with a more accurate density functional and basis set combination is straightforwardly automated using the lower level DFT data as start guesses. The additional refinement is especially

important as the results of the DFT calculations like the molecular structures and harmonic vibrational frequencies enter the calculation of thermodynamic properties and rate constants (*vide infra*). Since the accuracy of DFT energies is often not sufficient for the requirements on reaction energies and rate constants it is usually necessary to get more accurate electronic energies from wave function based methods as, for example, coupled cluster theory.

The reaction rate constants are calculated using transition state theory (TST). The partition functions needed for this are calculated using a modified version of the quasi-rigid rotor and harmonic oscillator (qRRHO) approximation[38,39]. Using the ideal gas law as the equation of state, the site specific reaction rate constants (excluding the symmetry numbers) can be calculated at constant pressure within TST according to

$$k(T) = \kappa \frac{k_B T}{h} \frac{Q_{\mathrm{TS}}/V}{\prod_{i=educts}(Q_{\mathrm{i}}/V)} \exp\left(-\frac{\Delta E_{el,act}}{RT}\right) \tag{1}$$

where R is the gas constant, k_B Boltzmann's constant, V the molar volume, T the temperature, h Planck's constant, and Q is the partition function which includes translational, vibrational, and rotational contributions. The vibrational partition function (q_{vib}) is here defined relative to the bottom of the potential energy surface as

$$q_{vib} = \prod_i \frac{\exp(-h\nu_i/2k_BT)}{1-\exp(-h\nu_i/k_BT)}, \tag{2}$$

where ν_i are the vibrational frequencies. κ in Eq. 1 is the transmission coefficient used to account for tunneling. It is calculated using the high temperature limit of the Eckart tunneling correction:[40]

$$\kappa = 1 - \frac{1}{24}\left(\frac{h\nu_{im}}{k_BT}\right)^2 \left(1 + \frac{RT}{\Delta E_{el,act}}\right), \tag{3}$$

where ν_{im} is the imaginary frequency and $\Delta E_{el,act}$ is the electronic activation energy without zero point vibrational correction. For a compact representation of the results, the parameters A, n, and E for the modified Arrhenius equation,

$$k(T) = AT^n \exp\left(-\frac{E}{T}\right) \tag{4}$$

where T and E are in units of K, are fitted to the rate constants calculated via Eq. 1 with a non-linear least square fitting for the temperature range of 300–3000 K.

3 SHOWCASE STUDY

In the following, we want to demonstrate the presented automated workflow on the basis of studying the combustion of PAHs under oxyfuel conditions. First, we summarize the computational settings used in our study in Section 3.1. In Section 3.3 the introduced filtering scheme is assessed and how the composition of the gas phase affects the number of reactions. Afterwards follows a qualitative analysis of found reaction types in Section3.5 and in Section 3.6 the quantitative assessment of reactions employing the automated refinement via DFT calculations.

3.1 Computational details

The MD simulations were done with our in-house python interface to the LAMMPS Molecular Dynamics Simulator[41] employing the reimplementation of the reactive Force-Field ReaxFF[42]. For the ReaxFF calculations the "CHO" parameterization for combustion as published in Ref. 14 was used. All MD simulations reported in this study use the following settings if not otherwise stated: Periodic boundary conditions in a cubic simulation box were applied and depending on the case the size varied from 20 – 25 Å. The simulations were run for 500 ps with a time step of 0.1 fs at 2000 K using a Nosé–Hoover thermostat. Each simulation was repeated 5 times using different initial random velocities.

For the manipulation of graphs the graph-tools library[34] is applied and the SQLite[43] database abstraction layer of web2py[44] is used to manage the relational database to store reaction events, reaction lists as well as the molecular species.

All quantum chemical calculations were performed with the TURBOMOLE program package[28] version 7.5[29]. For the DFT calculations the ridft module was used. Basis sets applied in the DFT calculation include the TZVP[45] and SVP[46] basis sets together with the corresponding auxiliary basis sets for the resolution of the identity approximation from Ref. 47. Furthermore, Grimme's D3 dispersion correction[48] was included. As functionals TPSS[49] and TPSSh[50] were chosen. The TPSSh/TZVP zero point vibrational frequencies have been scaled by the factor of 0.9874 that was optimized for TPSSh/6-31+G(d,p) to fit the experimental zero point vibrational energies (ZPVE)[51].

In addition, explicitly correlated coupled cluster singles doubles and perturbative triples, CCSD(F12*)(T*) calculations were carried out with the ccsdf12 module[52] based on preceding restricted open-shell Hartree-Fock (ROHF) calculations with the dscf module[53]. The cc-pVDZ-F12 orbital[54], auxiliary[55,56], and complementary auxiliary basis sets[57] are used with the recommended exponent[54] γ=0.9 for the correlation factor $\exp(-\gamma r_{12})$.

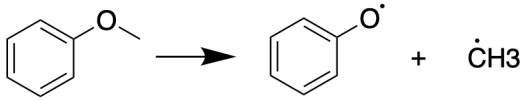
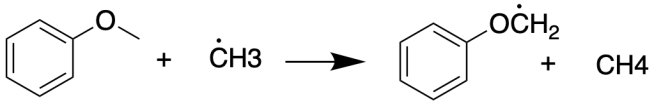
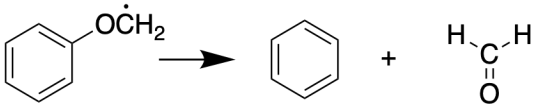
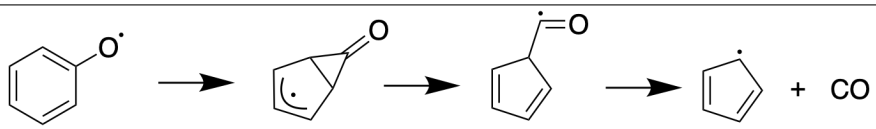
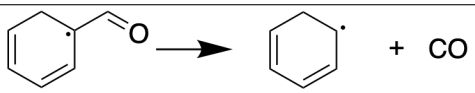
3.2 Verification

As a first step, we aim to reproduce the elementary reactions known in particular processes. For this purpose we choose the combustion and pyrolysis of anisole. The reference data is taken from Ref. 58. We performed MD simulations including one or several ansiole molecule(s) in the simulation box with and without O_2 atmosphere and depending on the cases OH radicals. The simulation box had a size of 20 Å. Straight out of the box, it is possible to reproduce the main reaction channel of the pyrolysis and oxidation in the main categories "initiations", "H-abstraction" and "radical decomposition" as described in Ref. 58. The reaction classes are depicted in Table 1 with examples.

The automated workflow detected the practically barrierless anisole initiation by methyl radical release. Additionally, the H abstraction reaction via a methoxy radical as illustrated in Table 1 was found. The release of formaldehyde from the H-abstracted anisole radical depicts the radical decomposition reaction class. There were also some reactions from the benzene and phenol oxidation sub-mechanisms. We face problems to locate elementary reactions involving cyclopentadienyl as reactant, but this is due to a relative inertness of this molecule using ReaxFF. Overall these examples reveal that *findR* can detect important elementary reaction steps and is useful for the study of combustion processes.

