Fast and automated identification of reactions with low barriers: the decomposition of 3-hydroperoxypropanal

Mads Koerstz[⋆], Maria H. Rasmussen[⋆] and Jan H. Jensen[†]

Department of Chemistry, University of Copenhagen, Denmark

⋆ These authors contributed equally to the work
† jhjensen@chem.ku.dk

Abstract

We show how fast semiempirical QM methods can be used to significantly decrease the computational expense for automated reaction mechanism discovery, using two different method for generating reaction products: graph-based systematic enumeration of all possible products and the meta-dynamics approach by Grimme (*J. Chem. Theory. Comput.* 2019, 15, 2847). We test the two approaches on the low-barrier reactions of 3-hydroperoxypropanal, which have been studied by a large variety of reaction discovery approaches and therefore provides a good benchmark. By using PM3 and GFN2-xTB for reaction energy and barrier screening the systematic approach identifies 64 reactions (out of 27,577 possible reactions) for DFT refinement, which in turn identifies the three reactions with lowest barriers plus a previously undiscovered reaction. With optimized hyperparameters meta-dynamics followed by PM3/GFN2-xTB-based screening identifies 15 reactions for DFT refinement, which in turn identifies the three reactions with lowest barrier. The number of DFT refinements can be further reduced to as little as six for both approaches by first verifying the transition states with GFN1-xTB. The main conclusion is that the semiempirical methods are accurate and fast enough to automatically identify promising candidates for DFT refinement for the low barrier reactions of 3-hydroperoxypropanal in about 15-30 minutes using relatively modest computational resources.

---

1   Introduction

Determining the mechanism of chemical reactions is one of the cornerstones of chemistry and computational quantum chemistry (QM) has been proven very helpful in this regard. QM is mainly used to test the viability of possible reaction mechanisms proposed by chemists through heuristic rules. Generating the proposed mechanism and verifying it computationally is generally done manually and requires expert knowledge. However, several very promising methodologies have been developed to automate the discovery of reaction mechanisms using QM. [1–13] Following Dewyer et al. [12] these methods can be categorized into four categories: 1) encoding mechanistic steps, 2) generate transition state (TS) guesses, 3) generate intermediates using graph theory, and 4) generate driving coordinates. Recently, quantum chemical meta-dynamics based approaches [9–11, 13, 14] have emerged as a fifth category. Most of these approaches are rather computationally expensive so they have generally been applied to relatively small molecules with less than ten heavy atoms. As a result, the methods have yet to be widely adopted with the exception of atmospheric and combustion chemistry. [2, 15]

In this paper we investigate whether fast semiempirical QM (SQM) methods can be used to decrease the computational requirements for automated reaction mechanism discovery, using two different methods for generating reaction intermediates: a graph-based method by Kim et al. [3] and the meta-dynamics based method by Grimme. [9] In general we are interested in finding all reactions that proceed at measurable rates at room temperature, which translates into barrier heights (activation free energies) of no more than 25-30 kcal/mol. To test the reaction prediction algorithms we choose a thoroughly studied system: 3-hydroperoxypropanal [16,17]. This system has been studied with multiple TS search algorithms: the freezing string method [6], the single-ended growing string method (SSM) [18], the double-ended growing string method (GSM), [19] the heuristic KinBot algorithm, [15] and the single-component artificial force induced reaction method (SC-AFIR) [20, 21] and it is therefore quite likely that all low-barrier reactions have been found. In this study we aim to find the three lowest-barrier reactions, which have barrier heights of 36, 40, and 46 kcal/mol (Figure 1). These barriers are higher than the target value of 25-30 kcal/mol, which may mean that some of the cutoffs and hyperparameters used in this study may need to be adjusted in future studies.
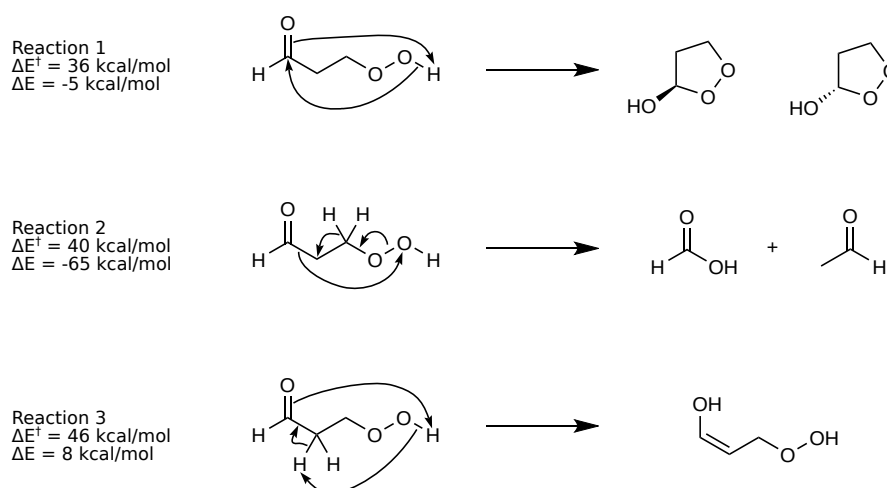
Figure 1: The three lowest barrier one-step reactions of 3-hydroperoxypropanal found by Grambow et al. and Maeda et al [16, 17]. The reactions correspond to R37/38, R69, and R53 in their notation.

2 Computational Methodology

2.1 The test system

The study from Grambow et al. [16] identified 75 one-step unique reactions starting from 3-hydroperoxypropanal (Table S1). However, we were only able to confirm 69 of the reactions by IRC starting from the transition state (TS) geometries supplied by the authors in SI (missing R10, R12, R31, R32, R34, and R62 in Table S1). To compare the product molecules our algorithms produce to those of Grambow et al. we construct a set of atom-mapped SMILES strings with a consistent (arbitrary) atom ordering. It is important to use atom-mapped SMILES to distinguish different reaction mechanisms that lead to the same product, such as reactions R1-R5 in Table S1. In a few cases, such as R15-R17, there are several resonance forms for one of the products and here we select the form with the smallest number of rotatable bonds. This saves as much 3D structural information as possible and avoids rotation around bonds that are not completely free to rotate during the reaction path estimation. Thus, we use the same atom order for the reactant SMILES that serve as input to our two reaction discovery algorithms. Since the atoms are labelled, all tetrahedral centers including CH_2 and CH_3 groups are chiral (Figure S3, we refer to "label-induced" chirality as pseudo-chirality hereafter). When using RDKit to convert the SMILES representation to 3D coordinates the algorithm will generate random pseudo-chiralities for CH_2 and CH_3 groups, which represents a problem for TS search algorithms that interpolate between the 3D structures of reactants and products. Thus, we specify an arbitrary pseudo-chirality for each tetrahedral atom in the reactant (Figure 2) and make sure that these pseudo-chiralities are maintained where applicable in the products as described below. We hereafter, refer to atom mapped SMILES with added pseudo-chirality information simply as "mapped SMILES". The construction of a mapped SMILES string from a usual SMILES string is done automatically as part of the computational methodology.

2.2 Systematic search: Combinatorial Graph Enumeration

In the systematic approach [3], we attempt to systematically sample all possible products that are connected to the reactant with a transition state, as outlined in Figure 3. We start

[C:1](=[O:2])([C@:3]([C@:5]([O:8][O:11][H:12])([H:9])[H:10])([H:6])[H:7])[H:4]
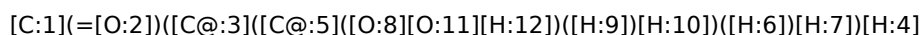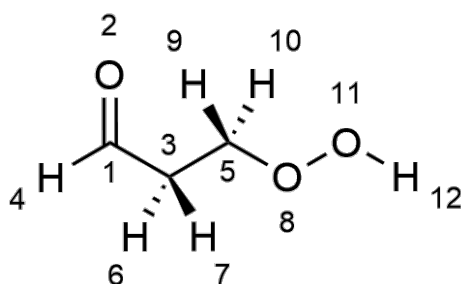
Figure 2: Reactant with assigned atom mapping and defined spatial orientation based on the input reactant SMILES.

by generating the atom connectivity matrix (AC, also known as an adjacency matrix) for the reactant: $\mathbf{R}$. This is a binary $N_{atom} \times N_{atom}$ symmetric matrix where the presence and absence of bonds are indicated by 1's and 0's, respectively. The corresponding AC for product i ($\mathbf{P}_i$) is derived from $\mathbf{R}$ through a conversion matrix $\mathbf{C}_i^{d,f}$

$$\mathbf{P}_i = \mathbf{R} + \mathbf{C}_i^{d,f} \,. \tag{1}$$

Here superscript d and f refers to the number of dissociated and formed bonds, respectively. Each conversion matrix element can only have the values 1, -1, and 0, which corresponds to a bond formation, dissociation, and no bond change.

To systematically create all possible products, we create a set of all conversion matrices. However, to avoid a computational explosion and to make it more likely that the product is connected to the reactant by a single TS, we constrain the maximum number of bond dissociations and formations to three, respectively. In other works [3, 6, 19], the same type of maximal bond formation/dissociation constraints have been used, but are set to two instead. We choose three bond changes because reaction 2 in Figure 1 cannot be obtained using two bond changes. We also constrain the chemical distance (CD) between the reactant and product to 5 bond changes, that is:

$$CD(R, P_i) = \sum_{l > j} |R_{lj} - P_{i,lj}| \leq 5 \,. \tag{2}$$

Thus, the full set of conversion matrices is given by:

$$\{\mathbf{C}\} = \{\mathbf{C}^{f,d} : (f, d) \leq 3 \text{ and } f + d \leq 5\} \,. \tag{3}$$

With the full set of conversion matrices, we iteratively create the product AC matrices ($\{\mathbf{P}_i\}$). Once an AC matrix is constructed, we check that the number of neighbours for each atom is less than or equal to its maximally allowed valence (C $\leq$ 4, H $\leq$ 1, O $\leq$ 3). If just one atom in the product has more neighbours than allowed, it is discarded. However, if the product AC matrix is valid, we use the program xyz2mol [22, 23] to convert the product AC matrix $\mathbf{P}_i$ into an RDKit mapped molecular graph. To avoid duplicate products, we only save the products graphs with unique SMILES. Before we compare the SMILES, we ensure that we use the resonance form with least rotatable bonds.

