

Fast and automated identification of reactions with low barriers using meta-MD simulations

Maria H. Rasmussen¹ and Jan H. Jensen^{1,2}

¹Department of Chemistry, University of Copenhagen, Denmark

²E-mail: jhjensen@chem.ku.dk, Twitter: @janhjensen

December 14, 2021

Abstract

We test our meta-molecular dynamics (MD) based approach for finding low-barrier (<30 kcal/mol) reactions (*SciPost Chem.* 2021, 1, 003) on uni- and bimolecular reactions extracted from the barrier dataset developed by Grambow et al. (*Scientific Data* 2020, 7, 137). For unimolecular reactions the meta-MD simulations identify 25 of the 26 products found by Grambow et al., while the subsequent semiempirical screening eliminates an additional four reactions due to an overestimation of the reaction energies or estimated barrier heights relative to DFT. In addition, our approach identifies an additional 36 reactions not found by Grambow et al., 10 of which are <30 kcal/mol. For bimolecular reactions the meta-MD simulations identify 19 of the 20 reactions found by Grambow et al., while the subsequent semiempirical screening eliminates an additional reaction. In addition, we find 34 new low-barrier reactions. For bimolecular reactions we found that it is necessary to "encourage" the reactants to go to previously undiscovered products, by including products found by other MD simulations when computing the biasing potential as well as decreasing the size of the molecular cavity in which the MD occurs, until a reaction is observed. We also show that our methodology can find the correct products for two reactions that are more representative of those encountered in synthetic organic chemistry. The meta-MD hyperparameters used in this study thus appears to be generally applicable to finding low-barrier reactions.

1 Introduction

Understanding how molecular systems react, i.e. which reactions are possible under what conditions, is an essential part of chemical research and computational methods for exploring reaction space in an automated manner are continually being proposed. [1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11] These methods can be divided into three different categories: (semi-)exhaustive searches,[5, 3, 1, 2, 11] reaction template methods,[10, 12, 13, 14, 15], and meta-molecular dynamics (meta-MD) based approaches.[10, 16, 11] Examples of (semi-)exhaustive searches include graph enumeration of products[5, 3, 1, 2, 11] and enumeration of reaction coordinates.[14, 15] For these methods the size of the search space, and hence the computational cost, grows quickly with the size of the molecules. The reaction template approaches lie at the other extreme in terms of computational efficiency, and work by investigating only pre-determined reaction types. Though efficient and used extensively in atmospheric and combustion chemistry, this approach can be hard to generalise to other areas such as synthetic organic chemistry, though a recent attempt is encouraging.[13] The meta-MD approaches explore reactivity via biasing potentials that force reactions and exploration of conformational space. The meta-MD approach can also be combined with the reaction coordinate approach as shown by Lavigne et al.[9] In the non-exhaustive approaches the key question is whether they identify *all* relevant reactions for the problem at hand. At room temperature, this typically means all reactions with barriers less than ca 30 kcal/mol.

Recently, we demonstrated that a combination of product generation using meta-MD and barrier estimation + TS guess generation using the RMSD-PP method [16], both relying on semiempirical GFN2-xTB

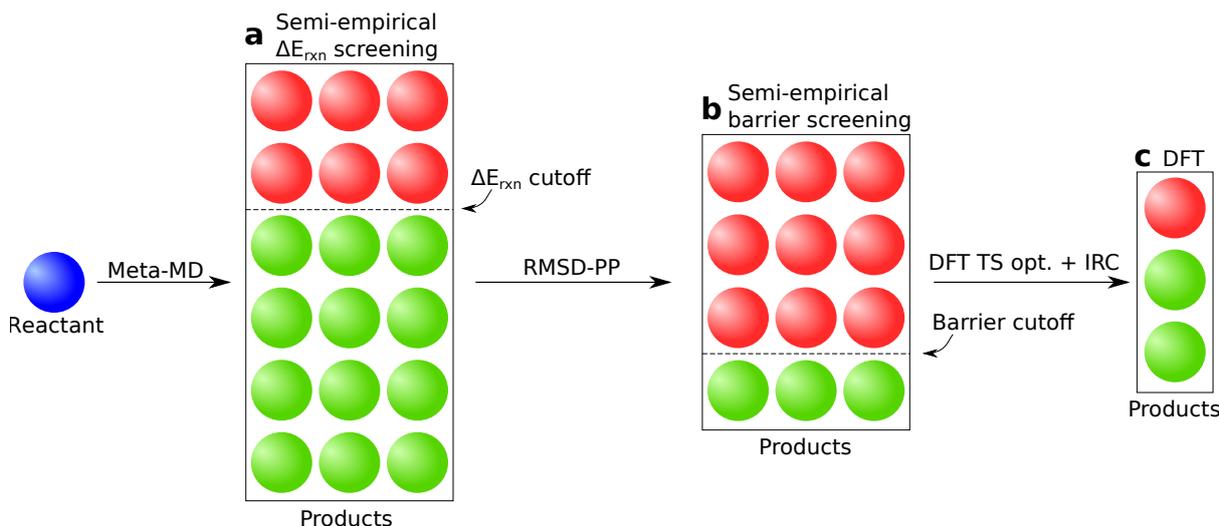


Figure 1: Schematics of the method used to predict low-barrier reactions employed in this study. A cutoff value of 40 kcal/mol is used to screen for low-barrier reactions.

[17] calculations, could efficiently suggest the three lowest barrier elementary reactions of 3-hydroperoxypropanal.[11] The results depend strongly on the parameters that control the biasing potential in the meta-MD and it is unclear how well this parameter set generalizes to other unimolecular reactants and to bimolecular reactions in general. Here we test the performance of this parameter set on elementary unimolecular reactions involving 163 reactant molecules and 20 elementary bimolecular reactions, both extracted from the recently published database of elementary reactions on two different DFT levels.[18] In addition we test the method on two multi-step reactions related to organic synthesis.