Additionally, we have compared the reaction rate constants calculated at the temperature range 700-1100 K and shown in Fig. 5 for the example radical decomposition, hydrogen abstraction, and the first step of the bicyclic intermediate formation from the phenoxy radical as part of the phenol oxidation sub-mechanism. Overall, it can be seen that the relative ratio of reaction rate constants follow similar trends as with the reference ones, which is important, for example, to obtain realistic branching ratios. The automatically calculated activation energy and reaction energy of the decomposition of the phenoxymethyl

Table 1: Identified reaction pathways which are in agreement with the anisole pyrolysis and oxidation mechanism used in Ref. 58

Reaction	Reaction class
	anisole initiation
	hydrogen abstraction
	radical decomposition
	phenol oxidation
	benzene oxidation

to phenyl and formaldehyde ($C_6H_5OCH_2 \rightarrow C_6H_5 + CH_2O$) are 145 kJ/mol and 102 kJ/mol, respectively using the CCSD(F12*)(T*) energies, and similar to the reference values from Ref 59 calculated using G3X composite method as 162 kJ/mol and 106 kJ/mol respectively. The reaction rate constants from this study is also shown in Figure 5. In this study, the activation and reaction energies of the H abstraction from anisole with a methyl radical ($C_6H_5OCH_3 + CH_3 \rightarrow C_6H_5OCH_2 + CH_4$) were calculated as 52.6 kJ/mol and -27.8 kJ/mol. These are similar to values reported in the previous studies, for example, using M06-2X they were respectively calculated as 59.0 kJ/mol and -29.9 kJ/mol[60]. It should be noted that the reference kinetic parameters taken for the H-abstraction that are used for comparison in Figure 5 were obtained by Mulcahy et al.[61] at the pressure and temperature range of 0.01-0.004 bar and 453-539 K. The reaction of formation of bicyclic intermediate from the phenoxy radical (C_6H_5O) is compared with two different reference rate constants calculated using CBS-QB3 method at 1 atm and 300-2000 K range by Carstensen and Dean[62] and with that from the theoretical and experimental studies of Nowakowska et al.[58].

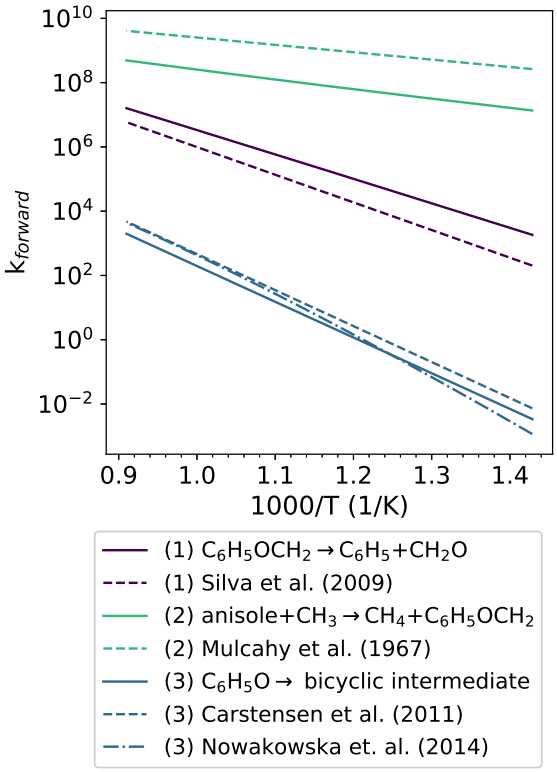


Figure 5: Comparison of the rate constants calculated at 1 bar in this study (shown in solid lines) with the reference data at the temperature range 700-1100 K.

The most important point is that these reactions were identified and could automatically be refined. Depending on the available computational resources it is always possible to refine the identified reactions further using more accurate electronic structure methods and/or additional corrections going beyond the harmonic rigid rotor approximation. Especially the use of composite schemes is attractive since our database layout makes it easy to retrieve the computed data on various LoTs.

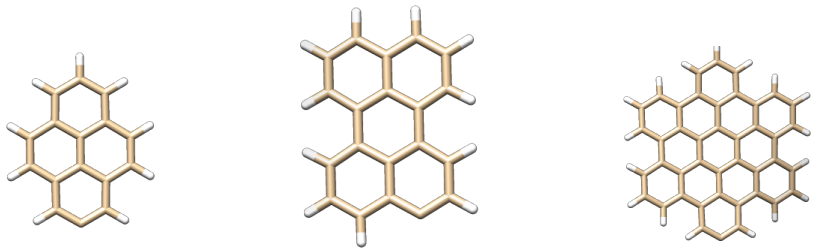


Figure 6: Systems (without O_2 and OH atmosphere) studied by ReaxFF MD simulation and analysis by *findR*.

3.3 Reaction screening

For the initial exploration, we used the three systems shown in Fig. 6 as examples. We run simulations in an atmosphere enriched with OH radicals (40 O_2 molecules and 10 OH radicals) as well as in a pure O_2 atmosphere to investigate if these conditions lead to different reaction types. In all cases we started from a PAH, where already one H atom was abstracted to enhance reactivity. The H abstraction process was not studied in detail here since we already investigated it in a recent publication[38].

First we want to address the question of how many unique reactions according to the defined uniqueness measure can be identified in the MD simulations and how this will increase performance in the subsequent analysis on the basis of for example quantum chemical calculations. Table 2 gives an overview of the number of non-redundant and total reactions found in the MD simulations in the OH enriched atmosphere, which were repeated each 5 times with different initial velocities. We further differentiate in the filtered reactions if the molecular C graph changed upon reaction.

In all cases a large reduction in the number of reactions is observed. Especially for the larger systems savings above 90% are possible. This translates directly to less computational costs for the post-processing with more accurate quantum chemical methods, e.g. DFT, and during this stage reactions can again be added to the database if it turns out that for a specific example it is important to account for different conformations.

Table 2: Statistics of filtered reactions in OH enriched atmosphere. Given are the found unique reactions $n_{\bar{r}}$ as well as all found reactions n_r. Additionally, the number of reactions is given in which the molecular C-graph changed upon reaction $n_{\bar{r}}^{(C)}$ and thus an important step in the combustion process happened. The screening ratio "Screen.", defined as the fraction of reactions that are skipped, given in %.

	(a) $C_{16}H_9$					(b) $C_{20}H_{11}$					(c) $C_{42}H_{17}$				
run	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5
n_r	268	1348	853	1265	696	1061	775	2204	483	912	5943	7974	2458	159	1405
$n_{\bar{r}}$	43	143	169	134	147	71	209	260	91	554	195	281	100	48	72
Screen.	84	89	80	89	79	93	73	88	81	39	97	96	96	70	95
$n_{\bar{r}}^{(C)}$	10	43	49	44	49	13	84	83	46	63	51	95	22	13	17

3.4 Influence of radical enriched atmosphere

Next, we address the question of how a radical enriched atmosphere can trigger reaction events in the sense that more reactions occur. Unfortunately, performing ReaxFF simulations at experimentally realistic conditions is computationally too demanding and previous studies employed too high temperatures, which leads to a bias towards fragmentation instead of oxidation reactions. To assess how the radical atmosphere results in more reactions also MD simulations in a pure O_2 atmosphere were performed and are compared to the previous ones. The statistics over the found reactions are listed in Table 3. While the average number of all reactions increased, the number of unique reactions and in particular unique reactions including a change in the C-graph decreased in most cases. The on average larger number of reaction events can be explained by O_2 attaching to the PAHs and detaching again in a subsequent step. These non-reactive collisions of O_2 have also been observed in hydrocarbon oxidation simulations by Chenoweth et al.[14]. Furthermore, one can make a few more general observations: In those cases where many unique reaction events occur, some radical species in the atmosphere were created. If no gas phase radicals are involved, the reactions occur mainly in close proximity to the initial radical site at the PAH. At least at the ReaxFF level of theory one can conclude that the formation of radical species in the gas phase is an

important step in the combustion process.

Table 3: Statistics of filtered reactions in pure O_2. Given are the found unique reactions $n_{\bar{r}}$ as well as all found reactions n_r. Additionally the number of reactions is given in which the molecular C-graph changed upon reaction $n_{\bar{r}}^{(C)}$ and thus an important step in the combustion process happened. The screening ratio Screen. which represents the skipped reactions is given in %.