The product graphs contain no information about stereochemistry including "label-induced" stereochemistry as discussed above (Figure S3). The stereo information is assigned to each product graph by enumerating all possible stereoisomers using RDKit. For each stereoisomer,
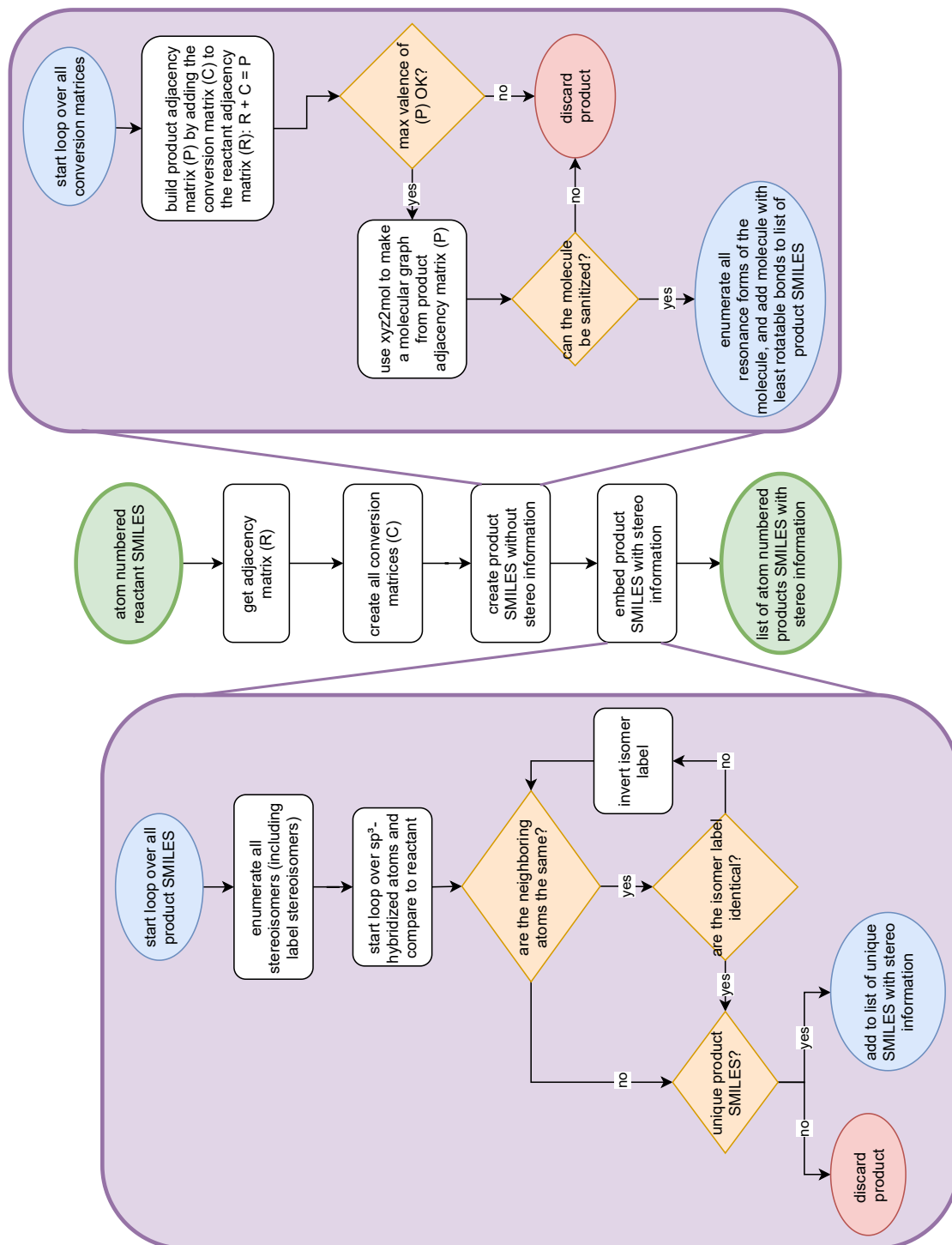
Figure 3: Workflow for the systematic search for hypothetical products
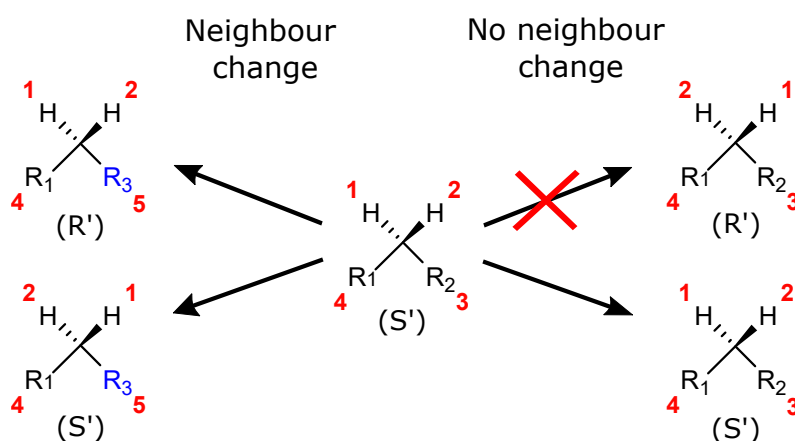
Figure 4: Example of which chiral tags are allowed if the neighbouring atoms change or do not change.

we make sure that the chirality of all non-reacting atoms are preserved during the reaction. The non-reacting atoms are identified by comparing all the reactants sp^3 hybridized atoms to the corresponding atom in the product and checking whether the immediate neighbours are identical. If the chirality of any non-reacting atom is different from the reactant, we discard the stereoisomer as shown in Figure 4 (for a concrete example see Figure S4).

2.3 Meta-dynamics search

In the meta-dynamics search a molecular dynamics (MD) simulation is performed using a GFN2-xTB energy function that allows for bond formation and dissociation. [9] The energy function is augmented by a biasing potential that rewards changes in the molecular structure. Our main question is whether this approach tends to produce molecules that are connected by low barriers. In addition to the usual MD parameters (time-step, length of simulation, temperature etc.), parameters describing the additional biasing potential must be specified. The biasing potential has the form [9]

$$E_{bias}^{RMSD} = \sum_{i}^{n} k_{push} \cdot e^{-\alpha \Delta_i^2}, \tag{4}$$

where k_{push} controls the strength of the biasing potential, defining how strongly the system should be 'pushed' out of previously visited structures. Choosing too small a value, we approach the unbiased MD limit where no reactions take place (or at least unreasonable long simulations are needed for them to occur). On the other hand, choosing too high a value can lead to chemically irrelevant reactions. n is the number of structures saved (which the system is pushed out of). Here, a new structure is added to the biasing potential every picosecond. Δ_i is the root mean squared deviation (RMSD) between the current structure and structure i in Cartesian space. α describes the width of the biasing potential and thus determines how far from the saved structures the system is still affected by the biasing potential. A high value of α prescribes a narrow biasing potential meaning that the system can 'escape' the biasing potential by smaller changes to the structure (e.g. conformational changes). A low value of α means that very large changes in the structure are needed to avoid being affected by the biasing potential.

In addition to the energy potential and the RMSD potential, a spherical wall potential is

used to confine the reactant system using a logfermi pontential.

$$E_{bias}^{wall} = k_{wall} \sum_i^{N_{atoms}} \log\left(1 + e^{-\beta(R-r_i)}\right).$$

(5)

k_{wall} and β determine the strength and steepness of the potential, respectively and we use the values suggested by Grimme [9] ($k_{wall} = 0.019E_h$ and $\beta = 10$ Bohr^{-1}). r_i is the position vector of atom i and R is the radius of the sphere, which is related to the input coordinates by a scaling-factor s. The simulations are run for 50 ps with 0.4 fs time steps, a thermostat temperature of 300 K, double hydrogen weight and with no bonds constrained (SHAKE off) unless otherwise noted. A maximum of 100 structures (n) are used for the RMSD criteria.

In order to find the best set of values for k_{push}, α, and s, i.e. the values for which the MD simulations find the three target products in Figure 1, different parameter combinations are tested using the work flow presented in Figure 5. The mapped reactant SMILES is embedded using `RDKit` [24] followed by an MMFF94 force field optimization, also done using the `RDKit`, after which the structure is written to file. The structure is then relaxed for 2 ps without the RMSD biasing potential (but with the wall biasing potential) and the end structure in the trajectory is used in the 100 50-ps meta-dynamics runs.

Each of the 100 trajectories are then analyzed in the following way:

- Every 250th structure on the trajectory (corresponding to every 100 fs) is converted to an atom-mapped SMILES using `xyz2mol` [22]. In cases where several resonance structures are possible, we select the one with the fewest rotatable bonds.

- Pseudo-chirality information is added to the -CH$_2$ and -CH$_3$ groups to obtain mapped SMILES as described above. In practice, this is done by first finding CH$_2$ and CH$_3$ groups by substructure matching. One of the hydrogens in each is then temporarily exchanged with a fluorine atom which allows `RDKit` to recognize the chirality (Figure S5).

- The mapped SMILES string is then saved (along with the Cartesian coordinates) if it is different from the last saved entry.

- All saved structures are then geometry optimized with GFN2-xTB [25,26] and the SMILES extracted and analyzed as described above. The geometry optimisation is important since it occasionally results in a change in bonding.

- The final list of SMILES strings based on optimized structures correspond to a (possible) multi-step reaction, where each step is a potential elementary reaction. These reactions are collected and counted across all 100 trajectories with the median number of time-steps it took for the reaction to happen.

- In this study, we are only interested in identifying possible single-step reactions starting from the input reactant (Figure 2). Therefore, the products of interest in the list are those that immediately follow the reactant. This is usually the second SMILES string in the list (the first being the reactant) but occasionally the reactant is regenerated during the MD so it is necessary to check the entire list.

2.4 Identification of reactions with low barriers

Once the set of potential products have been determined, by either meta-dynamics simulations or the systematic search, the low-barrier reactions are identified using the the procedure

Parameters:
Size of wall potential (scaling factor: s)
Strength of the biasing potential (k_{push})
Width of the biasing potential (alpha)
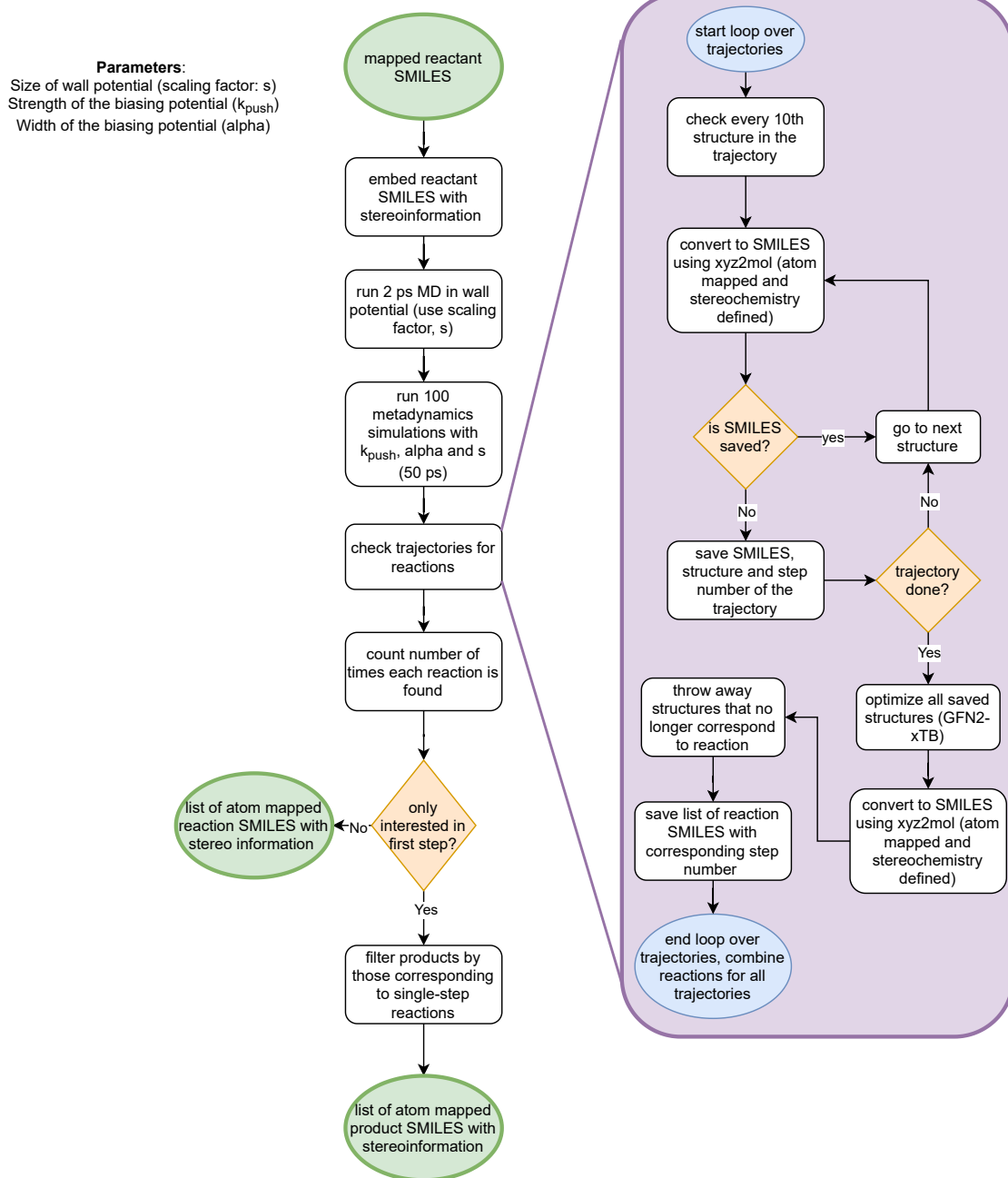
Figure 5: workflow for the automated procedure used to collect reactions based on the meta-dynamics approach

list of atom numbered
product SMILES with
stereoinformation

extract unique
fragments from
products. Start loop
over fragments

calculate reaction
energies for each
reaction

embed fragment
using RDKit

eliminate reactions
with reaction energy >
30 kcal/mol

do systematic
conformer search for
each fragment

get barrier estimate
with RMSD-PP
(GFN2-xTB) for
remaining reactions

do geometry
optimization and
frequency calculation
to confirm minima
(PM3)

eliminate reactions
with barrier > 50
kcal/mol

extract energy of
most stable
conformer: E_i^{lowest}

calculate reaction
energies with DFT
(same procedure as
with PM3)

calculate reaction energy

$$\sum_i^{N_{frags}} E_i^{lowest} - E_{reactant}$$

optimize TS with DFT
using RMSD-PP
guess structure

test TS with IRC.
Eliminate if incorrect

list of remaining
reactions described
with DFT barriers and
reaction energies

Figure 6: Flowchart describing how the potential products are screened

outlined in Figure 6.