2 Methods

Our method for predicting the kinetically important elementary reactions[11] is based on three steps (Figure 1): (1) the generation of possible single-step products which is based on meta-molecular dynamics (meta-MD) simulations followed by (2) screening of the proposed products based on reaction energies and estimated barrier heights computed at the semiempirical (GFN2-xTB) level of theory followed by (3) validation at the DFT level of theory. In this study we use a cutoff of 40 kcal/mol when screening reaction energies and barrier estimates. Our approach to product generation is based on the meta-MD approach by Grimme [16] which is a way of increasing the likelihood of a reaction occurring during an MD simulation by penalising to previously visited structures. This is done by adding additional terms to the energy and gradient that depend on the RMSD from previously visited structures during the current MD simulation or previous MD simulations (selected by the user). The additional energy terms depend on the hyper-parameters k_{push} , α , and s and we use values that were optimised to promote unimolecular chemical reactions[11] ($k_{push} = 0.05$, $\alpha = 0.3$, $s = 0.8$) along with an additional set found as part of this study ($k_{push} = 0.03$, $\alpha = 0.7$, $s = 0.6$). For the bimolecular reactions, we change the procedure a bit. Initial runs showed greatly increased run times compared to single-fragment reactants using the original parameters. To decrease run times we decrease s (which scales the size of the molecular cavity) by 0.02 every 5 ps as long as no reaction has occurred, thereby forcing the reactant molecules closer together.

We generate different random Cartesian coordinates ("embedding" in RDKit) for the reactants for each of the meta-MD runs. If the reactant consists of multiple molecules, they are first embedded randomly on top of each other using RDKit [19] with a subsequent force field optimization of each fragment. The second molecule is then moved in a random direction by a distance, d :

$$d = 0.5 \cdot (D_{max,1} + D_{max,2}) + 2\text{\AA} \quad (1)$$

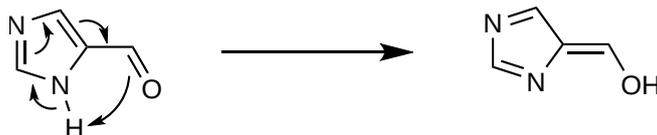


Figure 2: The reaction not found by either meta-MD parameter sets. The reaction energy is $\Delta E = 19$ kcal/mol and the barrier is $\Delta E^\ddagger = 27$ kcal/mol calculated with ω B97X-D3/def2-TZVP [18]

where $D_{max,i}$ is the maximum distance between any two atoms in molecule i . The coordinates of the reactants are energy minimised with GFN2-xTB before starting the meta-MD. If a change in atomic connectivity is detected the meta-MD simulation is skipped and the barrier estimation is done using the unoptimised reactant structure. We perform 100 meta-MD simulations for each reactant unless otherwise noted. Our algorithm checks for changes in atomic connectivity every 5 ps and the simulation is stopped when this is detected. The output of the meta-MD simulations is then a list of all unimolecular reactions and a database of the reactant and product structures.

Barrier estimates and transition state (TS) guess structures are based on the RMSD-PP procedure by Grimme [16] and is described in detail in [20] and [11]. Instead of embedding the reactant and product structures from the SMILES saved during the product generation as done in [11], we use the optimized structures from the meta-MD procedure as input to the RMSD-PP procedure. For the barrier estimate, the RMSD-PP is run five times, and the lowest barrier estimate is used, as described in [11], except two of the five runs are done starting the procedure from product \rightarrow reactant instead of reactant \rightarrow product. If available, different conformers of the reactant and product is used in each of the five runs. If the reaction is found in five or more of the 100 meta-MD runs, the reactant/product structures are extracted from a different meta-MD simulation in each of the five RMSD-PP calculations. If the reaction is found four times during the meta-MD simulations, two of the RMSD-PP calculations use reactant/product structures from the same meta-MD simulation, while the remaining three RMSD-PP calculations use reactant/product structures from three different runs and so on for cases where the reaction is found three, two or one time during the 100 meta-MD runs. For the computation of reaction energies and barriers we need energies of the reactant and products. The lowest energy structure encountered for each molecule (each unique canonical SMILES) during the geometry optimizations that follow the meta-MD runs is used.

The last step of the procedure is the DFT-refinement (ω B97X-D/def2-TZVP) of the reactions found at the semiempirical level of theory that have reaction energies and barriers below 40 kcal/mol. Again, our validation procedure is as described in [20] and [11] except that we only test one of the five possible TS guess structures that can be extracted from the five RMSD-PP runs in order to reduce the computational cost of the DFT part of the procedure. The barriers and reaction energies at DFT level of theory are computed as stated in [18], adding zero-point vibrational energies to the electronic energies of reactant, product and transition states (TSs) before calculating the barrier as the energy difference between TS and reactant and reaction energy as the energy difference between product and reactant.

All Density Functional Theory (DFT) calculations are performed using Gaussian 16 [21]. The meta-MD calculations and the RMSD-PP barrier estimates are performed with version 6.1.4 of the `xtb` program [16] using the GFN2-xTB method.[17] All structure-to-SMILES and structure-to-adjacency matrix (AC) ($N_{atoms} \times N_{atoms}$ dimensional matrix with elements either 1 or 0 depending on whether the atom-pair is bound or not) conversions are done using `xyz2mo1`. [22]

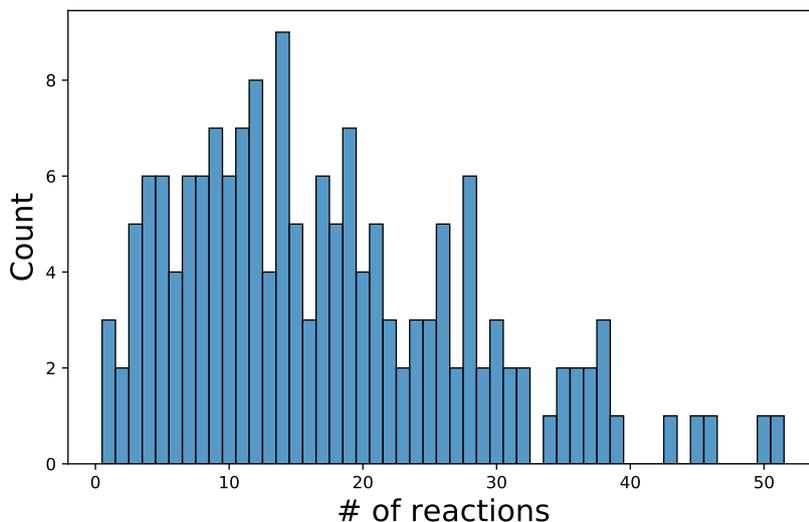


Figure 3: Distribution of the number of different reactions found during the 100 meta-MD runs (with default parameter set) for the 163 single-fragment reactants shown in Figures S1