	(a) $C_{16}H_9$					(b) $C_{20}H_{11}$					(c) $C_{42}H_{17}$				
run	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5
n_r	1643	2494	1166	809	978	3829	3733	50	498	3933	601	3942	1842	1311	1673
$n_{\bar{r}}$	103	32	11	34	14	11	8	9	41	7	15	7	63	77	17
Screen.	94	99	99	96	99	99.9	99.9	82	92	99.9	98	99	97	94	99.9
$n_{\bar{r}}^{(C)}$	38	10	3	8	4	2	2	20	1		3	0	20	28	8

3.5 Identification of common elementary reactions

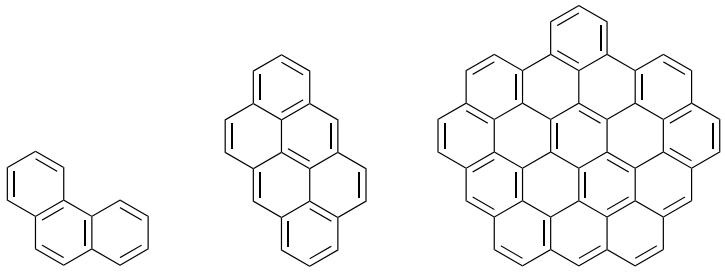


Figure 7: Model systems for the molecular dynamic simulations to identify key elementary reactions in the combustion process.

A key feature of *findR* is the identification of elementary reactions. Our current focus lies on the atomistic modelling of combustion reactions of char models. Therefore, it is essential that *findR* can identify common key reaction mechanisms. To assess this, we created an additional test set of systems depicted in Fig. 7 where we again varied the composition of the atmosphere. An overview can be found in Table 4. Each MD was done 5 times using

Table 4: Composition of the atmosphere in the molecular dynamic simulation as well as the studied system given in Fig. 7. For each system we repeated the simulation 5 times with different initial random velocities. Later in the text these simulations are referred to as run 1, run 2 etc.

Molecular Dynamic Simulation	System	hydroxyl radicals	oxygen molecules	water molecules
MD 1	(1)	10	40	-
MD 2	(1)	15	60	-
MD 3	(1)	10	40	10
MD 4	(2)	12	40	-
MD 5	(3)	12	40	-

different initial velocities. For this purpose, we generated reaction networks for each of the simulations and inspected them by hand to identify common reaction steps. This type of analysis only allows a qualitative description. In Section 3.6 we will present a more elaborate way to analysis reactions by means of post-processing with DFT.

Before discussing some common reaction steps that were found in the MD simulations, we want to discuss a few more general observations: For larger PAHs more reaction events are observed, which can be trivially explained by the larger number of reaction sites. Furthermore, we observed that the PAHs are decomposed starting from the exterior ends. Using different pathways, H atoms are abstracted, and radical species attach to the PAHs and initiate rearrangements. In these rearrangements some common reaction steps are observed like the formation of five-membered rings as well as the opening of 6-rings. In the following, we will list a few typical reactions found in the simulations. We give the examples as found in the MD, but note that they represent general reactions that can occur on outer rings of PAHs or graphene-like areas in char.

Shrinkage of a 6-membered ring usually happens via the formation of an additional bond that divides it into a 5- and a 3-membered ring. In subsequent steps the latter is opened and a fragment, e.g. CO_2 or formic acid, is released. Examples of these are shown in Figure 8

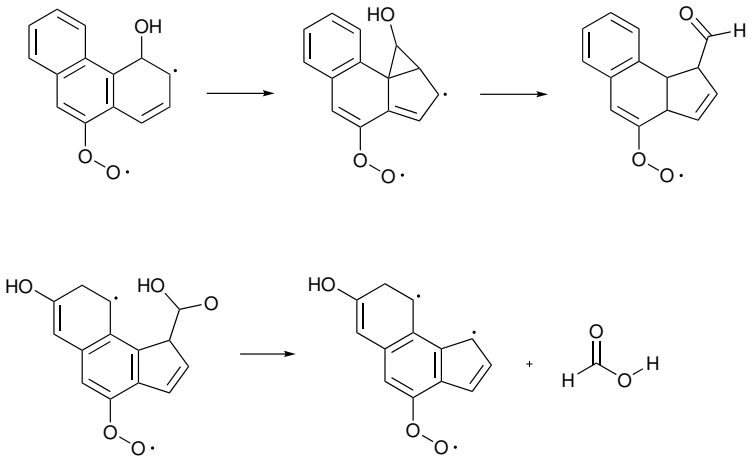


Figure 8: Top: Formation of a 5-membered ring via ring shrinkage. Bottom: Release of formic acid from a 5-membered ring.

Ring extension usually proceeds via insertion of O into the ring structure after a preceding formation of an epoxide as shown in Figure 9.

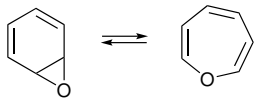


Figure 9: Benzene-oxide/oxepin tautomers.

The opening of a 6-ring occurs due to homolysis after oxidation of two neighbouring carbon positions to a bicarbonyl such as represented in Figure 10.

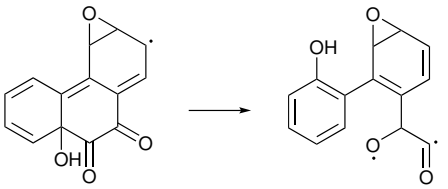


Figure 10: An example ring opening reaction.

We observed several times that formic acid is a favoured leaving group in a OH radical enriched atmosphere, while in atmospheres with lower OH radical concentration CO and CO_2 are favoured leaving groups as depicted in Figure 11.

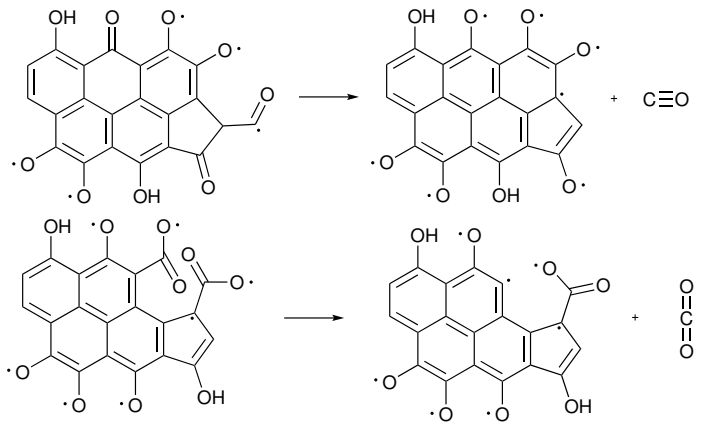


Figure 11: The release of CO and CO_2.

3.6 Quantum Chemical Calculations

Table 5: Educt, transition state, and product structures optimized at the TPSSh-D3/TZVP level for some reactions detected in the run 3 trajectory of MD3. Spin densities obtained from ROHF calculations are shown as isosurface plots. In the last two columns we give the CCSD(F12*)(T*) results for activation and reaction energies including the ZPVE corrections from TPSSh-D3/TZVP, $\Delta E = \Delta E_{\mathrm{el,CCSD(F12*)(T*)/cc-pVDZ-F12}} + \Delta E_{ZVPE,\mathrm{TPSSh-D3/TZVP}}$