The first step is to compute reaction energies and eliminate reactions with reaction energies larger than a threshold (here 30 kcal/mol) since the reaction energy provides a lower bound on the barrier height. The products for several of the reactions consist of more than one molecule (or fragment) and many of the molecules are identical since atom mapping does not affect the reaction energy. Thus, to reduce the number of calculations, we identify all unique fragments and perform PM3 calculations on the fragments and calculate the reaction energy as:

$$\Delta E = \sum_{i}^{N_{frags}} E_i^{\text{lowest}} - E_{reactant} \, . \tag{6}$$

We also test other SQM methods, including GFN1-xTB, GFN2-xTB, and PM6, but PM3 performed best when compared to DFT results (Figure S2). When we compute the energies for each fragment, we use the energy for the most stable conformer. The fragments' most stable conformer is found by systematically sampling conformers by rotating all rotatable bonds by ±120 degrees and then performing a PM3 structure optimization in `MOPAC`. Then we pick the most stable conformer based on the enthalpy. We then remove all reactions with reaction energies above 30 kcal/mol. This value is chosen because we normally would define low-barrier reactions as being below 30 kcal/mol and we want to investigate how many structures this cutoff eliminates. As we show below, this cutoff does not eliminate the reactions of interest for this particular molecule.

We are interested in low-barrier reactions, which we in this study define as barriers below 50 kcal/mol (Figure 1). To screen the remaining reactions on their barriers, we use barrier estimates based on the RMSD-PP approach using GFN2-xTB [9] as described in [27]. As barrier estimates were seen to vary between runs, we calculate a barrier estimate for each reaction five times and keep the lowest one. In case all of the barrier estimate computations fail, we discard the reaction. We then discard all reactions with barriers above 50 kcal/mol.

Finally, the remaining reactions are checked with DFT (B3LYP/6-31+G(d) [28–32]) as in the original studies [16, 17]. The reaction energies are computed as described above and the barriers and transition states (TSs) are computed and verified by IRCs [27]. The TS is found for all of the cases where the barrier estimate is below 50 kcal/mol. A linear interpolation (10 points from maximum energy structure to both neighbours in the RMSD-PP path) is performed and the interpolated structures are subjected to single point energy calculations using DFT. The maximum energy structure based on DFT calculations is used as initial guess for the TS structure in a DFT TS search using `Gaussian` [opt=(calcall, ts, noeigen)]. The procedure is the same for GFNx-xTB verification of the TSs using a custom script to interface `xtb` with `Gaussian`.

The PM3 calculations are performed with `MOPAC2016` and the DFT calculations are performed with `Gaussian16`. [33] The meta-dynamics calculations and the RMSD-PP barrier estimates are performed with version 6.1.4 of the `xtb` program. All usage of `RDKit` is performed with version 2020.09.05.

3 Results and Discussion

3.1 Systematic search: Combinatorial Graph Enumeration

The main hyperparameter for the combinatorial graph enumeration is the maximum number of bonds broken or formed. Other works [3, 6, 19] have suggested setting this number to two. However, by only allowing two bonds to be formed or broken, we do not find Reaction 2 (Figure 1), which not only has the second-lowest barrier but is also the thermodynamic product. If we instead allow all the molecules to form and break three bonds but constrain the total number of changes to five, we find 74 of the 75 products that Grambow et al. found. The missing reaction (R62 in Table S1) has more than five bonds broken and formed and the barrier is 80 kcal/mol. This approach results in a total of 27,577 unique reactions involving 4,095 unique fragments, for which we perform a systematic conformational search. For many of the reactions, none of the conformers successfully converged with PM3 while for other reactions, the atom connectivity changed during the optimizations and were therefore discarded. Thus, only 1,138 of the fragments calculations resulted in a fragment energy allowing us to compute reaction energies for 9,056 of the 27,577 reactions. As seen in the distribution of reaction energies in Figure 7 most of the reactions have high reaction energies.
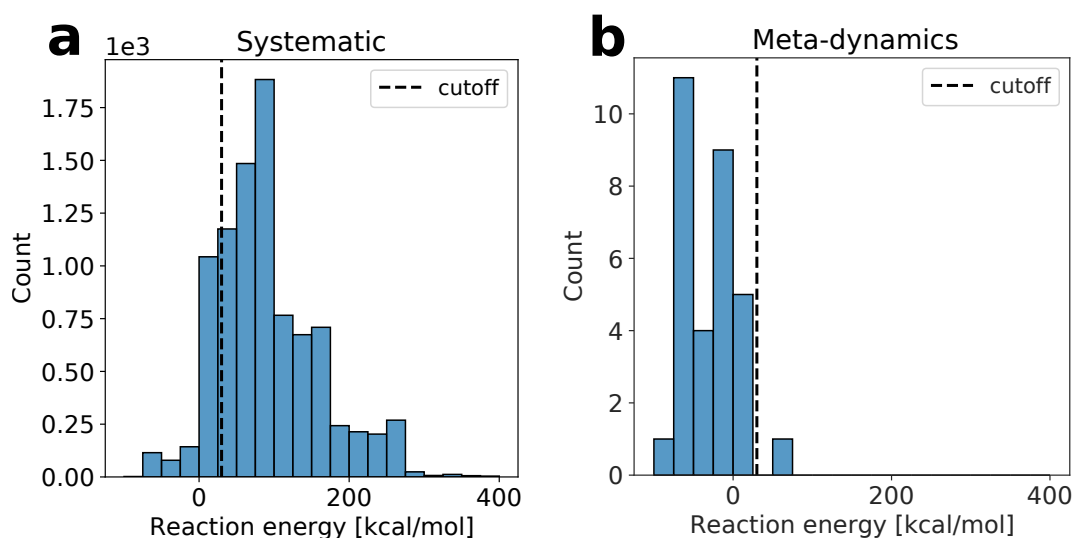
Figure 7: Distribution of the reaction energies for (a) the 9,056 reactions in the systematic search and (b) the 31 reactions from the meta-dynamics search. The chosen cutoff value of 30 kcal/mol is indicated. For the systematic search 1,583 reactions are below this cutoff. For the meta-dynamics search, 30 reactions are below this cutoff (only one discarded)

Of the 9,056 reactions only 1,583 have reaction energies below 30 kcal/mol and the barriers of these reaction were estimated using our RMSD-PP-based method [27] (Figure 8). Of these 1,583 reactions 64 of the reactions have estimated barriers below 50 kcal/mol and are thus checked with DFT. For seven of the 64 reactions TSs connecting the intended reactant and product are found using DFT, as verified by IRCs. Six of them correspond to the reactions depicted in Figure 1: two of them correspond to Reaction 1 (the two enantiomers of the product), two of them correspond to Reaction 2 (there are two ways of moving an H from the -CH$_2$ group, with TSs that are mirror images) and two of them correspond to Reaction 3 (either of the hydrogens on the reacting -CH$_2$ group can move resulting in TSs that are mirror
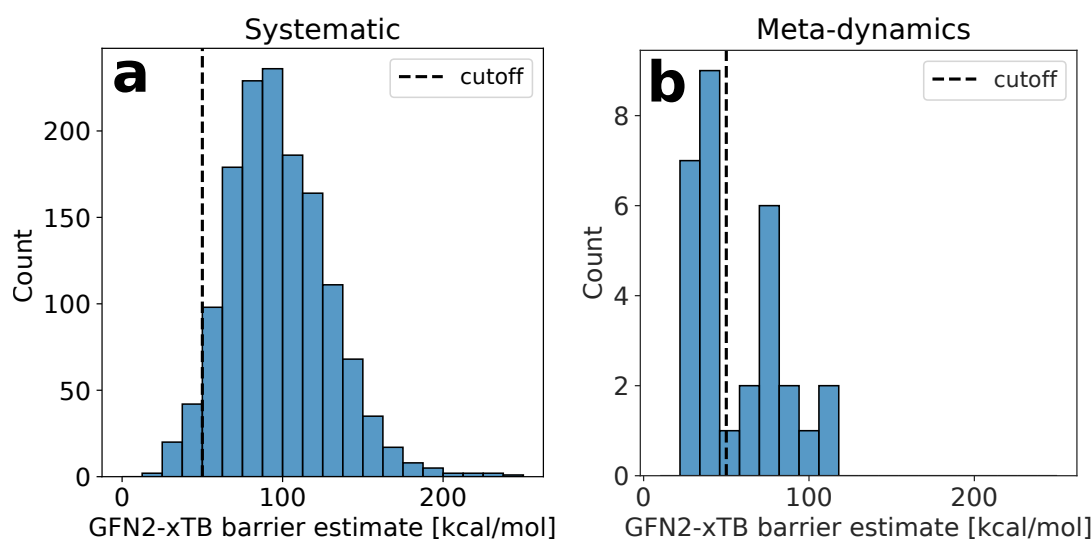
Figure 8: Distributions of estimated barriers for (a) the systematic search and (b) the search based on meta-dynamics. A total of 1,583 reactions were tested in the systematic search. For 149 of them the search for a barrier estimate failed. 1416 of the reactions had barriers above the cutoff at 50 kcal/mol, while 64 had barrier estimates below the cutoff. A total of 30 reactions were tested in the search based on meta-dynamics. For 1 of them the search for a barrier estimate failed. 15 of the reactions had barriers below the cutoff.
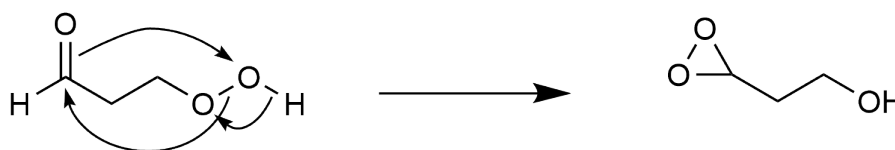
Figure 9: New reaction found with a barrier of 62 kcal/mol and a reaction energy of 19 kcal/mol during the systematic search

images). The remaining reaction, depicted in Figure 9, has a barrier of 62 kcal/mol and a reaction energy of 18.8 kcal/mol and was not found in any of the previous studies [16, 17]. In addition, R28 showed up during one of the searches for another reactant and product pair and is discussed further below.

If instead the 64 reactions with estimated barriers less than 50 kcal/mol are first checked (i.e. TS optimisation followed by IRC) with GFN2-xTB the number of DFT searches can be lowered significantly. We tested three different criteria for further DFT refinement: 1) The TS connects 3-hydroperoxypropanal to the correct product, 2) The TS connects to 3-hydroperoxypropanal, and 3) the TS connects to either 3-hydroperoxypropanal or the correct product. Here the correct product is that used to create the TS guess using the RMSD-PP method. Using criterion 1, 2, and 3 reduces the number of DFT refinements to 5, 19, and 30, respectively. Using criterion 1, Reaction 2 is not identified as the corresponding TS is shown to connect 3-hydroperoxypropanal to formaldehyde and oxiranol (N1a, Figure S6) whereas using criterion 2 or 3 reactions 1-3 and the new reaction (Figure 9) are all identified by the DFT refinement. If GFN1-xTB is used instead, criterion 1, 2, and 3 reduces the number of DFT refinements to 6, 17, and 33, respectively and the subsequent DFT refinements identify Reactions 1-3, but not the new reaction.

We note that Grambow et al. [16] reports very different barriers for formation of the R and S product for Reaction 1 (34-39 and 67 kcal/mol, respectively), while we find the same low barrier ranges for both reactions. Note that we in general compute slightly higher barriers as our barriers are calculated compared to the lowest conformer of the reactant instead of the conformer that the IRC ends up in.

3.2 Meta-dynamics search

The main hypothesis we wish to test regarding the meta-dynamics approach is whether it can be used to identify reactions with low barriers. As outlined above the meta-dynamics approach relies on three main hyperparameters, k_{push}, α, and s, that modulate the effect that structural changes have on the course of the trajectory. Increasing k_{push} and decreasing α increases the effect of structural or conformational changes on the trajectory as measured by the RMSD from previous structures. Similarly, increasing s increases the radius of the spherical boundary potential surrounding the molecule(s) and is most important when the number of molecular fragments changes.