3 Results and discussion

3.1 Unimolecular reactions

3.1.1 The low-barrier reaction dataset

The low-barrier dataset used in this study is extracted from the dataset created by Grambow et al.[18] The reactants in the Grambow dataset consists of all molecules in the GDB-7 dataset [23] with less than seven heavy atoms plus ca. 430 randomly selected molecules with seven heavy atoms. Reactions and the corresponding transition states (TSs) are located by performing several hundred single-ended growing string method (GSM[14]) searches from each reactant at the B97-D3/def2-mSVP level of theory, followed by TS refinement at the ω B97X-D3/def2-TZVP level of theory. The result is 16,365 and 11,961 unimolecular reaction barriers at the B97-D3/def2-mSVP and ω B97X-D3/def2-TZVP level of theory, respectively. The corresponding number of low-barrier reactions with barriers below 30 kcal/mol is 199 and 30, involving 163 (Figure S1) and 27 different reactants, respectively. We would thus expect that applying our meta-MD search to these 163 reactants (Figure S1) should identify these 30 reactions (Table S1) if we use ω B97X-D3/def2-TZVP for the DFT refinement.

Since the D3 dispersion correction is not available with the ω B97X functional in Gaussian16 we start by reoptimising the 30 TSs at the ω B97X-D/def2-TZVP level of theory and verifying them by performing intrinsic reaction coordinate (IRCs) calculations. For two reactions (R1108 and R7201, using the notation of Grambow et al.) the IRCs go to the correct reactant, but to a different stereoisomer. Since that barrier is below 30 kcal/mol we use the newly found structure to initiate the meta-MD instead of the one proposed by Grambow et al. For five reactions (R1084, R2399, R2523, R6490, and R18816), the IRC does not lead to the reactant proposed by Grambow et al. We subsequently find the TS for R2399 using our meta-MD approach, but we exclude R1084, R2523, R6490, and R18816 from our low-barrier dataset. Thus, we expect that applying our meta-MD search to the 163 reactants described above should identify 26 reactions with barriers below 30 kcal/mol at the ω B97X-D/def2-TZVP level of theory.

3.1.2 Meta-MD based search for low-barrier reactions

Using the default parameter set ($k_{push} = 0.05$, $\alpha = 0.3$ and $s = 0.8$) we find 20 of the 26 reactions. For three of the reactions (R3725, R7207, and R8701) the meta-MD failed to generate the corresponding product structures. Another reaction (R2514) is eliminated due to having a GFN2-xTB reaction energy of 53 kcal/mol, which is significantly higher than the corresponding ω B97X-D3/def2-TZVP-value of -3 kcal/mol. The final two reactions (R4612 and R9011) are eliminated due to high estimated barrier heights of 48 and 47 kcal/mol, respectively - considerably higher than the corresponding ω B97X-D3/def2-TZVP-values of 26 and 30 kcal/mol. The actual barrier heights at the GFN2-xTB level are 42 and 45 kcal/mol, which indicates that the problem lies primarily with the GFN2-xTB method itself and not the barrier-estimation method. If we perform additional meta-MD simulations with a slightly different parameter set ($k_{push}=0.03$, $\alpha = 0.7$, $s = 0.6$) we locate R3725 and R7207, but R3725 is subsequently eliminated due to a high reaction energy (42 kcal/mol) compared to a DFT value of -21 kcal/mol. So, using two set of parameters, the meta-MD simulations identify 25 of the 26 low-barrier reactions reactions found by Grambow et al., but four of these are subsequently eliminated due to overestimated reaction energies and estimated barrier heights by GFN2-xTB. Only R8701 (Figure 2) is not found by the meta-MD, presumably due to a combination of high barrier and reaction energy (27 and 19 kcal/mol at the DFT level) plus a relatively small change in structure on going from reactants to products, resulting in a weak biasing potential.

The four false negatives that result from errors in the GFN2-xTB reaction energies and estimated barrier height (R2514, R3725, R4612, and R9011) all contain a N-N triple bond in the products, indicating a systematic error in the GFN2-xTB method. In the case of reaction energies such errors can be efficiently corrected by, for example, the connectivity-based hierarchy method.[24, 25] Another option is simply to use DFT instead of GFN2-xTB for the subsequent screening. The meta-MD approach produces between 1 and 51 reactions per reactant (Figure 3) which is practical to check with DFT, and certainly with DFT//GFN2-xTB single point calculations. In our current approach 1257 and 2392 reactions are eliminated based on semiempirical reaction energies and estimated barriers, respectively, so that only 316 reactions are checked with DFT out of a total of 3965 candidate reactions.

Our meta-MD based approach also identifies 10 low-barrier reactions, shown in Figure 4, not found by Grambow et al., as well as 26 new reactions with barriers above 30 kcal/mol (Table S2 and S3). All low-barrier reactions involve new reactants not represented in the low-barrier dataset, which indicates that the reactions in that dataset are likely the ones with the lowest possible barriers for each reactant. Eleven of the new high-barrier reactions (Table S2) have barriers that are lower than the lowest barrier found by Grambow et al. for those reactants. Fourteen of the 36 new reactions (N6, N11, N13, N14, N16, N17, N18, N22, N23, N24, N25, N29, N35, and N36) are found by Grambow et al. at the B97-D3/def2-mSVP level of theory but these TS structures are apparently not of sufficient quality for the ω B97X-D3/def2-TZVP refinement. For the remaining reactions, it is not clear whether the new reactions are not found by Grambow et al. due to the growing string method (GSM) itself or the small basis set used for the GSM calculations. One reaction (N32 in Table S3) corresponds to a change in chirality due to a hydrogen transfer and such reactions do not appear to have been reported by Grambow et al., so it is possible that they found it but did not report it.