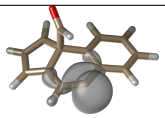
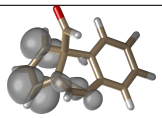
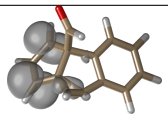
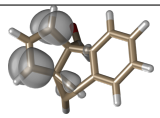
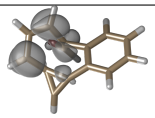
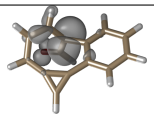
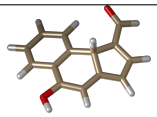
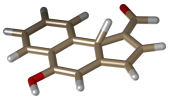
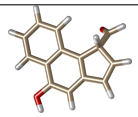
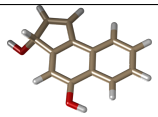
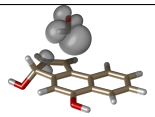
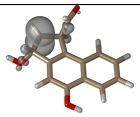
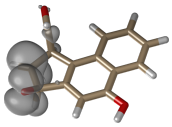
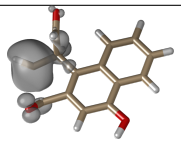
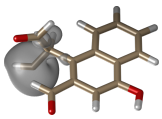
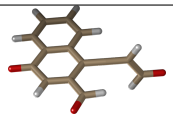
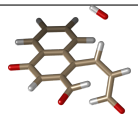
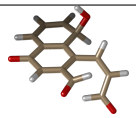
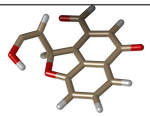
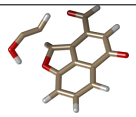
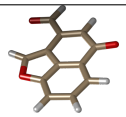
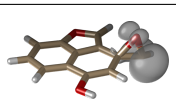
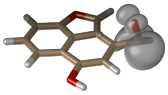
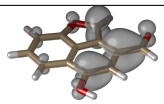
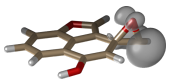
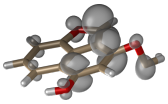
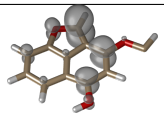
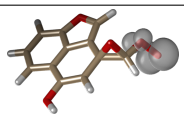
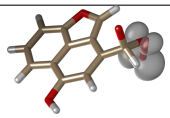
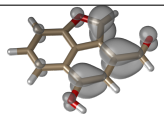
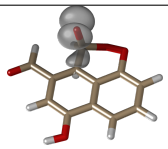
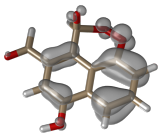
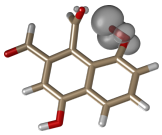
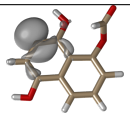
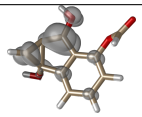
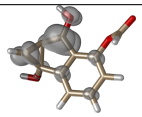
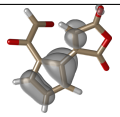
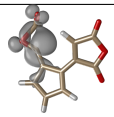
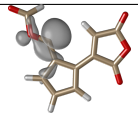
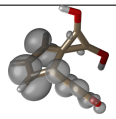
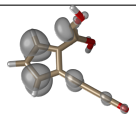
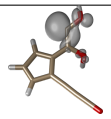
	Educt(s)	Transition State	Product(s)	ΔE_{act} kJ/mol	ΔE_{rxn} kJ/mol
1	 m_s=1/2	 m_s=1/2	 m_s=1/2	122.3	116.3
2	 m_s=1	 m_s=1	 m_s=1	82.5	16.9
3	 m_s=0	 m_s=0	 m_s=0	77.6	-54.0
4	 m_s=0 CHO m_s=1/2	 m_s=1/2	 m_s=1/2	20.1	-77.8
5	 m_s=1	 m_s=1	 m_s=1	80.4	62.5
6	 m_s=0 H_2O m_s=0	 m_s=0	 m_s=0	163.1	64.7
7	 m_s=0	 m_s=0	 m_s=0 ethynol m_s=0	7.4	-86.1

Table 6: The species geometry optimized at TPSSh-D3/TZVP level as obtained from the run 3 trajectory of MD3. The ROHF spin densities are shown. $\Delta E = \Delta E_{\mathrm{el,CCSD(F12*)(T*)/cc-pVDZ-F12}} + ZVPE_{\mathrm{TPSSh-D3/TZVP}}$ (cont.)

	Educt(s)	Transition State	Product(s)	ΔE_{act} kJ/mol	ΔE_{rxn} kJ/mol
8	 m_s=1/2	 m_s=1/2	 m_s=1/2	12.9	-258.1
9	 m_s=1/2	 m_s=1/2	 m_s=1/2	140.7	47.6
10	 m_s=1/2	 m_s=1/2	 m_s=-1/2 O_2 m_s=1	46.8	-328.2
11	 m_s=1/2	 m_s=1/2	 m_s=1/2	41.6	1.1
12	 m_s=1/2	 m_s=1/2	 m_s=1/2	202.5	201.5
13	 m_s=1/2	 m_s=1/2	 m_s=1/2	191.8	111.4
14	 m_s=1/2	 m_s=1/2	 m_s=1/2	52.4	-29.5

So far, we did a mostly qualitative assessment of the reaction events found by *findR*. Let us now focus on the next step, the automated refinement with quantum chemical methods, which allows for a more quantitative assessment.

As mentioned before, due to the approximate nature of ReaxFF, it is not always straightforward to refine the reactions found in the trajectories of MD simulations with static DFT or other quantum chemical calculations. In the following, we want to discuss some of these challenges on the basis of a particular trajectory we analysed. For this purpose we selected the trajectory from run 3 for MD 3 (cmp. Table 4) as example since the PAHs in this MD are still relatively small and allow even Coupled Cluster calculations. Furthermore, this trajectory contains some examples for the specific types of challenges we faced and want to address in this work. Tables 5–6 summarize for this trajectory the reactions that were identified and successfully refined with the automated workflow until there was only one ring left. In Tables 5–6, reactions in which the molecular graph does not change, such as CHO or OH internal rotations, and practically barrierless reactions, such as many cases of O_2 addition or elimination, are not shown. The assignment of the reaction directions is in general arbitrary (as explained in Section 2.6). Only if the molecular graphs of the IRC end points (x^{IRC}) match those of the educts and products ($x^{QM,ref}$) that were obtained by refinement of the corresponding ReaxFF structures, the reaction directions are defined in accordance with the MD trajectory. The activation and reaction energies given in Tables 5–6 include the zero point vibrational energy (ZPVE) corrections at the TPSSh-D3/TZVP level of theory and the electronic energies from the ROHF-CCSD(F12*)(T*)/cc-pVDZ-F12 calculations. The rate constants for all elementary reaction steps shown in these tables were calculated according to eq. 1 and parameters for the modified Arrhenius equation (eq. 4) were fitted to them. The Arrhenius parameters are given in the Supporting Information.

One of the challenges for the automatic refinement is related to unimolecular reactions. Among the 73 unique unimolecular reactions found in the example trajectory it happened for 53 of them that the educt and product structures extracted from the trajectory converged during the ReaxFF geometry optimization to equilibrium structures with the same molecular graph, indicating an intermediate bond breaking or formation that might have been reverted after a few time steps rather than a true reaction, which could be detected during the refinement. The reason why the educt and product structures from ReaxFF are the same is that these structures are, as explained in section 2.4, extracted directly from the trajectories of the MD simulations that are run at 2000 K where the molecular structure undergoes large

deformations that, even though they lead to changes in the ReaxFF bond order, are often not real reactions in the sense of transitions from one potential well over a saddle point to another well on the PES. That is, as the educt and product are not located through an intrinsic reaction coordinate search from an optimized transition state, it is not guaranteed that they are close to two different equilibrium structures, i.e. minima on the PES, which are connected by a transition state. Nevertheless, the Python module for the automated refinement tries for these cases to locate a saddle point on the DFT PES starting from the pre-optimized ReaxFF transition state structure in order to find the saddle point on the DFT PES as well as the educt and product structures that are connected to the located transition state. Reactions 1, 2, 12, and 14 are such examples, where geometry optimizations of the structures taken from the MD trajectory ended up at the ReaxFF level in the same equilibrium structures, but where the ReaxFF pre-optimized transition state structures were successfully used as start guesses for the optimization of a transition state at the DFT level. As it can be seen in Table 5, they correspond to either C-C bond formation or breaking. For all four of these reactions the products are, however, structurally rather close to the transition state, which partially explains why on the ReaxFF PES no minima were found that correspond to the product structures on the DFT PES.