We performed 100 50-ps meta-dynamics simulations for each of 27 different combinations of k_{push}, α, and s and identify the product of the first reaction (if any) that occurs in each trajectory. Table 1 shows the results for a subset of the 27 combinations, while the results for all 27 combinations can be found in SI (Table S2). For k_{push}=0.05, α=0.3, and s=0.8, all 100 trajectories resulted in a reaction, and in one case two reactions occurred leading to a total of 101 products (N_{total}). Ten of the trajectories lead to Reaction 1, while eight and eight of the trajectories lead to Reactions 2 and 3, respectively. N_{99} is the minimum number of trajectories needed to find at least one example of each reaction with 99% certainty and is the smallest N for which

$$0.99 \geq \left(1 - (1 - p_1)^N\right) \cdot \left(1 - (1 - p_2)^N\right) \cdot \left(1 - (1 - p_3)^N\right), \tag{7}$$

where p_1 is the probability of finding Reaction 1 (e.g. 10/100 = 0.10 for k_{push}=0.05, α=0.3, and s=0.8, cf. Table 1). We select k_{push}=0.05, α=0.3, and s=0.8 as the best combination of k_{push}, α, and s since that results in the lowest value of N_{99} for this molecule. The results of these 100 trajectories are thus analysed further.

Table 1: Table of how many times each of the three lowest barrier reactions were found for each parameter set. Each parameter set was tested in 100 simulations (50ps) and N_{total} states how many times a reaction with the required reactant occurred. N_{99} is the estimated number of times the meta-dynamics should be run for each parameter set in order to be 99% sure of finding all 3 reactions.

s	k_{push}	α	reaction 1	reaction 2	reaction 3	N_{total}	N_{99}
0.8	0.03	0.7	1	0	0	3	-
0.8	0.03	0.5	2	3	0	25	-
0.8	0.03	0.3	4	7	8	88	114
0.8	0.05	0.7	5	6	5	48	107
0.8	0.05	0.5	12	4	8	101	114
0.8	0.05	0.3	10	8	8	103	65

Overall, the 100 trajectories resulted in 31 different products, 30 of which are below 30 kcal/mol (Figure 8b). Fifteen of these 30 reactions have estimated reaction barriers lower than 50 kcal/mol and these were subjected to DFT refinement. DFT TSs are located for all
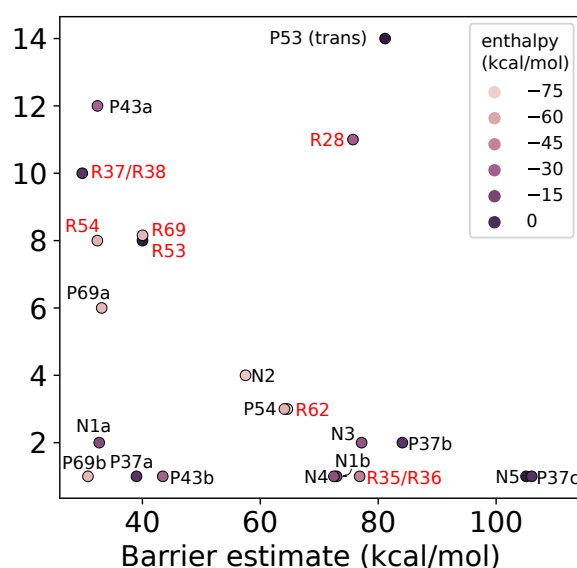
Figure 10: Number of times each reaction found in the meta-dynamics search was observed versus the GFN2-xTB barrier estimate. R69 is actually exactly on top of R53, but was moved slightly up to reveal both points. The points are colored according to the PM3 reaction enthalpies. The structures corresponding to each point is shown in Figures S7 and S8 for barrier estimate below and above 50 kcal/mol, respectively.

15 reactions, but only six connect to 3-hydroperoxypropanal and these six correspond to the three reactions with lowest barrier (Figure 1). Using GFN2-xTB criterion 1, 2, and 3 reduces the number of DFT refinements to 4, 12, and 14, respectively. Again, using criterion 1, Reaction 2 is not identified whereas using criterion 2 or 3 Reactions 1-3 are all identified by the subsequent DFT refinement. If GFN1-xTB is used instead, criterion 1, 2, and 3 reduces the number of DFT refinements to 6, 11, and 14, respectively and the subsequent DFT refinements identify Reactions 1-3.

Our results suggest that hyperparameter settings can be found for which reactions with low barrier heights are more likely to occur. To test this further we plot the estimated barrier heights vs the number of occurrences among the 100 trajectories shown in Figure 10. There are two main categories of products: those found by Grambow et al. [16] and Maeda et al. [17] (labels starting with "R") and other structures that are similar to those structures but differ in atom labels (labels starting with "P") or completely novel structures (labels starting with "N"). As we will demonstrate only the R points correspond to elementary reactions for which the barrier estimate reflects the lowest barrier from the reactant. Thus, these are the only points for which we expect an inverse correlation between barrier estimate and number of occurrences in the trajectories. There are seven such points in the plot (R37, R54, R69, R53, R62, R35, and R28) and with the exception of R28 there is a good inverse correlation meaning that products of reactions with low barriers are observed more often. The most likely reason that R28 is observed frequently despite having a relatively large barrier estimate is that the product consists of three separate fragments (ethenone, formaldehyde, and water), which will be favored due to the large change in RMSD relative to the reactant. While there are three other products consisting of three fragments (R29, R30, and R31) only R28 has a relatively low barrier (50 vs 70-88 kcal/mol at the DFT level) as well as a much lower reaction energy (-25 vs 25-45 kcal/mol at the DFT level).

The most common occurrence is "P53 (trans)", the trans-isomer of R53. Inspection of a
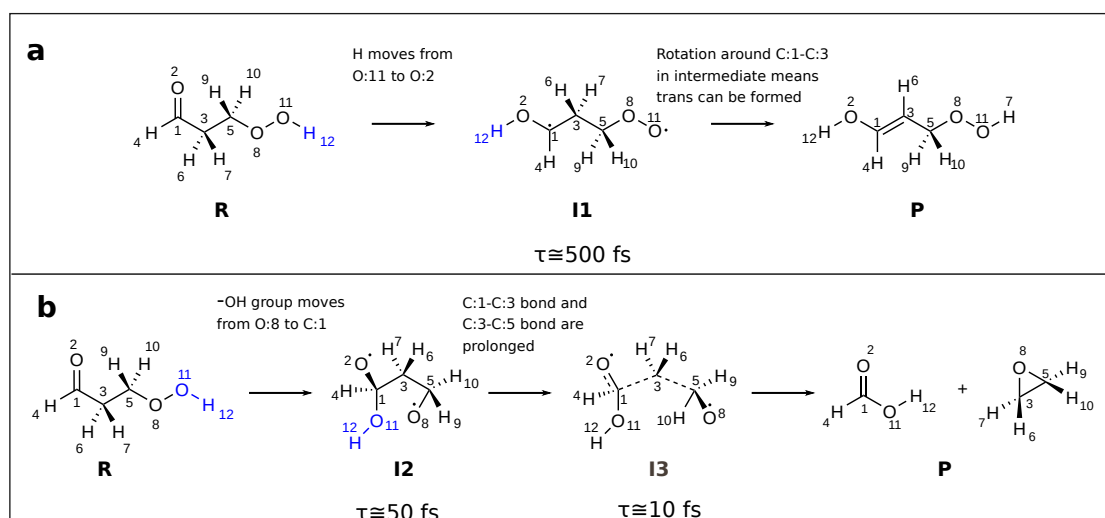
Figure 11: Examples of how the meta-dynamics trajectories look for (a) R53trans and (b) P43a. The unstable intermediates are labelled with approximate lifetimes (τ)

representative trajectory shows that an unstable intermediate (I1 in Figure 11a) is formed by transferring the -OOH hydrogen to the carbonyl group), which allows for free rotation about the C_1-C_3 bond and thus, ultimately, a roughly equal probability of the cis and trans isomer. The second most common occurrence is P43a, which is the same as R43 (formic acid + oxirane) but with a different pseudo-chirality at the C_3 group (Figure 11b), indicating a different mechanism of formation from the reactant. Inspection of a representative trajectory shows that an unstable intermediate (I2 in Figure 11b) is formed by transferring the -OOH hydroxyl group to the carbonyl carbon. I2 is then converted to a very short-lived intermediate where both CC bonds are significantly elongated, so that the pseudo-chirality of C_3 can change easily before going to P43a.

3.3 Timings

To obtain optimal timings we terminate the meta-dynamics simulations when a reaction has occurred (checking every 5 ps). The 94,820 PM3 geometry optimisations require only 6 minutes on Intel Xeon E5-2643 v3 (3.4 GHz) nodes with a total of 100 cores, compared to 10 minutes for the 100 meta-dynamics simulations.

The 1,583 and 30 reaction barrier estimates for the systematic and meta-dynamics approach, respectively, take about 29 minutes and 2 minutes, respectively, on the same cluster. The final DFT refinement step involves 64 and 15 molecules for the systematic and meta-dynamics searches, respectively. However, this can be reduced to six DFT refinements by pre-screening using GFN1-xTB as described above, although in that case the new reaction is not discovered by the systematic approach. In conclusion, our best estimate is thus that the three lowest barrier-reactions can be identified for DFT refinement in about 12 minutes using the meta-dynamics approach and in about 35 minutes using the systematic approach on 100 cores.

4 Conclusions and outlook

We show how fast semiempirical QM methods can be used to decrease the computational requirements for automated reaction mechanism discovery, using two different methods for generating reaction intermediates: graph-based systematic enumeration of all possible products developed by Kim et al. [3] and the meta-dynamics based method by Grimme. [9] We test the two approaches on the reactions of 3-hydroperoxypropanal, which have been studied by a large variety of reaction discovery approaches [16, 17] and therefore provides a good benchmark. Our aim is to find the three reactions with barriers below 50 kcal/mol (Figure 1).

The systematic approach generates 27,577 reactions, involving 4,095 fragments, that can be made by forming and breaking three bonds but constrain the total number of changes to five. Only 1,138 of the 4,095 fragments are stable upon PM3 geometry optimisation, resulting in 9,056 reactions for which the reaction energy can be computed. GFN2-xTB barriers are estimated for the 1,583 reactions that have reaction energies below 30 kcal/mol and the 64 reaction with estimated barriers less than 50 kcal/mol are refined at the DFT level and verified with an IRC. This results in four reactions connecting 3-hydroperoxypropanal to products: the three reactions with barriers below 50 kcal/mol previously identified by Grambow et al. [16] and a previously undiscovered reaction (Figure 9) with a barrier of 62 kcal/mol.

In the meta-dynamics approach the energy is augmented by two terms: a term favoring changes in the molecular structure as measured by the change in RMSD from earlier structures in the trajectory and a confining potential. The effect of each term on the dynamics simulations are controlled by several hyperparameters. After finding an optimum set of hyperparameters for reaction discovery, 100 50-ps meta-dynamics simulations identify 31 different reactions, of which all but one have PM3 reaction energies below 30 kcal/mol. Fifteen of the 30 reactions have estimated barrier heights below 50 kcal/mol and subsequent DFT refinement of these yield the three target reactions with barriers below 50 kcal/mol. Seven products are observed that can be made from 3-hydroperoxypropanal in one elementary step, and for these reactions there is a good inverse correlation between the barrier and the frequency with which they are observed. The one exception a reaction (R28) where the product consists of three separate molecules, which therefore results in a disproportionate change in RMSD compared to products consisting of one or two molecules thereby favoring the reaction. The remaining products are most likely made via unstable intermediates that are not minima on the potential energy surface.

The determination of products by the systematic and meta-dynamics searches require 15-30 minutes using modest computational resources (100 CPU cores). One difference in efficiency is the greater number of reactions that are refined by DFT in the systematic search. However, subsequent calculations show that most of these refinements can also be done at the semiempirical level, which would make the efficiency of the methods relatively similar for this molecule.

The main conclusions are thus that the semiempirical methods are accurate and fast enough to automatically identify promising candidates for DFT refinement for the low barrier reactions of 3-hydroperoxypropanal within 15-30 minutes using modest computational resources. The main question is how the systematic and meta-dynamics search algorithms scale with system size and whether the optimum hyperparameters found for this molecule also perform well for other types of reactions. We are currently investigating both questions and preliminary results suggest that the answer to the latter question is yes.

Acknowledgments

This work was supported by a research grant (00022896) from VILLUM FONDEN.

References

[1] S. Maeda, K. Ohno and K. Morokuma, *Systematic exploration of the mechanism of chemical reactions: the global reaction route mapping (GRRM) strategy using the ADDF and AFIR methods*, Phys. Chem. Chem. Phys. **15**, 3683 (2013), doi:10.1039/C3CP44063J.

[2] P. L. Bhoorasingh, B. L. Slakman, F. Seyedzadeh Khanshan, J. Y. Cain and R. H. West, *Automated transition state theory calculations for high-throughput kinetics*, J. Phys. Chem. A **121**, 6896 (2017), doi:10.1021/acs.jpca.7b07361.

[3] Y. Kim, J. W. Kim, Z. Kim and W. Y. Kim, *Efficient prediction of reaction paths through molecular graph and reaction network analysis*, Chem. Sci. **9**, 825 (2018), doi:10.1039/C7SC03628K.