Two of the new low-barrier reactions (N4 and N7) proceed without a barrier at the GFN-xTB level of theory, while the DFT barriers are 21 and 12 kcal/mol, respectively and both reactions are endothermic at the DFT level of theory. Both reactions involve a N-N triple bond, though that is also the case for several other new low barrier reaction that worked fine.

3.2 Bimolecular reactions

Grambow et al. [18] only searched for elementary reactions starting from single reactant molecules, but they found a lot of products with two fragments. From these back-reactions, we extract a set of 20 target reactions with barriers below 30 kcal/mol, where two molecules react to create a single molecule

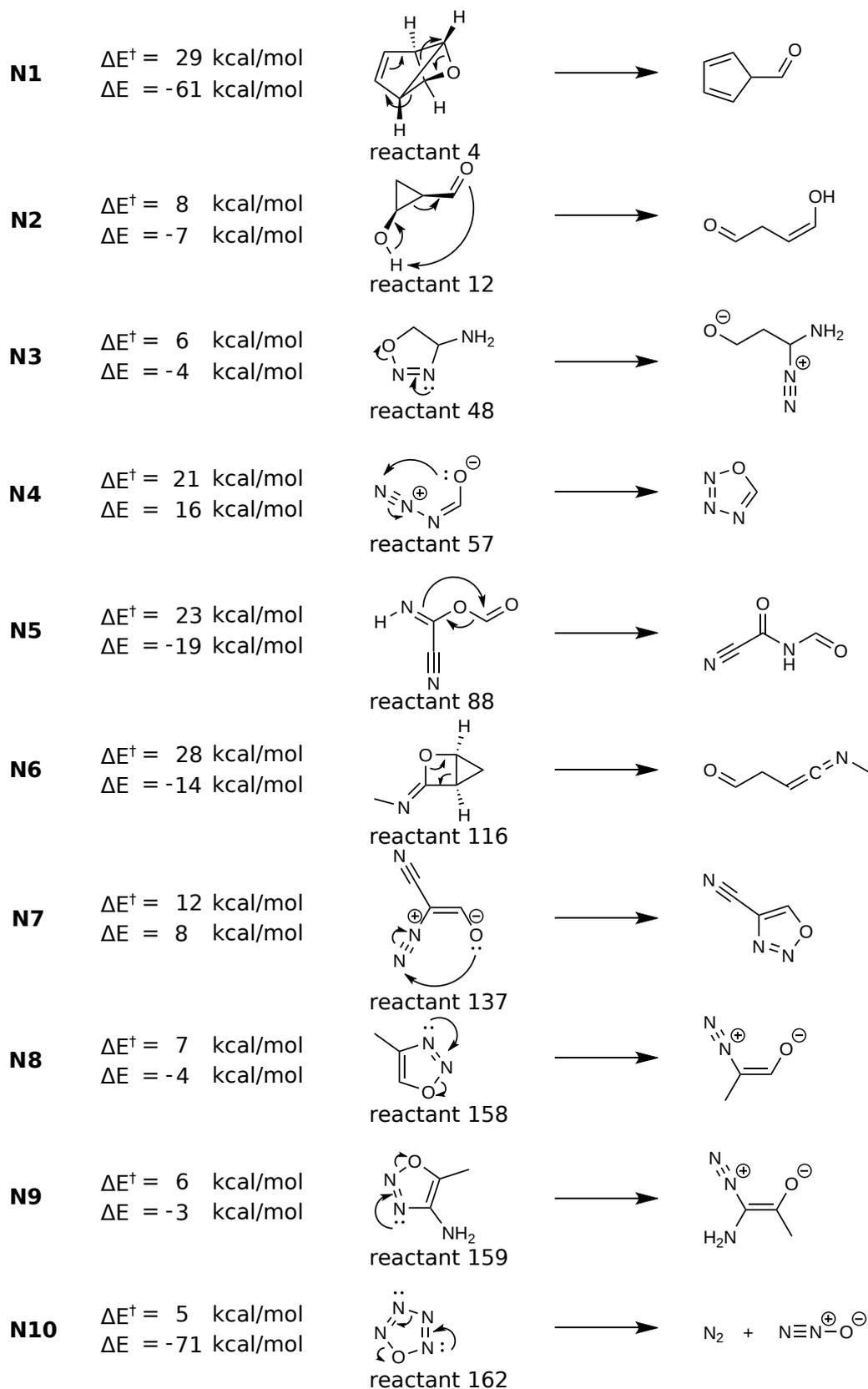


Figure 4: 10 new reactions found with barriers below 30 kcal/mol. The stated barriers (ΔE^\ddagger) and reaction energies (ΔE) are computed at the ω B97X-D/def2-TZVP level of theory.