For those unimolecular reactions where the transition state guess from the ReaxFF structure was not a good initial structure, the reaction path calculation either located a conformational change such as rotation of a -OH group or it was not carried out because educt and product were the same (see Section 2.6). The automated refinement workflow located for 24 of the 73 unique unimolecular reaction events a saddle point on the DFT PES. 14 of them belong to reactions with different molecular graphs for the educt and product.

Other difficulties arise when the MD trajectory deviates qualitatively from (minimum energy) reaction pathways on the DFT PES. For example, the educt of reaction 1 in Table 5 is on the MD trajectory formed through several elementary reaction steps where the automated workflow has been challenged. The first elementary step has been in the MD trajectory a simultaneous addition of OH and O_2. As the automated refinement with DFT has not been implemented for termolecular reactions, this step has been skipped in the proof-of-concept study. The following step, H abstraction from the OH group with another OH radical, results

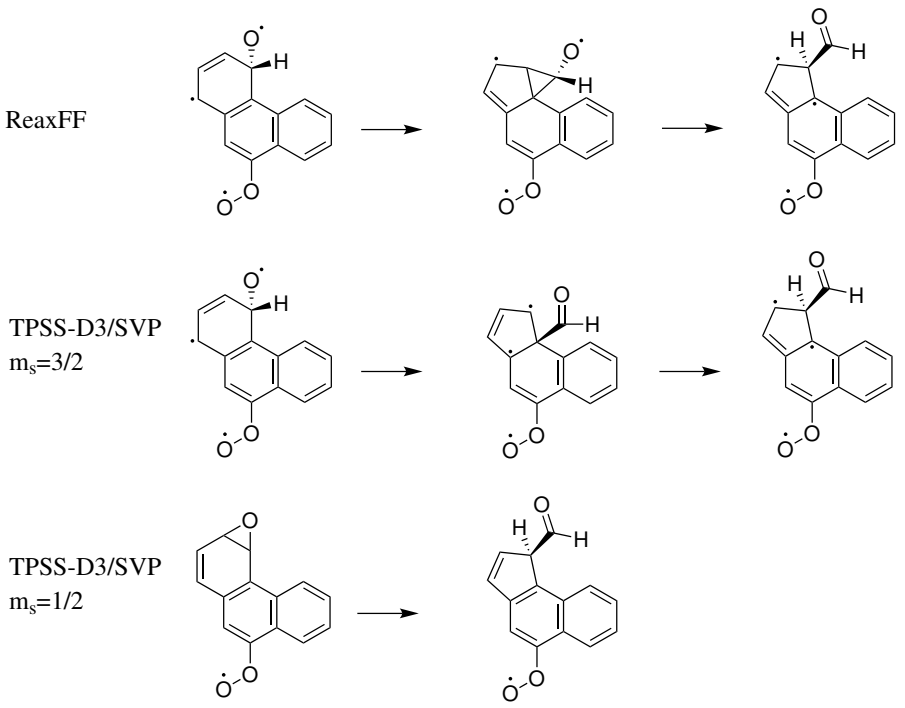


Figure 12: The elementary steps of the formation of the formyl radical following the formation of an oxy-radical or an epoxide.

in an oxy radical (R-O·) which is stable only with three unpaired electrons (and a water molecule). On the PES for one unpaired electron, the equilibrium structure converges to an epoxide. Starting from the oxy radical or the epoxide, the following elementary steps of the formyl group formation are as shown in Figure 12. A problem for the H abstraction from the R-OH group by a hydroxy radical is the inability of ReaxFF to correctly describe radicals. As shown in Figure 13, the ReaxFF optimized educt has a structure that is intermediate between a water molecule bonded to an oxy radical and a hydroxyl radical bonded to an OH functional group, because 111 pm is obviously too short to be an intermolecular bond. Indeed, when optimized at the DFT level, a water molecule separates from the rest. In the future one can potentially also avoid these difficulties in the first place by using MD simulations based on methods like the extended tight binding methods[63,64] (xTB), which should more closely resemble the DFT PES. For the outlined structure in Figure 13 we also optimized the molecular species obtained from the ReaxFF MD simulation on the GFN0-xTB, GFN1-xTB and GFN2-xTB level. All structures optimized on the xTB level show a

better agreement with the DFT optimized ones than ReaxFF. This does especially hold for GFN1-xTB and GFN2-xTB.

It can occur that ReaxFF optimized species exhibit unusual bonding, which is then resolved upon optimization on the DFT level. If this optimization leads to fragmentation of the species, the reaction is currently flagged for inspection, but in the future it is planned to identify the reactive center(s) and register new reactions in the database. To keep track of which reactions originate from each other, our database layout has the `origin` cell which can relate different reactions.

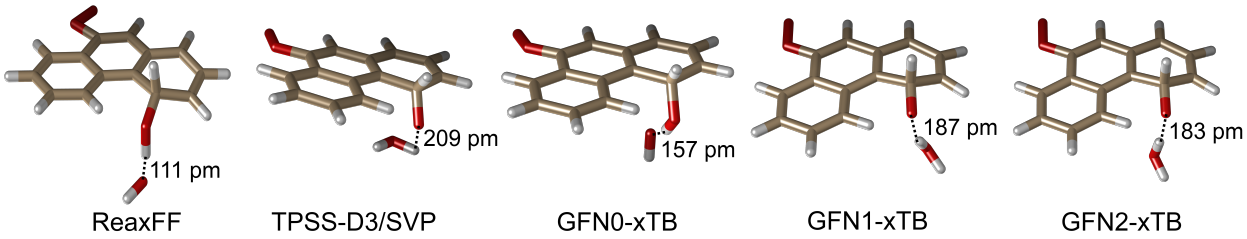


Figure 13: Comparison of the ReaxFF, DFT ($m_s = 3/2$), and xTB optimized educt structures for the detected dehydration reaction.

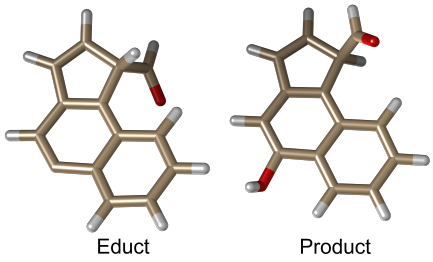


Figure 14: The educt and product structures of the detected OH addition, as geometry optimized with ReaxFF.

During the ReaxFF simulations, functional groups connected to the ring systems by single bonds, e.g. alcohol (-OH) or aldehyde (-CHO) groups, rotate or that stereo centers isomerize in some occasions. The latter results in educt and product structures for a detected reaction where these functional groups are flipped relative to the plane of the aromatic rings as shown in Figure 14. The reason, why the stereoisomerism is not detected by *findR* is that, as explained in section 2.4, only the composition and the bonding of the atoms (based

on the molecular graphs) are used to detect the reaction events. Due to the two simultaneous reactions happening at the same time and the large structural change during the stereoisomerisation, the default number of iteration steps used during the woelfling calculations (maximum of 40 iterations) as well as the number of intermediate structures along the path might not be enough to optimize a good approximate reaction path. But still, the structures with maximum energy along the approximate reaction path obtained with the woelfling program may be sufficiently good start structures for the location of saddle points and converge to a transition state structure upon trust region image minimization. Reaction 3, migration of an H atom to a neighbouring carbon, was located this way. The detected reaction event from the MD trajectory was the addition of an OH radical, whose detected educt and product structures are shown in Figure 14, however, the formyl group flips along the trajectory very shortly after the OH is chemisorbed. Also, reaction 4, addition of the formyl radical was found similarly during the automated DFT refinement. The reaction event detected in the MD trajectory was again the addition of an OH radical. Thus, the quantum chemical refinement can in those cases resolve the issues related with *findR* not differentiating stereoismers. In the future a configurational sampling step for each involved reaction species can be included. The setup of the database is general enough to register new reactions at any step and to relate them to each other, which can be used to insert conformational changes as additional elementary reaction steps.