[4] J. P. Unsleber and M. Reiher, *The exploration of chemical reaction networks*, Annu. Rev. Phys. Chem. **71**, 121 (2020), doi:10.1146/annurev-physchem-071119-040123.

[5] C. Robertson, I. Ismail and S. Habershon, *Traversing dense networks of elementary chemical reactions to predict minimum-energy reaction mechanisms*, Chem. Syst. Chem. e1900047 (2020), doi:10.1002/syst.201900047.

[6] Y. V. Suleimanov and W. H. Green, *Automated discovery of elementary chemical reaction steps using freezing string and Berny optimization methods*, J. Chem. Theory Comput. **11**, 4248 (2015), doi:10.1021/acs.jctc.5b00407.

[7] S. Habershon, *Automated prediction of catalytic mechanism and rate law using graph-based reaction path sampling*, J. Chem. Theory Comput. **12**, 1786 (2016), doi:10.1021/acs.jctc.6b00005.

[8] D. Rappoport, C. J. Galvin, D. Y. Zubarev and A. Aspuru-Guzik, *Complex chemical reaction networks from heuristics-aided quantum chemistry*, J. Chem. Theory Comput. **10**, 897 (2014), doi:10.1021/ct401004r.

[9] S. Grimme, *Exploration of chemical compound, conformer, and reaction space with meta-dynamics simulations based on tight-binding quantum chemical calculations*, J. Chem. Theory Comput. **15**, 2847 (2019), doi:10.1021/acs.jctc.9b00143.

[10] R. J. Shannon, E. M. Nunez, D. V. Shalashilin and D. R. Glowacki, *Chemdyme: Kinetically steered, automated mechanism generation through combined molecular dynamics and master equation calculations*, (2021), arXiv:2104.02389.

[11] C. Lavigne, G. dos Passos Gomes, R. Pollice and A. Aspuru-Guzik, *Automatic discovery of chemical reactions using imposed activation* (2020), doi:10.26434/chemrxiv.13008500.v1.

[12] A. L. Dewyer, A. J. Argüelles and P. M. Zimmerman, *Methods for exploring reaction space in molecular systems*, WIREs Comput. Mol. Sci. **8**, e1354 (2018), doi:10.1002/wcms.1354.

[13] D. Polino and M. Parrinello, *Kinetics of aqueous media reactions via ab initio enhanced molecular dynamics: The case of urea decomposition*, J. Phys. Chem. B **123**, 6851 (2019), doi:10.1021/acs.jpcb.9b05271.

[14] R. J. Shannon, S. Amabilino, M. O'Connor, D. V. Shalishilin and D. R. Glowacki, *Adaptively accelerating reactive molecular dynamics using boxed molecular dynamics in energy space*, J. Chem. Theory Comput. **14**, 4541 (2018), doi:10.1021/acs.jctc.8b00515.

[15] R. V. de Vijver and J. Zádor, *KinBot: Automated stationary point search on potential energy surfaces*, Comp. Phys. Commun. **248**, 106947 (2020), doi:10.1016/j.cpc.2019.106947.

[16] C. A. Grambow, A. Jamal, Y.-P. Li, W. H. Green, J. Zádor and Y. V. Suleimanov, *Unimolecular reaction pathways of a γ-ketohydroperoxide from combined application of automated reaction discovery methods*, J. Am. Chem. Soc. **140**, 1035 (2018), doi:10.1021/jacs.7b11009.

[17] S. Maeda and Y. Harabuchi, *On benchmarking of automated methods for performing exhaustive reaction path search*, J. Chem. Theory Comput. **15**, 2111 (2019), doi:10.1021/acs.jctc.8b01182.

[18] P. M. Zimmerman, *Single-ended transition state finding with the growing string method*, J. Comput. Chem. **36**, 601 (2015), doi:10.1002/jcc.23833.

[19] P. M. Zimmerman, *Automated discovery of chemically reasonable elementary reaction steps*, J. Comput. Chem. **34**, 1385 (2013), doi:10.1002/jcc.23271.

[20] S. Maeda, Y. Harabuchi, M. Takagi, T. Taketsugu and K. Morokuma, *Artificial force induced reaction (AFIR) method for exploring quantum chemical potential energy surfaces*, Chem. Rec. **16**, 2232 (2016), doi:10.1002/tcr.201600043.

[21] S. Maeda, T. Taketsugu and K. Morokuma, *Exploring transition state structures for intramolecular pathways by the artificial force induced reaction method*, J. Comput. Chem. **35**, 166 (2013), doi:10.1002/jcc.23481.

[22] J. H. Jensen, *xyz2mol* (2020), https://github.com/jensengroup/xyz2mol.

[23] Y. Kim and W. Y. Kim, *Universal structure conversion method for organic molecules: From atomic connectivity to three-dimensional geometry*, Bull. Korean Chem. Soc. **36**, 1769 (2015), doi:10.1002/bkcs.10334.

[24] G. Landrum, *Rdkit: Open-source cheminformatics* (2020).

[25] C. Bannwarth, S. Ehlert and S. Grimme, *GFN2-xTB-An accurate and broadly parametrized self-consistent tight-binding quantum chemical method with multipole electrostatics and density-dependent dispersion contributions*, J. Chem. Theory Comput. **15**, 1652 (2019), doi:10.1021/acs.jctc.8b01176.

[26] S. Grimme, C. Bannwarth and P. Shushkov, *A robust and accurate tight-binding quantum chemical method for structures, vibrational frequencies, and noncovalent interactions of large molecular systems parametrized for all spd-block elements (z = 1-86)*, J. Chem. Theory Comput. **13**, 1989 (2017), doi:10.1021/acs.jctc.7b00118.

[27] M. H. Rasmussen and J. H. Jensen, *Fast and automatic estimation of transition state structures using tight binding quantum chemical calculations*, PeerJ Phys. Chem. **2**, e15 (2020), doi:10.7717/peerj-pchem.15.

[28] A. D. Becke, *Density-functional exchange-energy approximation with correct asymptotic behavior*, Phys. Rev. A Gen. Phys. **38**, 3098 (1988), doi:10.1103/PhysRevA.38.3098.

[29] C. Lee, W. Yang and R. G. Parr, *Development of the Colle-Salvetti correlation-energy formula into a functional of the electron density*, Phys. Rev. B Condens. Matter **37**, 785 (1988), doi:10.1103/PhysRevB.37.785.

[30] A. D. Becke, *Density-functional thermochemistry. III. The role of exact exchange*, J. Chem. Phys. **98**, 5648 (1993), doi:10.1063/1.464913.

[31] R. Ditchfield, W. J. Hehre and J. A. Pople, *Self-consistent molecular-orbital methods. IX. An extended Gaussian-yype basis for molecular-orbital studies of organic molecules*, J. Chem. Phys. **54**, 724 (1971), doi:10.1063/1.1674902.

[32] W. J. Hehre, R. Ditchfield and J. A. Pople, *Self-consistent molecular orbital methods. XII. Further extensions of Gaussian-type basis sets for use in molecular orbital studies of organic molecules*, J. Chem. Phys. **56**, 2257 (1972), doi:10.1063/1.1677527.

[33] M. J. Frisch et al., *Gaussian~16 Revision A.03*, Gaussian Inc. Wallingford CT (2016).

Supporting Information

The code and data resulting from this study can be found here:

- code: https://github.com/jensengroup/ReactionDiscovery

- data: https://sid.erda.dk/wsgi-bin/ls.py?share_id=DyEK4eZmrD

Benchmark of SQM methods

To determine which SQM method that gives us the most reliable reaction energies, we compare the reaction energies obtained with PM3, PM6, GFN1-xTB, and GFN2-xTB to DFT B3LYP/6-31+G(d). From the Grambow et al. article [16] we have 69 elementary step reactions for the peroxide we can use as our benchmark set.

To compare the reaction energies, we split the products from the Grambow dataset into a list of unique fragments as determined by their SMILES. We then compute the reaction energy as in Eq. 6.
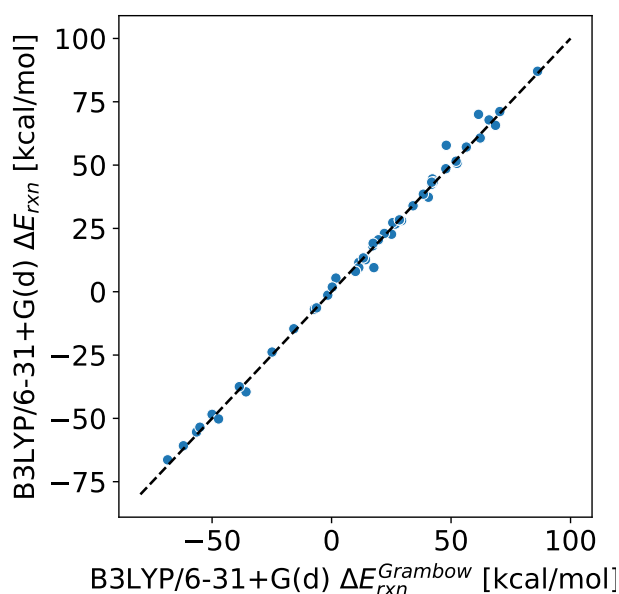
Figure S1: Comparison between Grambows DFT reaction energies, and the reaction energies computed by summation of fragment energies. Reactions 10, 12 , 31, 32, 34, 64 are removed from the benchmark since we can not confirm the reactions by IRC.

We choose the lowest energy conformer for each of the five methods (4 SQM and 1 DFT) by computing the free energy for the systematically sampled conformers. The conformers are sampled by systematically rotating each rotatable bond by ± 120 degrees.