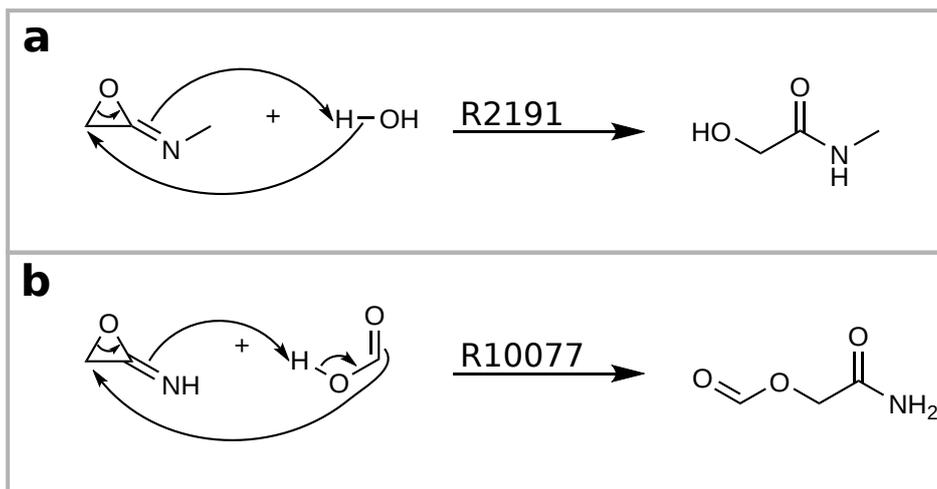


Figure 5: **(a)** The reaction not found by any meta-MD parameter sets (R2191). The reaction energy is $\Delta E = -46$ kcal/mol and the barrier is $\Delta E^\ddagger = 19$ kcal/mol calculated with ω B97X-D3/def2-TZVP [18]. **(b)** Target reaction R10077 which is found with meta-MD using the secondary parameter set, but where the GFN2-xTB barrier estimate from RMSD-PP is too high (60 kcal/mol). The reaction energy is $\Delta E = -39$ kcal/mol and the barrier is $\Delta E^\ddagger = 10$ kcal/mol calculated with ω B97X-D3/def2-TZVP [18]

(Table S4). The barriers range between 8 and 28 kcal/mol and reaction energies range between -58 and 8 kcal/mol. Unlike for the unimolecular reactants above, these bimolecular reactants were not subjected to a thorough reaction discovery search so it is more likely that there will be additional reactions with barriers below 30 kcal/mol. As with the unimolecular reactions above, we re-optimize the TS structures provided by Grambow et al. at the ω B97X-D/def2-TZVP level of theory and confirm that they connect the stated reactant and product with an IRC.

Again, we first run 100 meta-MD simulations for each of the 20 reactant pairs with the primary parameter set ($k_{push}=0.05$, $\alpha = 0.3$, $s = 0.8$), which results in a total of 588 elementary reactions (average of 29.4 per reactant). Among these reactions are 16 of the 20 target reactions but R129, R2191, R10077 and R7854 (Table S4) are not found. The lower success rate compared to unimolecular reactions (16/20 vs 23/26 for this parameter set) is likely due to the increased number of reactions per reactants (29.4 vs 17.3 reactions per reactant on average). For a reactant system with only a single viable path (low-barrier single-step product) we expect a high probability that at least one of the 100 meta-MD simulations will go to that product. However, for a reactant system with, say, 29 low-barrier single-step reactions there is a good chance that at least one of the 29 reactions will not be found by meta-MD in any of the 100 simulations. Assuming all 29 reactions are found with equal likelihood there is only a 40% chance that 100 meta-MD runs will find all 29 reactions, compared to a 96% chance for 17 low-barrier reactions. To account for this, we try to "encourage" the reactants to go to previously undiscovered products, by including products found by other MD simulations when computing the biasing potential. After filtering the 588 reactions found in the first set of runs with RMSD-PP estimated barriers, 318 reactions are left for DFT refinement. For the second set of runs (initiated with products from the first set of runs), after filtering based on both RMSD-PP barrier estimates and duplicates of reactions from the first run, 174 reactions are left for DFT refinement. We find one additional target reaction in this way: R129. Two (R7854 and R10077) of the remaining three target reactions can be found with meta-MD by changing the parameter set to our secondary choice ($k_{push} = 0.03$, $\alpha = 0.7$ and $s = 0.6$).

With the three different kind of runs tested here we find 19 of the 20 target reactions shown in Table S4. The one reaction not found by meta-MD (R2191) is shown in Figure 5(a). Though R10077 (Figure 5(b)) was found by meta-MD with our secondary parameter set, it was predicted to have a too high barrier (≈ 60 kcal/mol) using the GFN2-xTB RMSD-PP estimate. We note that these two target reactions (Figure 5) both involve oxiranimines and have similar mechanisms. Unlike the reactions causing problems in the search from single molecule reactants above, these two reactions have low

reaction energies at the GFN2-xTB level of theory (-37 kcal/mol for R2191 and -29 kcal/mol for R10077).

The DFT refinement step localizes the 18 target reactions remaining at this point as well as 34 new reactions with barriers below 30 kcal/mol. Twenty-five of the new reactions are located from the reactions found using our default parameter set while the remaining nine reactions are found by penalising products found by the first meta-MD runs. The 34 new reactions are spread across 12 of the 20 reactants. Figure 6 shows the lowest-barrier reaction for each of these 12 reactants and the remaining 22 new low-barrier reactions can be found in Figure S2. The new reactions represent a range of reaction energies between -47 and 20 kcal/mol and barriers ranging from 0.5 to 29 kcal/mol. We find both reactions where the two reactant molecules react with each other and unimolecular reactions where one of the reactant molecules goes through an isomerization reaction.

3.3 Application to organic chemistry

The reactions tested thus far involve relatively small molecules, often with functional groups not usually seen in organic chemistry. We thus test our methodology on two reactions, one unimolecular and one bimolecular, that are more representative of those encountered in synthetic organic chemistry. Our goal here is simply to check whether the correct products can be found with the current parameters rather than an exhaustive computational study of the reaction mechanisms.

3.3.1 A unimolecular reaction

Inspired by Lavigne et al. [9] we study an important step of the synthesis of Berkelyone A reported by Elkin et al. (Figure 7).[26] We choose the same protonated epoxide reactant structure as Lavigne et al. for the starting point of our analysis (Figure 7). In practice several protonation sites must be investigated but this process can easily be automated. Figure 8 highlights some pathways found using meta-MD for product generation and RMSD-PP for barrier estimates at the GFN2-xTB level. As our goal with this example is to check the ability of meta-MD + RMSD-PP to give insight into more complicated multi-step reactions compared to the elementary reactions studied until now we skip the DFT validation step and report the GFN2-xTB energies and barrier estimates. Thus, the energetics presented in Figure 8 are not expected to be quantitatively accurate.