Even without simultaneous stereoisomerization, it can happen for unimolecular reactions that the implemented multi-step refinement of the reactions at the DFT locates alternative reactions where for example the initial educts and products ($x^{QM,ref}$) are different from the end points of the IRC pathways (x^{IRC}) starting from the transition state. For example, reaction 6 (cmp. Table 5) is found during the reaction path optimization of an OH addition: when the OH radical approaches the molecule, it passes through a structure close to the transition state shown in Table 5 as maximum on the energy profile along the reaction path. The optimized transition state, however, belongs to an addition of a water molecule. Another such example is reaction 11, which was located starting from the pre-optimized ReaxFF transition state. In that case the DFT structures for the IRC end points match to the molecular graphs of the non-optimized structures from the MD trajectory, while the

initial geometry optimizations with ReaxFF for these structures converged for educt and product to the same minimum, which could then not be used to locate a reaction.

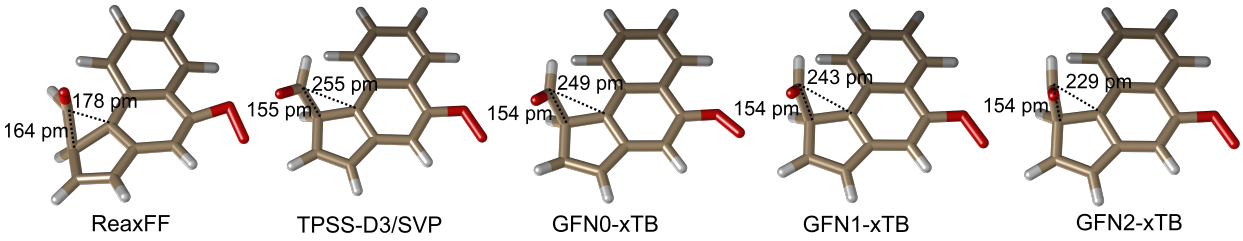


Figure 15: Comparison of the ReaxFF, DFT ($m_s = 1/2$), and xTB optimized structures.

Challenging for the DFT refinement are cases where the ReaxFF optimized equilibrium structures are far from minima on the DFT PES with the same molecular graph or where corresponding minima do not at all exist on the DFT PES. An example for such a case is shown in Figure 15, which is at the ReaxFF level a stable structure with 5- and 3-membered rings. As mentioned above in the text and in Figure 12, in the MD simulation the formation of the formyl group occurs through ring shrinkage of a 6-membered ring to 5- and 3-membered rings. However, neither on the $m_s = 3/2$ or $m_s = 1/2$ DFT PESs such a 3-membered ring exists as an equilibrium structure. On the DFT PES such a structure is instead rather close to a transition state. The depicted structure in Figure 15 has also been geometry optimized with xTB methods starting from the MD structure. For this species as well, the structures optimized on the xTB level show a better agreement with the DFT optimized ones than those obtained with ReaxFF. In a previous study Frenklach et al.[65] reported that for 4-phenanthrol a similar shrinkage of a 6-membered ring through expulsion of a -CHO group happens in a single step. It is known that the reaction mechanisms at different reactive sites depend on the rigidity of the molecule at the reactive site[66] and ReaxFF might not be able to capture these subtle differences. Therefore, use of MD simulations based on xTB methods might be a solution to such cases. As on the DFT PES this reaction happens at this reaction site in a single step, it is not possible to find an equilibrium structure at the DFT surface whose graph would not be different from that of the start structure obtained with ReaxFF. Hence, in such cases elementary reaction steps on the ReaxFF PES do not have a direct counterpart on the DFT PES. For these cases the equilibrium structures found for

the educts and products on the DFT PES are added as new species to the database and the saddle point(s) connecting them are located starting with a reaction pathway optimization with these species as start and end points.

Another example where the ReaxFF optimized equilibrium structures are far from minima on the DFT PES is shown on the left of the Figure 16. The program *findR* detected an O_2 addition to the ReaxFF optimized structure shown in the upper left of Fig. 16, which has the same molecular graph as that of the educt of reactions 8 and 9. For the detected reaction, the epoxide ring is in the ReaxFF optimized educt structure almost parallel to the plane of the aromatic rings. The unrelaxed scan of the CCCH torsional angle (see right part of Figure 16) shows that this ReaxFF structure is far from the equilibrium structure at the DFT level which has the same molecular graph. For the DFT equilibrium structures where the CCCH torsional angle either is around 160 ° or around 340 °, the epoxide group is almost perpendicular to the aromatic ring plane. Because of this mismatch between the ReaxFF and DFT equilibrium structures, a geometry optimization at the DFT level starting from the ReaxFF structure ends up in another equilibrium structure where the epoxide ring has opened to a formyl group. On the other hand, the transformation of the formyl group to the epoxide is highly endothermic (reverse reaction 8), so that the reaction found in the DFT refinement is much more likely than that originally detected on the MD trajectory calculated with ReaxFF. As can be seen in the following reactions, the ether is also converted back to a formyl group. Thus, for modelling the kinetics of combustion processes, the pathway via the unstable epoxide intermediates that are not found by the automated DFT refinement procedure does not play a significant role.

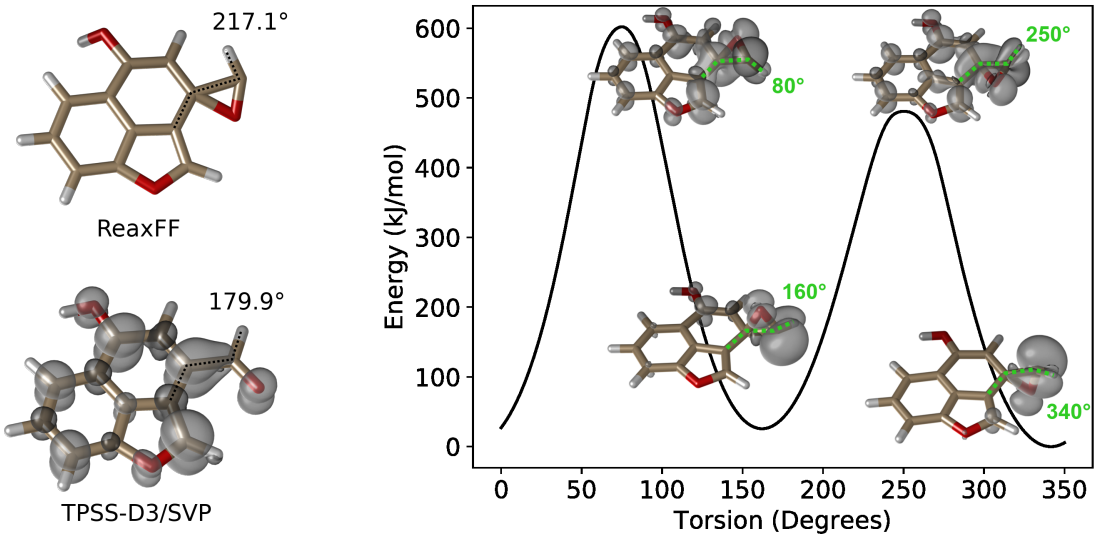


Figure 16: Left: Comparison of the ReaxFF and DFT optimized structures for an educt where the ReaxFF equilibrium structure was far from the minimum which describes the equilibrium structure with the same molecular graph on the DFT PES. Right: The unrelaxed scan of the CCCH torsional angle of the educt of the reactions 8 and 9 using TPSS-D3/SVP.