In Figure S2 are the four SQM reaction energies for the Grambow dataset plotted against the corresponding DFT reaction energies. When we compare the MAE for the four SQM methods it is clear that the MAE at 7.12 kcal/mol for PM3 is significantly lower than for PM6, GFN1-xTB, and GFN2-xTB with MAEs of 10.35, 10.97, and 12.93, respectively.
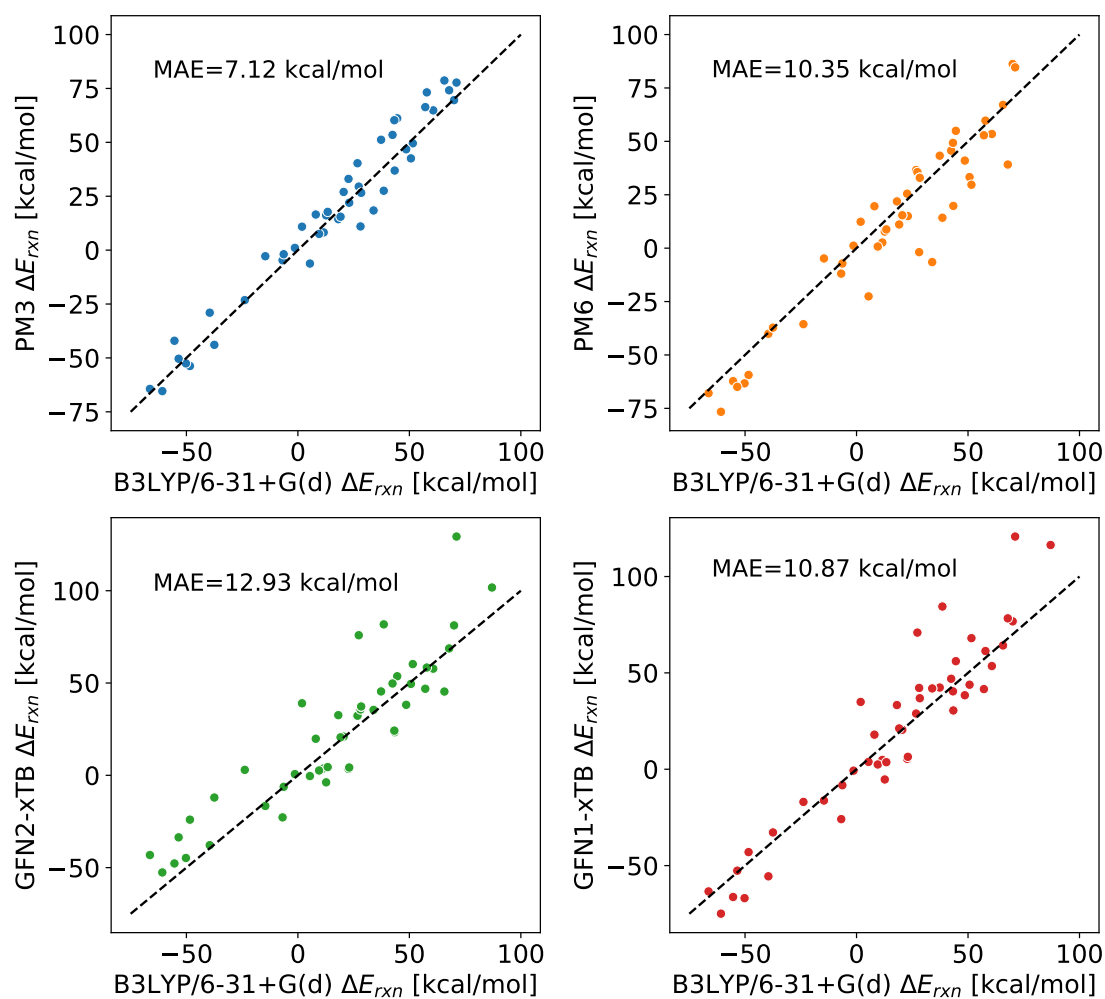
Figure S2: Comparison between DFT and SQM (PM3, PM6, GFN1-xTB, GFN2-xTB) reaction energies computed by summation of fragment energies. Reactions 10, 12, 31, 32, 34, 64 are removed from the benchmark since we can not confirm the reactions by IRC.
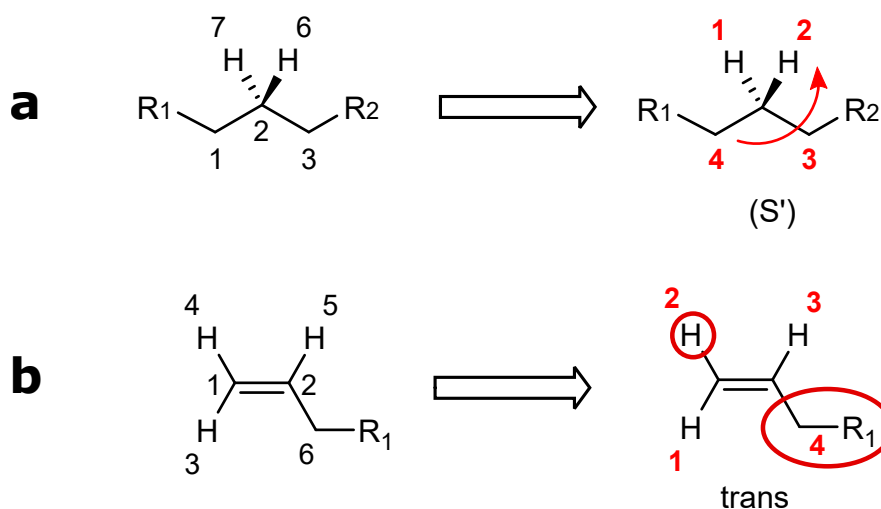
Figure S3: Illustration of a) R'/S' label chirality and b) cis/trans label isomerism. The assignment of label chirality generally follows the ordinary procedure for assigning chirality (Chen-Ingold-Prelog sequence rules), however, the priority of the neighbouring atoms are instead assigned as inversely proportional to the atom map number: rank $\propto 1/$(atom map number).
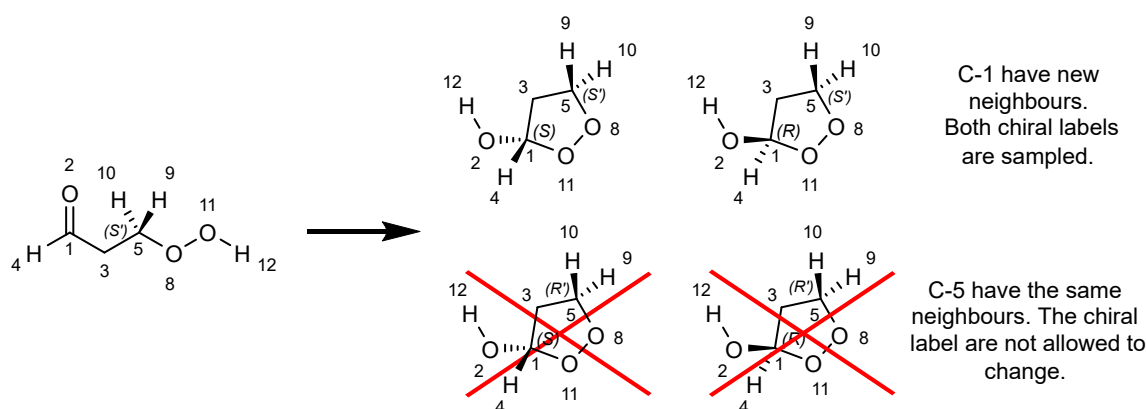