Doing meta-MD + RMSD-PP starting from the reactant structure (**R**) produces reactions involving ring-opening of the protonated epoxide to produce both secondary and tertiary carbocations, proton transfer from the epoxide to the carbonyl oxygen of the ester group, as well as tautomerization reactions. Instead of restarting the procedure from every one of the produced intermediates, we choose to follow the path involving the tertiary carbocation **I1** further. Continuing this process we create reaction profiles as presented in Figure 8. We locate a possible mechanism for the path to the product (**P**) through the five steps connected in black. We also show a pathway to one of the other (less stable) stereoisomers of the product (**Pa**, blue) as well as the path to the most stable product encountered (**Pb**, purple). We note that both the intermediate **I2** and **Pb** is also found in the study conducted by Lavigne et al. We also find paths to macrocycle structures (from **I1**) which is also found by them. The most stable intermediate (**I2**), a bicyclic ether, is a known byproduct when doing this type of epoxide-initiated cyclizations. [27, 28] The competition between carbon cyclisation (**P**) and oxygen cyclisation (**Pb**) is also a known problem for these reactions. [27, 28]

3.3.2 A bimolecular reaction example

Next we study the acid catalyzed synthesis of a benzimidazole derivative starting from ortho-phenylenediamine and benzoic acid (Phillips method).[29] The reaction is acid catalyzed so at each minimum along the path we try the different possible protonated structures and optimize with GFN2-xTB. Structures with energies < 30 kcal/mol relative to the lowest energy protonation structure are considered relevant for the analysis. The result of the search is summarized in Figure 9. For the reactant system, two protonation structures are possible: protonation at the nitrogen or at the carbonyl oxygen (which is

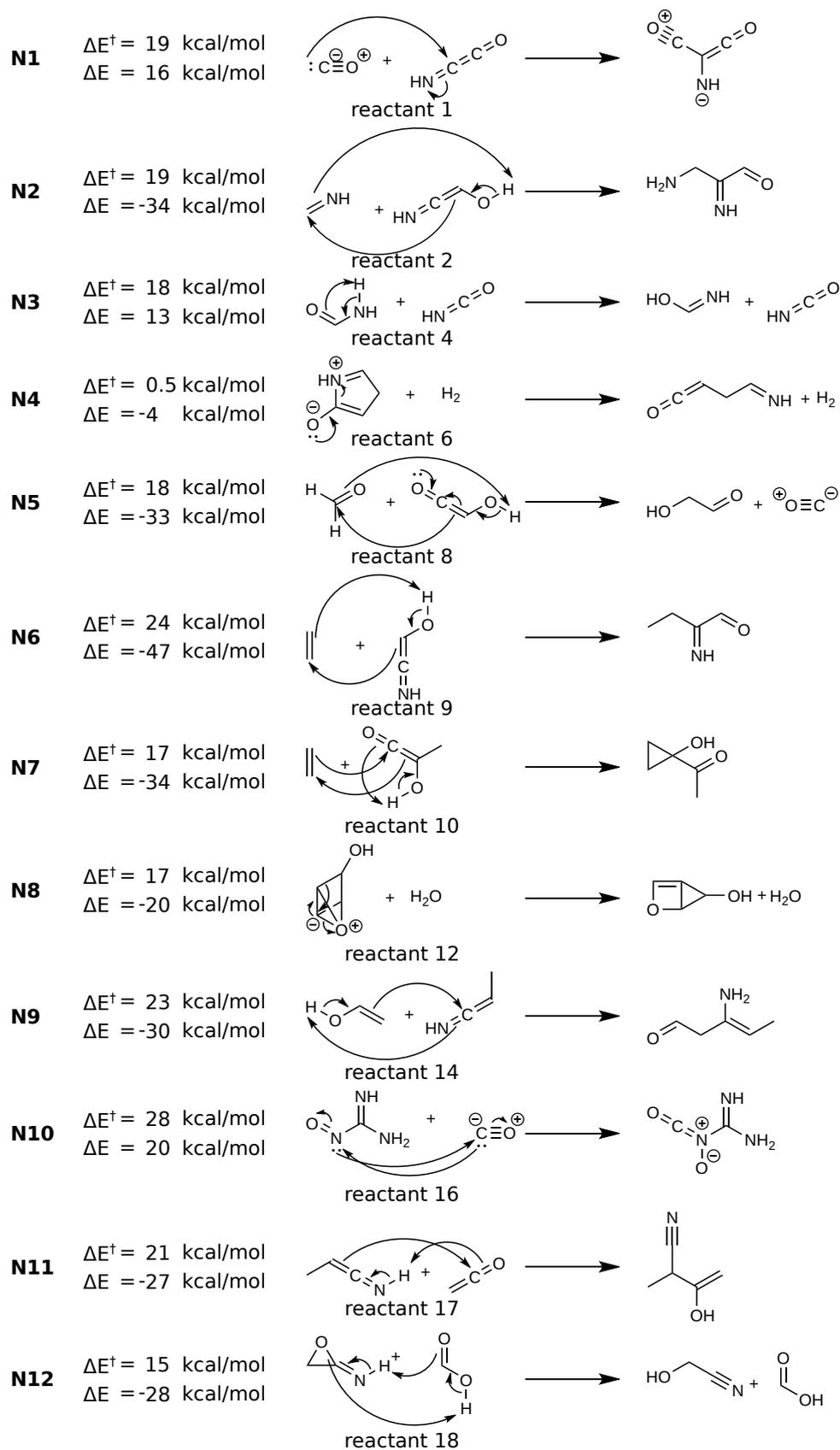


Figure 6: The lowest-barrier reaction per reactant for the 12 reactants, where new reactions below 30 kcal/mol were found.