For the majority of the relevant reactions, the multi-step refinement at the DFT level locates the reaction pathway and the transition state without problems. For example, reaction 5 is successfully located starting from the ReaxFF pre-optimized transition state. In reaction 7 the ethynol group (HCCOH) leaves the molecule. This reaction is successfully located starting from the ReaxFF pre-optimized transition state as well. The reaction 10, an O_2 elimination, is another reaction where the IRC end points match the reference educt and the product structures from ReaxFF. The transition state of this reaction is found through a start guess from the reaction path optimization. In reaction 13 the ketone group is converted to an ether group. Moreover, for this case the automated refinement procedure could locate the transition state on the DFT PES starting directly from the pre-optimized ReaxFF transition state structure.

Despite the problems that emerged during the automatic refinement, overall these problems do not circumvent our approach to identify important elementary reactions steps, and hence, identify important reaction classes for different active sites. Especially the observed

problems related to the inaccuracies of the ReaxFF PES can be lifted by using complementary methods which resolve the electronic structure such as semi empirical methods. Alternatively, Machine Learning (ML) Force Fields can be applied as well in the future. The introduced machinery could actually be a key ingredient in the selection of training points for ML and represent a form of active learning especially if the ML FF is used for MD and retrained based on the selected reactions.

4 CONCLUSIONS

We presented an automatic workflow which uses molecular dynamic simulations to sample possible elementary reactions which are refined with quantum chemical methods, in the current study DFT and coupled cluster, to obtain accurate thermodynamic and kinetic parameters. At the example of combustion reactions for polycyclic aromatic hydrocarbons, the capabilities of the workflow are demonstrated. The implemented automatic filtering technique reduces the number of reactions events that need to be refined with quantum chemical calculations dramatically by 70 to 90% and allows to restrict the latter to non-redundant reactions, which dominate the combustion process and involve C-C bond breaking. To focus on combustion, the simulation was performed at high temperature (2000 K) and the combustion parameterization for ReaxFF was employed.

To enhance the rate of observed reaction events, we started with PAHs containing already radical sites and an atmosphere enriched with OH radicals. The oxygen and radical concentration in the system influences the outcome of the observed reactions. For low hydroxyl radical concentrations, we observed that carbon monoxide and dioxide are the favoured products. With increasing hydroxyl radical concentration more formic acid-like species are formed. However, in all cases the steps for the decomposition of the carbon structure remained the same. The carbon rings are opened from the exterior side or functional groups are formed via oxygen and/or OH addition and later leave the system. At the enhanced temperatures of the MD simulations, we did not observe that for example epoxide groups form on the aromatic ring system itself. They have only found as short-living reaction intermediates at outer positions. The current simulations suggest that the PAHs decompose

starting from the exterior parts. In the future it will be interesting to study if this trend remains if hetero atoms or defects are included in the aromatic ring structure.

Besides these qualitative findings automated DFT calculations were carried out for a subset of the reactions. In this process the transition states and the connecting educts and products are optimized. In general, exploring chemical space via ReaxFF simulations is a fast and useful way to create a reaction pool. However, though more expensive, this study shows that more accurate methods for the MD simulations should also be considered.

The usage of ReaxFF MD simulations for sampling possible reaction events causes some challenges for a subsequent refinement with electronic structure calculations since the PESs for those two methods can differ to a large extent. We observed for example O—H—O bond sequences with bond lengths that are far from equilibrium structures on the more accurate PESs, or ReaxFF equilibrium structures which either do not exist as minima on the DFT PES or are far away from their DFT counterparts. Furthermore, some reactions have been detected on the ReaxFF level, where the automated refinement could not verify these reactions on the DFT PES. However, in some of those cases the ReaxFF transition states could still be used to locate transition states on the DFT PES for related reactions, which then could be added to the reaction database.

We started with ReaxFF MD as precursor to sample reactions because of its high computational efficiency and broad usage in the literature[67–70] for modelling char combustion. In future work we plan to generalize the implementation of the automated trajectory analyser and workflow for MD simulations with semi-empirical electronic structure methods like GFN2-xTB. Because these methods already use an electronic wave function, they will already provide information about the secondary spin quantum numbers which reduces the amount of information that needs to be filled in during the refinement of the reactions at the DFT level. Semi-empirical MD simulations will also potentially diminish the discrepancies between the structures and reactions found in the MD trajectories and those existing on the DFT PESs and will allow to study the influence of heteroatoms for which no suitable parameterization for ReaxFF exists. Furthermore, a future addition to our workflow could be the incorporation of configurational sampling techniques for the molecular species involved in the reactions to mitigate problems related to such changes in the refinement of reaction

pathways and the location of transition states.

ACKNOWLEDGMENTS

We are indebted to Kai Leonhard, Felix Schmalz, and Lukas Krep for helpful discussions and the other authors of ChemTraYzer for giving access to their code. We also thank to the German Research Foundation (DFG) for funding this research within the framework of the SFB/Transregio 129 "Oxyflame" (project number 215035359).

Data availability

Data available in article supplementary material and further data available on request from the authors.

References

1. C. Riplinger, B. Sandhoefer, A. Hansen, and F. Neese, The Journal of Chemical Physics **139**, 134101 (2013).

2. C. Riplinger, P. Pinski, U. Becker, E. F. Valeev, and F. Neese, The Journal of Chemical Physics **144**, 024109 (2016).

3. G. Schmitz and C. Hättig, The Journal of Chemical Physics **145**, 234107 (2016).

4. A. Hansen, C. Bannwarth, S. Grimme, P. Petrović, C. Werlé, and J.-P. Djukic, ChemistryOpen **3**, 177 (2014), ISSN 2191-1363.

5. M. R. Harper, K. M. Van Geem, S. P. Pyl, G. B. Marin, and W. H. Green, Combustion and Flame **158**, 16 (2011), ISSN 0010-2180.

6. J. W. Allen, A. M. Scheer, C. W. Gao, S. S. Merchant, S. S. Vasu, O. Welz, J. D. Savee, D. L. Osborn, C. Lee, S. Vranckx, et al., Combustion and Flame **161**, 711 (2014), ISSN 0010-2180.

7. G. R. Magoon, J. Aguilera-Iparraguirre, W. H. Green, J. J. Lutz, P. Piecuch, H.-W. Wong, and O. O. Oluwole, International Journal of Chemical Kinetics **44**, 179 (2012).

8. G. R. Magoon and W. H. Green, Computers & Chemical Engineering **52**, 35 (2013), ISSN 0098-1354.

9. C. W. Gao, J. W. Allen, W. H. Green, and R. H. West, Computer Physics Communications **203**, 212 (2016), ISSN 0010-4655.

10. G. N. Simm and M. Reiher, Journal of Chemical Theory and Computation **13**, 6108 (2017).

11. S. Habershon, The Journal of Chemical Physics **143**, 094106 (2015).

12. P. M. Zimmerman, Journal of Computational Chemistry **34**, 1385 (2013).

13. A. C. T. van Duin, S. Dasgupta, F. Lorant, and W. A. Goddard, The Journal of Physical Chemistry A **105**, 9396 (2001).

14. K. Chenoweth, A. C. T. van Duin, and W. A. Goddard, The Journal of Physical Chemistry A **112**, 1040 (2008).

15. M. Döntgen, M.-D. Przybylski-Freund, L. C. Kröger, W. A. Kopp, A. E. Ismail, and K. Leonhard, Journal of Chemical Theory and Computation **11**, 2517 (2015).

16. M. Döntgen, F. Schmalz, W. A. Kopp, L. C. Kröger, and K. Leonhard, Journal of Chemical Information and Modeling **58**, 1343 (2018).

17. L. C. Kröger, W. A. Kopp, M. Döntgen, and K. Leonhard, Journal of Chemical Theory and Computation **13**, 3955 (2017).

18. L.-P. Wang, A. Titov, R. McGibbon, F. Liu, V. S. Pande, and T. J. Martínez, Nature Chemistry **6**, 1044 (2014), ISSN 1755-4349.