Figure S4: Example of which chiral tags are allowed if the neighbouring atoms change or do not change.

A	[C:1](=[O:2])([C:3]([C:5]([O:8][O:11][H:12])([H:9])[H:10])([H:6])[H:7])[H:4]
B	[C:1](=[O:2])([C@:3]([C@:5]([O:8][O:11][H:12])([H:9])[H:10])([H:6])[H:7])[H:4]
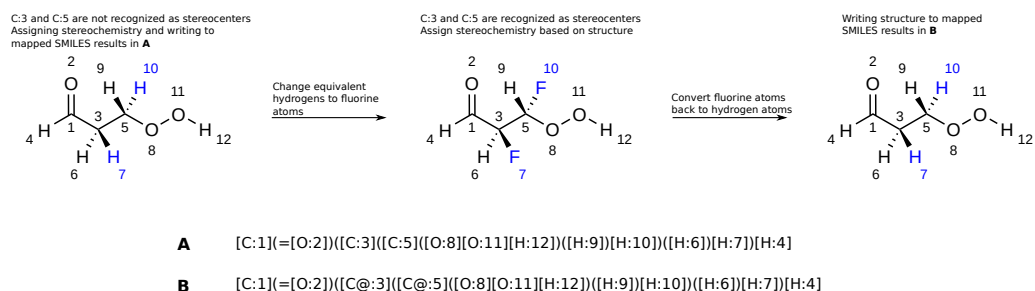
Figure S5: Assigning stereochemistry to non-chiral atoms using the RDKit
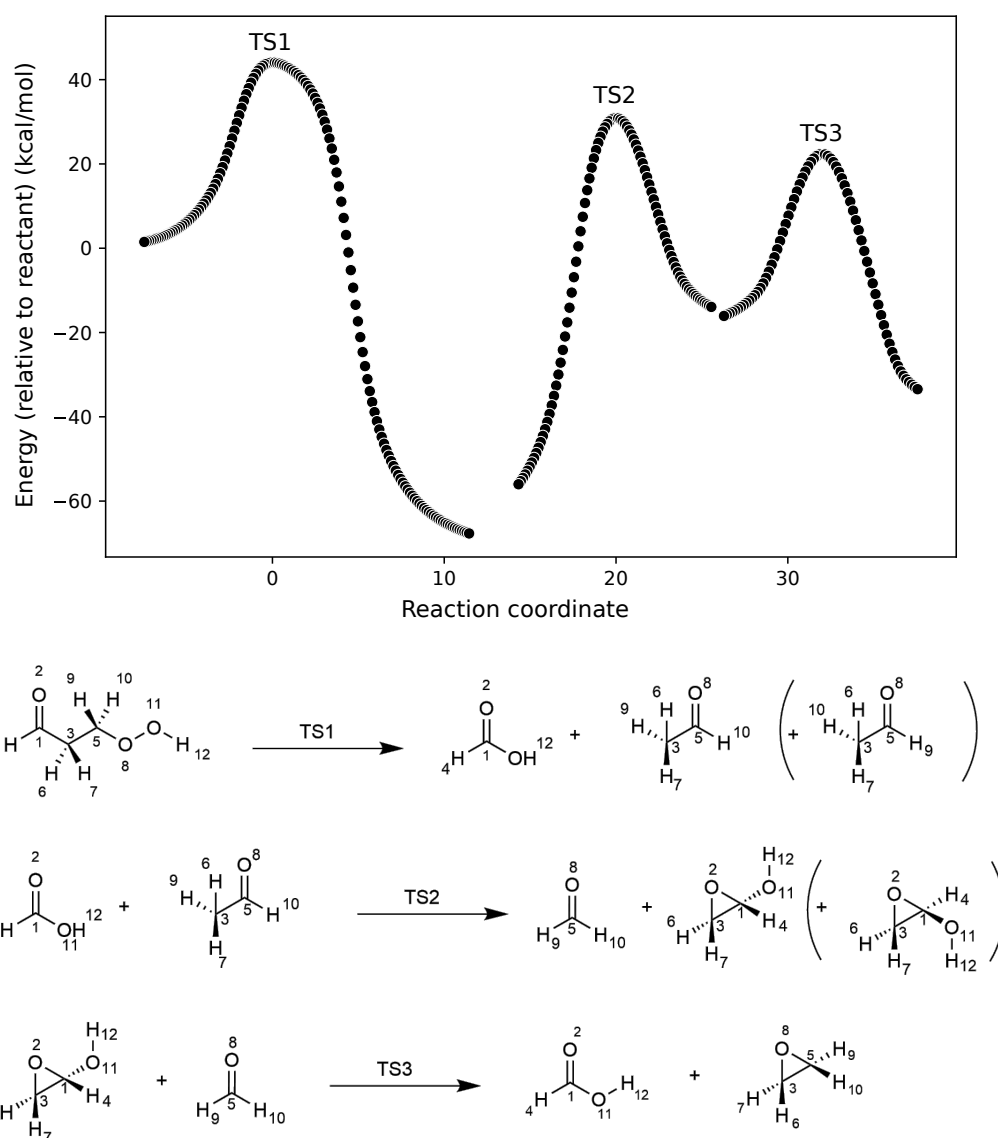
Figure S6: Path located from the optimization of TS guesses. TS1 connects 3-hydroperoxypropanal to R69, TS2 connects R69 to N1a, and TS3 connects N1a to P43a. P43a differs from R43 in that H6 and H7 are switched.
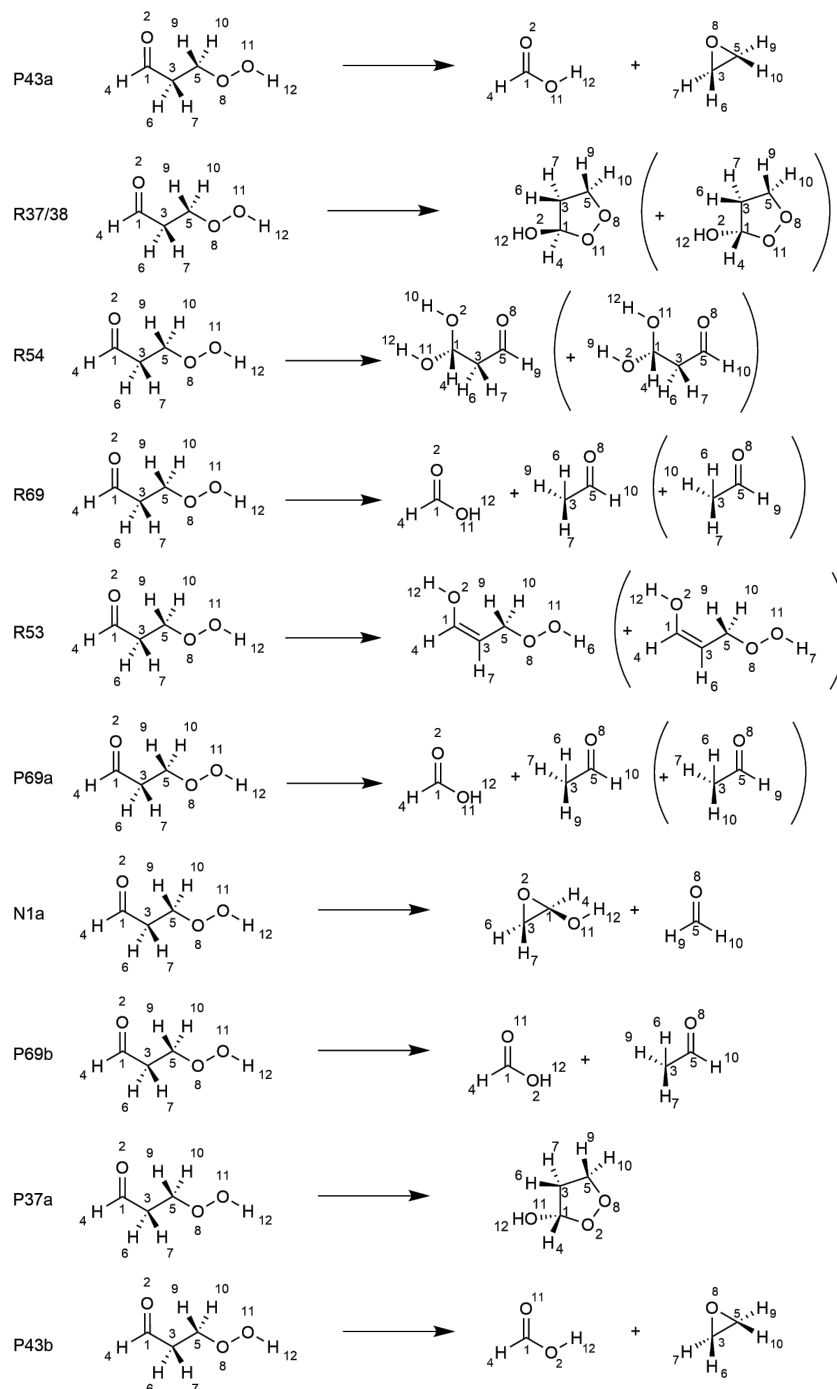
Figure S7: Reactions found in the meta-dynamics search, with a GFN2-xTB barrier estimate below 50 kcal/mol.
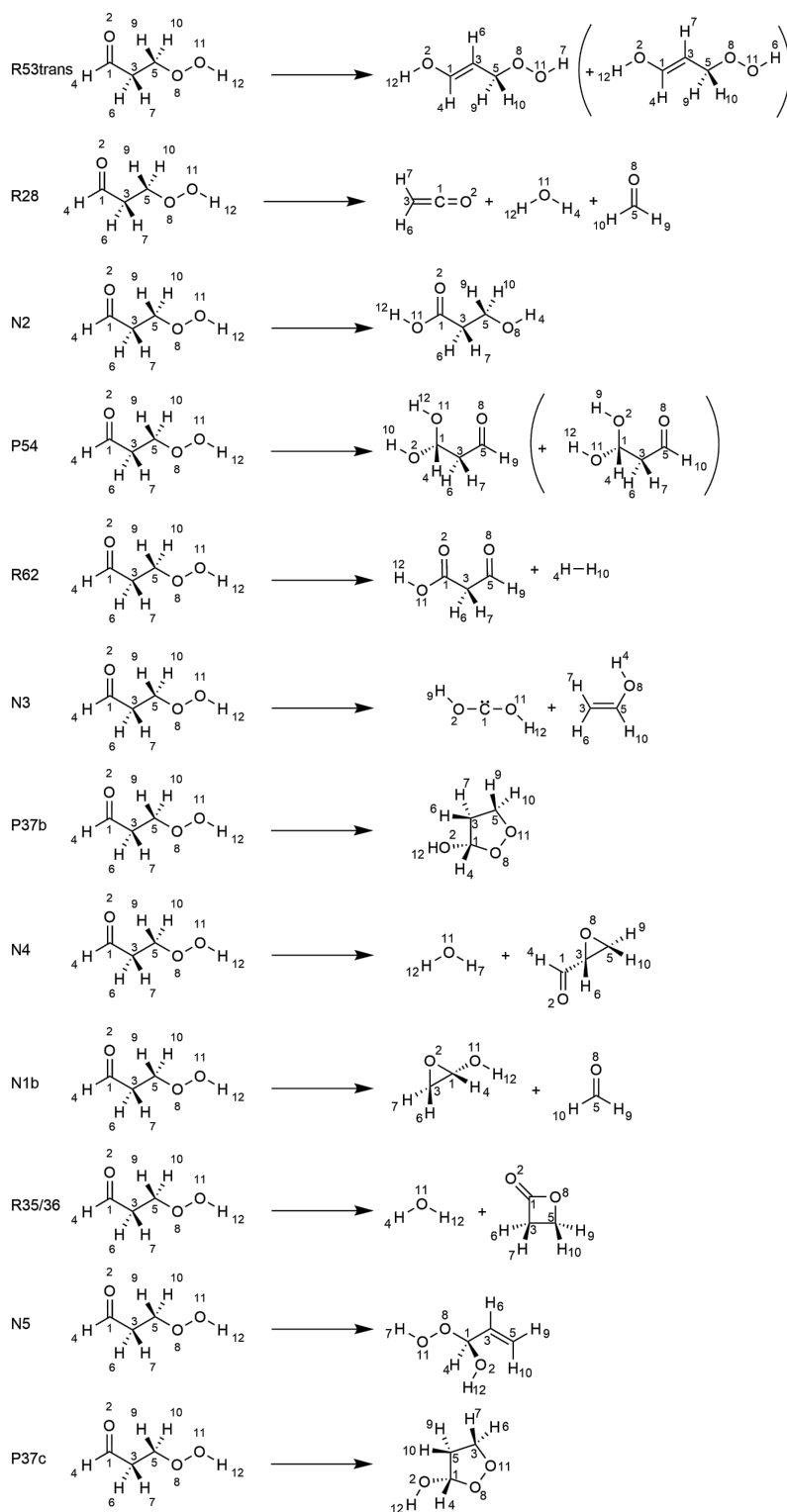
Figure S8: Reactions found in the meta-dynamics search, with a GFN2-xTB barrier estimate above 50 kcal/mol.
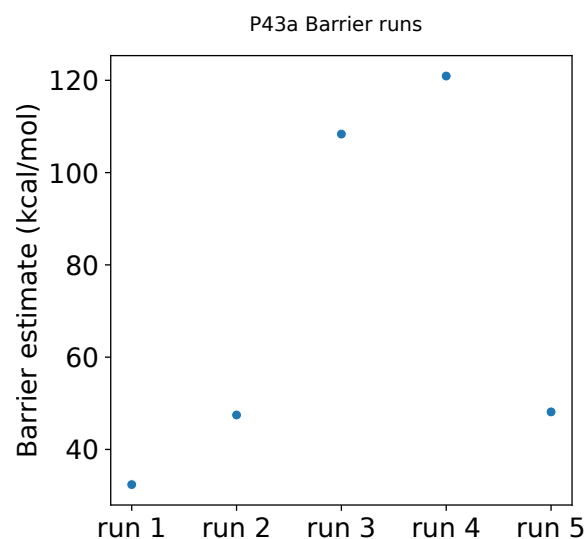
Figure S9: barrier estimates for each of the 5 GFN2-xTB RMSD-PP runs for P43a