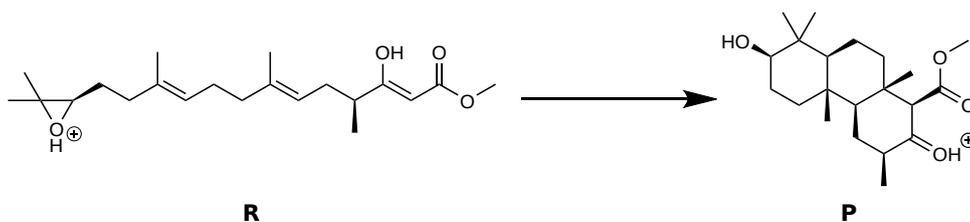


Figure 7: Step in the synthesis of Berkeleyone A [26]. This reaction was previously studied with imposed activation by Lavigne et al. [9]

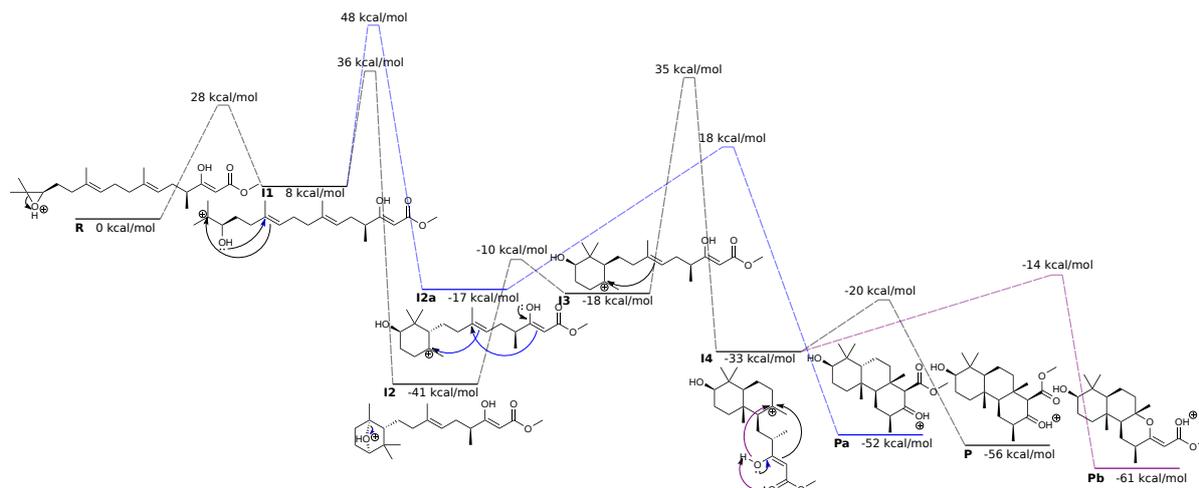


Figure 8: Some highlighted reaction paths found at GFN2-xTB level of theory. Energies are relative to the reactant (**R**)

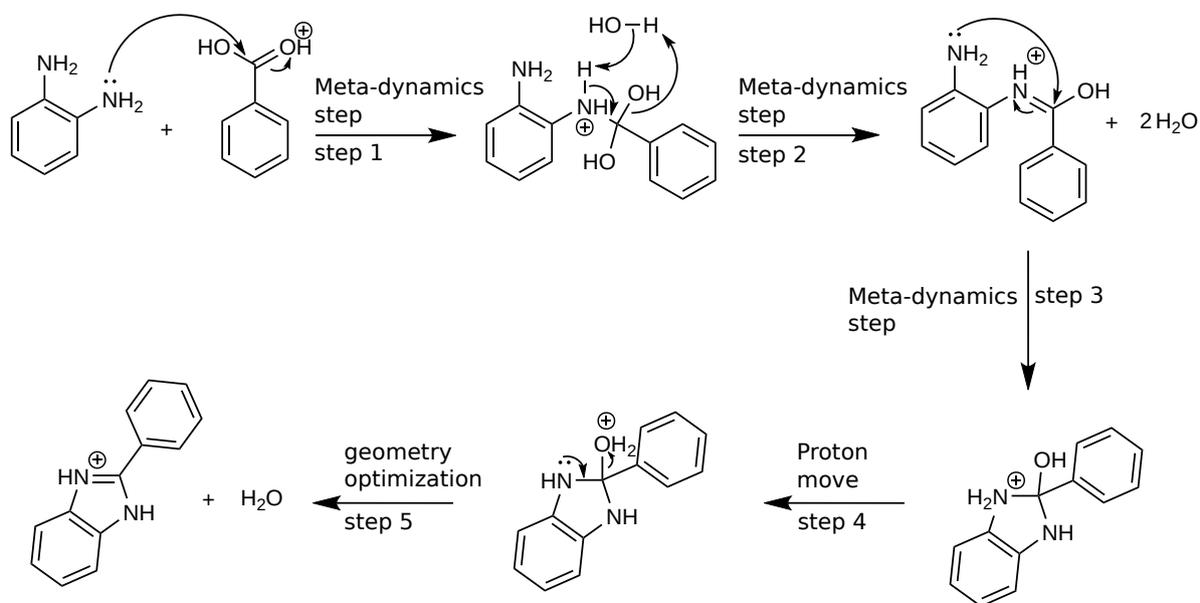


Figure 9: Summary of the reaction path found using meta-MD to follow the acid-catalyzed synthesis from ortho-phenylenediamine and benzoic acid to 2-phenyl benzimidazole

17 kcal/mol higher than at nitrogen). We follow the path from the carbonyl protonated structure and find step 1 by meta-MD (Figure 9, step 1) where the lone pair of one of the nitrogen atoms is used to attack the carbonyl carbon. The next step is elimination of water which can be found in two ways: either by manually transferring the proton to one of the hydroxyl groups, resulting in water eliminated upon energy minimization, or by adding a water molecule that can aid in the proton-transfer during the meta-MD simulation (Figure 9, step 2). Step 3 is a ring closure initiated by attack of the other nitrogen atom. The second water elimination can again be found by manually moving the proton from the ammonium to the hydroxyl group (step 4) and subsequent geometry optimization which eliminates the water molecule (step 5) creating the product 2-phenyl benzamidazole. Contrary to the first water elimination, this second water elimination was only found by manually moving the proton and not by the meta-MD runs with an additional water molecule (the backward reaction is the preferred path). The manual proton transfers can easily be automated.

3.4 Timings

The CPU-time requirements for the semiempirical calculations are relatively modest compared to the DFT refinement calculations. A single meta-MD simulation requires on average 5.4 minutes on a single core of a Intel Xeon E5-2643 v3 (3.4 GHz) for the reactants of the low-barrier reaction dataset. Larger molecules such as the Berkeleyone A precursor (Figure 7) require about 25 minutes, but the precise value depends greatly on how fast the reaction occurs; for the steps presented in Figure 8 the average run time for a meta-MD simulation was in the range 17-35 minutes. A single semiempirical barrier estimate typically takes about 14 seconds on the same type of core (13 minutes for the Berkeleyone A precursor) and we usually run five of these in parallel with different settings. For comparison, a typical ω B97X-D/def2-TZVP TS search takes about 6.5 hours on 2 cores.