19. M. Yang, L. Yang, G. Wang, Y. Zhou, D. Xie, and S. Li, Journal of Chemical Theory and Computation **14**, 5787 (2018), pMID: 30351922.

20. S. Maeda and K. Morokuma, Journal of Chemical Theory and Computation **7**, 2335 (2011).

21. W. M. C. Sameera, S. Maeda, and K. Morokuma, Accounts of Chemical Research **49**, 763 (2016).

22. C. W. Lee, B. L. H. Taylor, G. P. Petrova, A. Patel, K. Morokuma, K. N. Houk, and B. M. Stoltz, Journal of the American Chemical Society **141**, 6995 (2019).

23. P.-L. Kang and Z.-P. Liu, iScience **24**, 102013 (2021), ISSN 2589-0042.

24. The HDF Group, *Hierarchical data format version 5* (2000-2010), URL `http://www.hdfgroup.org/HDF5`.

25. L. P. Cordella, P. Foggia, C. Sansone, and M. Vento, in *In: 3rd IAPR-TC15 Workshop on Graph-based Representations in Pattern Recognition, Cuen* (2001), pp. 149–159.

26. L. P. Cordella, P. Foggia, C. Sansone, and M. Vento, IEEE Transactions on Pattern Analysis and Machine Intelligence **26**, 1367 (2004).

27. N. M. O'Boyle, M. Banck, C. A. James, C. Morley, T. Vandermeersch, and G. R. Hutchison, Journal of Cheminformatics **3**, 33 (2011), ISSN 1758-2946.

28. S. G. Balasubramani, G. P. Chen, S. Coriani, M. Diedenhofen, M. S. Frank, Y. J. Franzke, F. Furche, R. Grotjahn, M. E. Harding, C. Hättig, et al., JOURNAL OF CHEMICAL PHYSICS **152** (2020).

29. *TURBOMOLE V7.5 2020, a development of University of Karlsruhe and Forschungszentrum Karlsruhe GmbH, 1989-2007, TURBOMOLE GmbH, since 2007; available from* `https://www.turbomole.org`.

30. P. Deglmann, F. Furche, and R. Ahlrichs, Chemical Physics Letters **362**, 511 (2002), ISSN 0009-2614.

31. P. Deglmann and F. Furche, The Journal of Chemical Physics **117**, 9535 (2002).

32. T. Helgaker, Chemical Physics Letters **182**, 503 (1991), ISSN 0009-2614.

33. K. Fukui, Acc. Chem. Res. **14**, 363 (1981).

34. T. P. Peixoto, figshare (2014), URL `http://figshare.com/articles/graph_tool/1164194`.

35. P. Plessow, Journal of Chemical Theory and Computation **9**, 1305 (2013).

36. O. Weser, Master's thesis, Georg August Universität Göttingen (2017).

37. B. Helmich and M. Sierka, Journal of Computational Chemistry **33**, 134 (2012).

38. O. Yönder, G. Schmitz, C. Hättig, R. Schmid, P. Debiagi, C. Hasse, A. Locaspi, and T. Faravelli, The Journal of Physical Chemistry A **124**, 9626 (2020).

39. S. Grimme, Chemistry – A European Journal **18**, 9955 (2012).

40. I. Shavitt, The Journal of Chemical Physics **31**, 1359 (1959).

41. S. Plimpton, Journal of Computational Physics **117**, 1 (1995), ISSN 0021-9991.

42. H. M. Aktulga, J. C. Fogarty, S. A. Pandit, and A. Y. Grama, Parallel Comput. **38**, 245–259 (2012), ISSN 0167-8191.

43. R. D. Hipp, *SQLite* (2020), URL `https://www.sqlite.org/index.html`.

44. M. Di Pierro, *Chapter 6: The database abstraction layer.*

45. A. Schäfer, C. Huber, and R. Ahlrichs, The Journal of Chemical Physics **100**, 5829 (1994).

46. A. Schäffer, H. Horn, and R. Ahlrichs, J. Chem. Phys. **97**, 2571 (1992).

47. K. Eichkorn, O. Treutler, H. Öhm, M. Häser, and R. Ahlrichs, Chem. Phys. Lett. **242**, 652 (1995).

48. S. Grimme, J. Antony, S. Ehrlich, and H. Krieg, The Journal of Chemical Physics **132**, 154104 (2010).

49. J. Tao, J. P. Perdew, V. N. Staroverov, and G. E. Scuseria, Phys. Rev. Lett. **91**, 146401 (2003).

50. V. N. Staroverov, G. E. Scuseria, J. Tao, and J. P. Perdew, The Journal of Chemical Physics **119**, 12129 (2003).

51. G. I. Csonka, A. Ruzsinszky, and J. P. Perdew, The Journal of Physical Chemistry A **109**, 6779 (2005), pMID: 16834032.

52. C. Hättig, D. P. Tew, and A. Köhn, The Journal of Chemical Physics **132**, 231102 (2010).

53. M. Häser and R. Ahlrichs, Journal of Computational Chemistry **10**, 104 (1989).

54. K. A. Peterson, T. B. Adler, and H.-J. Werner, The Journal of Chemical Physics **128**, 084102 (2008).

55. C. Hättig, Phys. Chem. Chem. Phys. **7**, 59 (2005).

56. F. Weigend, A. Köhn, and C. Hättig, The Journal of Chemical Physics **116**, 3175 (2002).

57. K. E. Yousaf and K. A. Peterson, The Journal of Chemical Physics **129**, 184108 (2008).

58. M. Nowakowska, O. Herbinet, A. Dufour, and P.-A. Glaude, Combustion and Flame **161**, 1474 (2014), ISSN 0010-2180.

59. G. da Silva and J. W. Bozzelli, The Journal of Physical Chemistry A **113**, 6979 (2009), pMID: 19496593, https://doi.org/10.1021/jp902458d, URL https://doi.org/10.1021/jp902458d.

60. T. Zhang, C. Bhattarai, Y. Son, V. Samburova, A. Khlystov, and S. A. Varganov, Energy & Fuels **35**, 9994 (2021), https://doi.org/10.1021/acs.energyfuels.1c00858, URL https://doi.org/10.1021/acs.energyfuels.1c00858.

61. M. F. R. Mulcahy, B. G. Tucker, D. J. Williams, and J. R. Wilmshurst, Australian Journal of Chemistry **20**, 1155 (1967), URL https://doi.org/10.1071/CH9671155.

62. H.-H. Carstensen and A. M. Dean, International Journal of Chemical Kinetics **44**, 75 (2012), https://onlinelibrary.wiley.com/doi/pdf/10.1002/kin.20622, URL https://onlinelibrary.wiley.com/doi/abs/10.1002/kin.20622.

63. C. Bannwarth, S. Ehlert, and S. Grimme, Journal of Chemical Theory and Computation **15**, 1652 (2019).

64. P. Pracht, E. Caldeweyher, S. Ehlert, and S. Grimme (2019).

65. D. E. Edwards, D. Y. Zubarev, W. A. Lester, and M. Frenklach, The Journal of Physical Chemistry A **118**, 8606 (2014).

66. Özlem Yönder and C. Hättig, Fuel **241**, 851 (2019), ISSN 0016-2361.

67. Y. Qiu, W. Zhong, Y. Shao, and A. Yu, Powder Technology **361**, 337 (2020), ISSN 0032-5910.

68. C. Chen and X. Jiang, Phys. Chem. Chem. Phys. **22**, 20829 (2020).

69. F. Castro-Marcano, A. M. Kamat, M. F. Russo, A. C. van Duin, and J. P. Mathews, Combustion and Flame **159**, 1272 (2012).

70. M. Zheng, X. Li, J. Liu, Z. Wang, X. Gong, L. Guo, and W. Song, Energy & Fuels **28**, 522 (2013).