Reactions from Grambow et al.

Table S1: Table of the 75 studied reactions found by Grambow

	Reaction	Comments
R1		
R2		
R3		
R4		
R5		
R6		
R7		
R8		

R9		
R10		Following the IRC the proposed structure of Grambow continued to react. Same as R28
R11		
R12		Following the IRC, the product structure stated by Grambow continued to react. (same product as R59 but different mechanism)
R13		
R14		
R15		
R16		
R17		
R18		
R19		
R20		
R21		

R22		
R23		
R24		
R25		
R26		
R27		
R28		
R29		
R30		
R31		Following the IRC, the product proposed by Grambow continues to react to the indicated product
R32		Following the IRC leads to the same reaction as R53
R33		
R34		IRC converged after one step in each direction. The imaginary frequency is ≈ 100 cm^{-1}

R35	
R36	
R37	
R38	
R39	
R40	
R41	
R42	
R43	
R44	
R45	
R46	
R47	
R48	
R49	

R50		
R51		
R52		
R53		
R54		
R55		
R56		
R57		
R58		
R59		
R60		
R61		
R62		Following the IRC suggested by Grambow the product continues to react to the indicated product
R63		

R64		
R65		
R66		
R67		
R68		
R69		
R70		
R71		
R72		
R73		
R74		
R75		
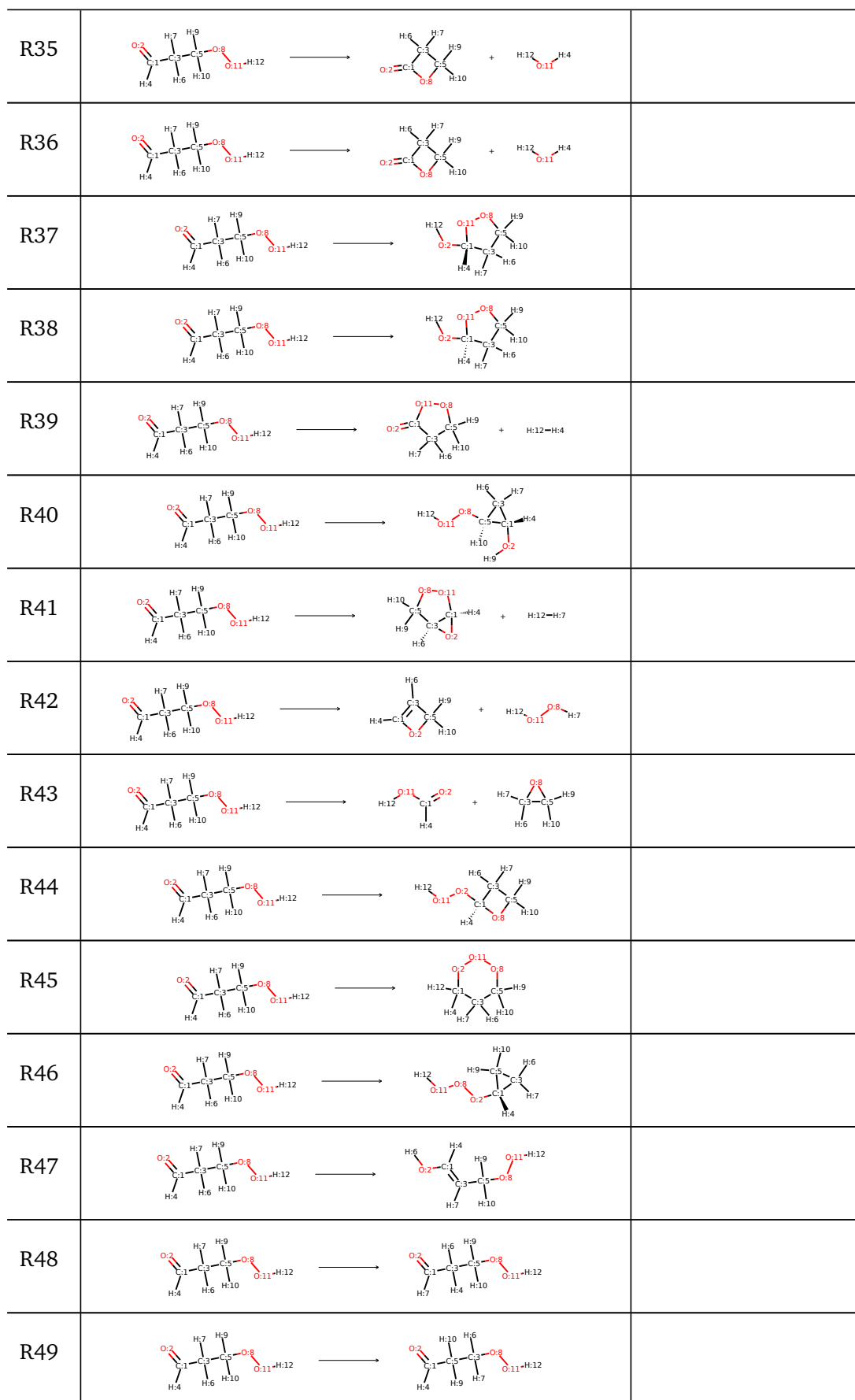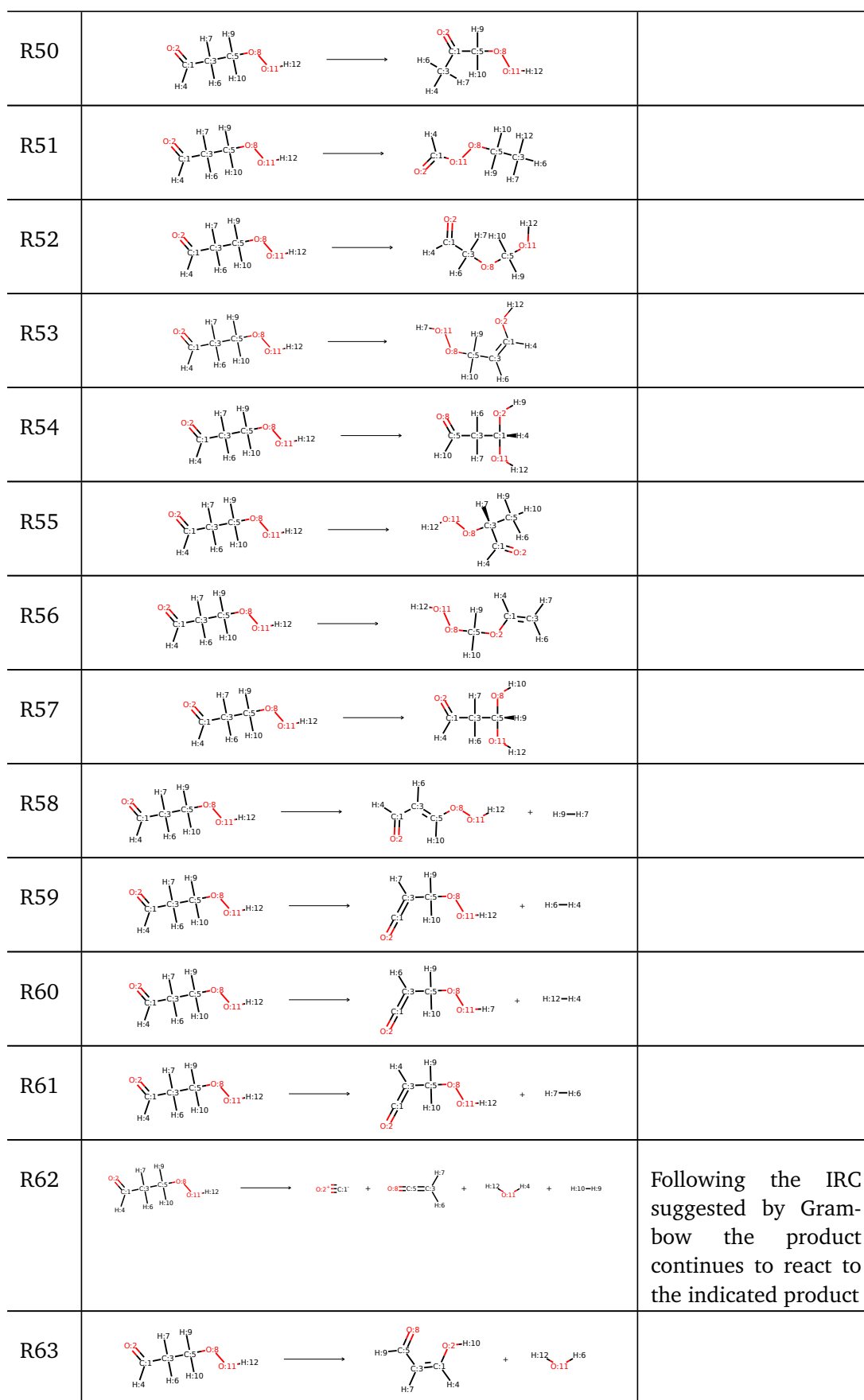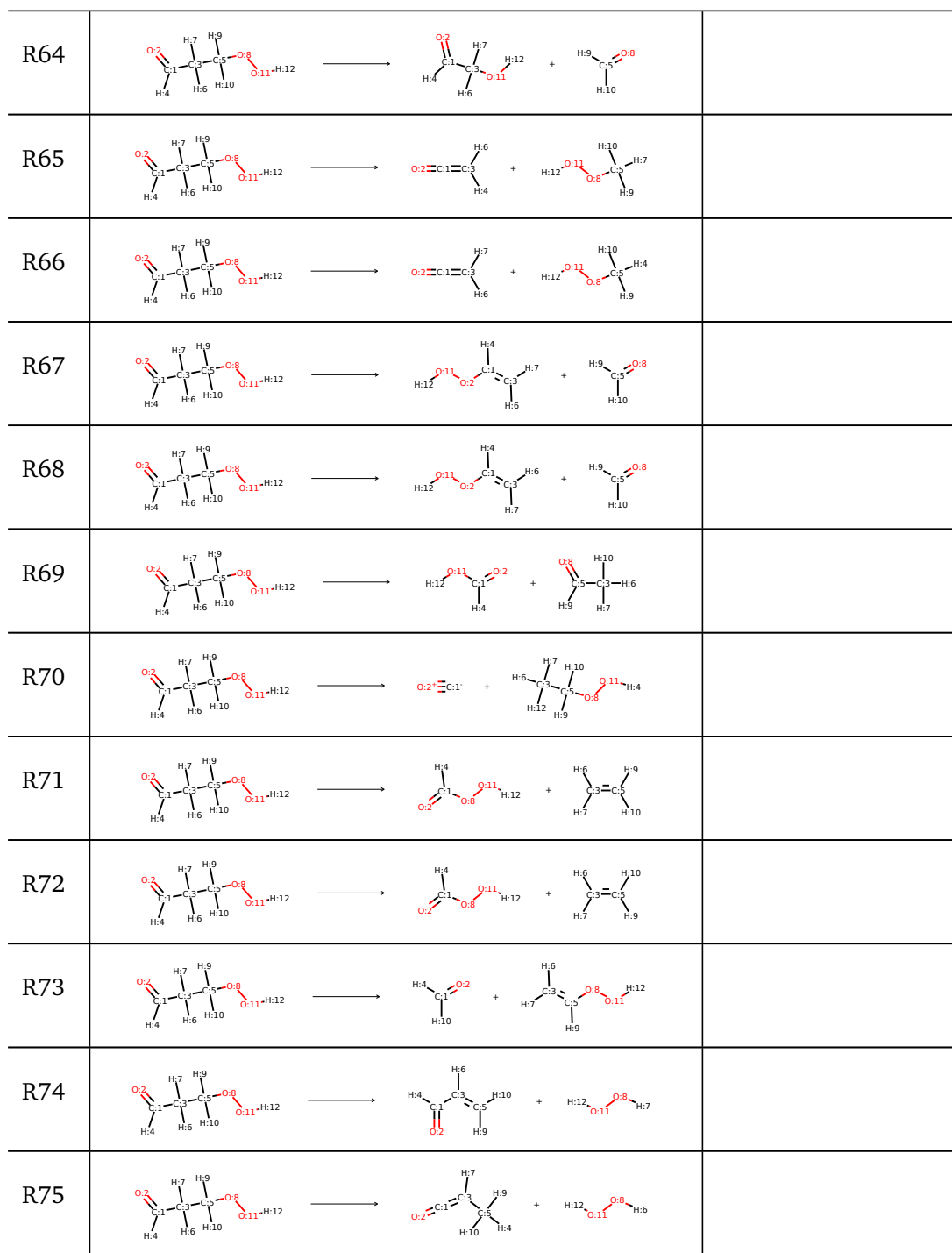
Table S2: Table of how many times each of the three lowest barrier reactions were found for each parameter set. Each parameter set was tested in 100 simulations (50ps) and N_{total} states how many times a reaction with the required reactant occurred. N_{99} is the estimated number of times the meta-dynamics should be run for each parameter set in order to be 99% sure of finding all 3 reactions.

s	k_{push}	α	reaction 1	reaction 2	reaction 3	N_{total}	N_{99}
0.6	0.01	0.7	0	0	0	0	-
0.6	0.01	0.5	0	0	0	0	-
0.6	0.01	0.3	0	0	0	0	-
0.6	0.03	0.7	1	0	0	6	-
0.6	0.03	0.5	6	1	1	34	527
0.6	0.03	0.3	3	3	2	95	236
0.6	0.05	0.7	16	2	3	75	233
0.6	0.05	0.5	9	4	1	100	459
0.6	0.05	0.3	1	2	0	100	-
0.8	0.01	0.7	0	0	0	0	-
0.8	0.01	0.5	0	0	0	0	-
0.8	0.01	0.3	0	0	0	0	-
0.8	0.03	0.7	1	0	0	3	-
0.8	0.03	0.5	2	3	0	25	-
0.8	0.03	0.3	4	7	8	88	114
0.8	0.05	0.7	5	6	5	48	107
0.8	0.05	0.5	12	4	8	101	114
0.8	0.05	0.3	10	8	8	103	65
1.0	0.01	0.7	0	0	0	0	-
1.0	0.01	0.5	0	0	0	0	-
1.0	0.01	0.3	0	0	0	0	-
1.0	0.03	0.7	0	0	0	4	-
1.0	0.03	0.5	2	3	2	16	264
1.0	0.03	0.3	3	10	6	87	152
1.0	0.05	0.7	4	12	2	43	229
1.0	0.05	0.5	9	11	5	100	91
1.0	0.05	0.3	4	7	6	101	116

Table S3: Table of how many times each of the three lowest barrier reactions were found for each parameter set. Each parameter set was tested in 100 simulations and N_{total} states how many times a reaction with the required reactant occurred. N_{99} is the estimated number of times the meta-dynamics should be run for each parameter set in order to be 99% sure of finding all 3 reactions.

s	k_{push}	α	reaction 1	reaction 2	reaction 3	N_{total}	N_{99}
0.6	0.01	0.7	0	0	0	0	-
0.6	0.01	0.5	0	0	0	0	-
0.6	0.01	0.3	0	0	0	6	-
0.6	0.03	0.7	22	2	5	99	228
0.6	0.03	0.5	15	3	5	100	153
0.6	0.03	0.3	3	3	2	100	236
0.6	0.05	0.7	19	3	6	100	152
0.6	0.05	0.5	9	4	1	100	459
0.6	0.05	0.3	1	2	0	100	-
0.8	0.01	0.7	0	0	0	0	-
0.8	0.01	0.5	0	0	0	0	-
0.8	0.01	0.3	0	0	0	5	-
0.8	0.03	0.7	7	7	11	79	74
0.8	0.03	0.5	6	10	9	101	77
0.8	0.03	0.3	4	7	13	101	113
0.8	0.05	0.7	12	9	8	101	61
0.8	0.05	0.5	12	4	8	103	114
0.8	0.05	0.3	10	8	8	103	65
1.0	0.01	0.7	0	0	0	0	-
1.0	0.01	0.5	0	0	0	1	-
1.0	0.01	0.3	0	1	0	8	-
1.0	0.03	0.7	5	5	11	82	104
1.0	0.03	0.5	5	8	12	102	91
1.0	0.03	0.3	3	10	6	101	152
1.0	0.05	0.7	8	14	4	101	114
1.0	0.05	0.5	9	11	5	100	91
1.0	0.05	0.3	4	7	6	101	116

Table S4: Table of how many times each of the three lowest barrier reactions were found for each parameter set. Each parameter set was tested in 100 simulations and N_{total} states how many times a reaction with the required reactant occurred with GFN0-xTB

s	k_{push}	α	reaction 1	reaction 2	reaction 3	N_{total}	N_{99}
0.6	0.005	0.7	0	1	0	101	-
0.6	0.01	0.7	0	6	0	100	-
0.6	0.03	0.7	0	6	0	100	-
0.6	0.03	0.5	0	7	0	100	-
0.6	0.05	0.7	0	11	0	100	-
0.8	0.03	0.7	0	0	0	100	-
0.8	0.03	0.5	0	1	0	101	-
0.8	0.03	0.3	0	0	0	100	-
0.8	0.05	0.7	0	1	0	100	-
0.8	0.05	0.5	0	0	0	100	-
0.8	0.05	0.3	0	0	0	100	-
1.0	0.03	0.7	0	0	0	100	-
1.0	0.03	0.5	0	0	0	100	-
1.0	0.05	0.7	0	0	0	100	-
1.0	0.05	0.5	0	1	0	100	-
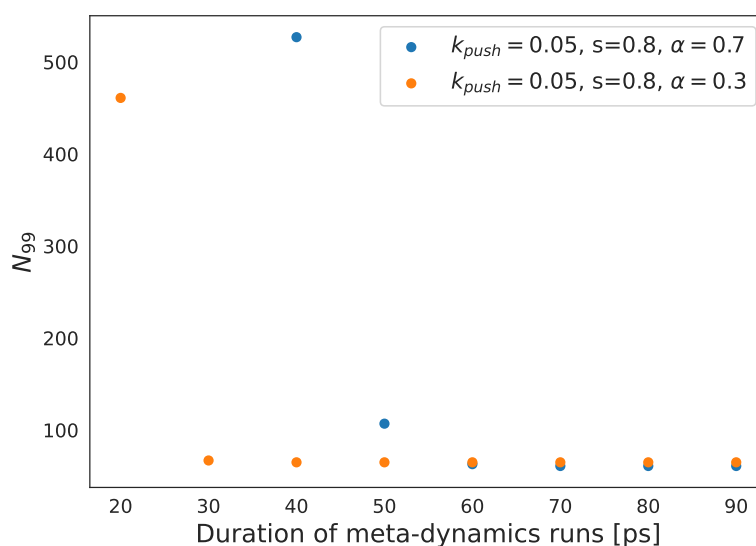
Figure S10: Showing the estimated number of required meta-dynamics runs to ensure 99 % probability of the three low-barrier reactions having been found versus the duration of each meta-dynamics run. Blue dots are for the most effective parameters when running 100 ps simulations (Table S3). Orange dots are for the most effective parameter set when running 50 ps simulations (Table S2).