4 Conclusions

We test our meta-MD based approach for finding low-barrier (<30 kcal/mol) reactions for uni- and bimolecular reactions extracted from the barrier dataset developed by Grambow et al.[18] Based on this dataset it should be possible to locate 26 low-barrier unimolecular reactions at the ω B97X-D/def2-TZVP level of theory starting from 163 reactants. Our method uses Grimme’s meta-MD approach, with carefully chosen hyperparameters, to identify possible products, which are subsequently screened using semiempirical reaction energies and barrier heights before being refined with DFT (Figure 1). The meta-MD simulations identify 25 of the 26 products found by Grambow et al., while the subsequent semiempirical screening eliminates an additional four reactions due to an overestimation of the reaction energies or estimated barrier heights relative to DFT, suggesting that DFT may thus be needed in the screening process. In addition, our approach identifies an additional 36 reactions not found by Grambow et al., 10 of which have barriers <30 kcal/mol. All low-barrier reactions involve new reactants not represented in the low-barrier dataset, which indicates that the reactions in that dataset are likely the ones with the lowest possible barriers for each reactant.

Grambow et al. [18] only searched for elementary reactions starting from single reactant molecules, but they found a lot of products with two fragments. From these back-reactions, we extract a set of 20 target low-barrier reactions where two molecules react to create a single molecule (Table S4). While these reactions are not necessarily the ones with the lowest barrier for a given pair of reactant molecules, our method should be able to identify them along with any reactions with lower barriers. The meta-MD simulations identify 19 of the 20 products found by Grambow et al., while the subsequent semiempirical screening eliminates an additional reaction due to an overestimation of the barrier height relative to DFT. In addition, we find 34 new low-barrier reactions. We found that it is necessary to ”encourage” the reactants to go to previously undiscovered products, by including products found by other MD simulations when computing the biasing potential as well as decreasing the size of the molecular cavity in which the MD occurs, until a reaction is observed.

The reactions in the Grambow et al. data set involve relatively small molecules, often with functional groups not usually seen in organic chemistry. We thus test our methodology on two reactions, one unimolecular and one bimolecular, that are more representative of those encountered in synthetic organic chemistry, with the goal to simply check whether the correct products can be found with the current parameters. The unimolecular reaction is a multi-step triple ring-closure (Figure 7)- an important step of the synthesis of Berkelyone A reported by Elkin et al.[26] We locate a possible mechanism for the path to the observed product through the five steps, together with other known biproducts. The bimolecular reaction is the acid catalyzed syntheses of benzimidazole derivatives starting from ortho-phenylenediamine and benzoic acid (Phillips method)[29], where we find all five steps in the generally accepted reaction mechanism. The meta-MD hyperparameters used in this study thus appears to be generally applicable to finding low-barrier reactions.

5 Acknowledgments

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

6 Supporting information

Additional tables and figures can be found in supporting information. The code can be found here <https://github.com/jensengroup/AutomatedReactionsMetaMD> and additional data can be found here <https://sid.erda.dk/sharelink/ewT1RzFbIT>

References

- [1] Paul M Zimmerman. “Automated discovery of chemically reasonable elementary reaction steps”. en. In: *J. Comput. Chem.* 34.16 (June 2013), pp. 1385–1392.
- [2] Yury V Suleimanov and William H Green. “Automated Discovery of Elementary Chemical Reaction Steps Using Freezing String and Berny Optimization Methods”. en. In: *J. Chem. Theory Comput.* 11.9 (Sept. 2015), pp. 4248–4259.
- [3] Scott Habershon. “Automated Prediction of Catalytic Mechanism and Rate Law Using Graph-Based Reaction Path Sampling”. en. In: *J. Chem. Theory Comput.* 12.4 (Apr. 2016), pp. 1786–1798.
- [4] J A Varela, S A Vázquez, and E Martínez-Núñez. “An automated method to find reaction mechanisms and solve the kinetics in organometallic catalysis”. en. In: *Chem. Sci.* 8.5 (May 2017), pp. 3843–3851.
- [5] Yeonjoon Kim, Jin Woo Kim, Zeehyo Kim, and Woo Youn Kim. “Efficient prediction of reaction paths through molecular graph and reaction network analysis”. en. In: *Chem. Sci.* 9.4 (Jan. 2018), pp. 825–835.
- [6] Amanda L Dewyer, Alonso J Argüelles, and Paul M Zimmerman. “Methods for exploring reaction space in molecular systems”. In: *WIREs Comput Mol Sci* 8.2 (Mar. 2018), e1354.
- [7] Christopher Robertson and Scott Habershon. “Fast screening of homogeneous catalysis mechanisms using graph-driven searches and approximate quantum chemistry”. en. In: *Catal. Sci. Technol.* 9.22 (Nov. 2019), pp. 6357–6369.
- [8] Jan P Unsleber and Markus Reiher. “The Exploration of Chemical Reaction Networks”. en. In: *Annu. Rev. Phys. Chem.* 71 (Apr. 2020), pp. 121–142.
- [9] Cyrille Lavigne, Gabriel dos Passos Gomes, Robert Pollice, and Alan Aspuru-Guzik. “Automatic discovery of chemical reactions using imposed activation”. en. In: *ChemRxiv* (Nov. 2020).
- [10] Robin J Shannon, Emilio Martínez-Núñez, Dmitrii V Shalashilin, and David R Glowacki. “ChemDyME: Kinetically Steered, Automated Mechanism Generation through Combined Molecular Dynamics and Master Equation Calculations”. en. In: *J. Chem. Theory Comput.* 17.8 (Aug. 2021), pp. 4901–4912.
- [11] Mads Koerstz, Maria H Rasmussen, and Jan H Jensen. “Fast and automated identification of reactions with low barriers: the decomposition of 3-hydroperoxypropanal”. In: *SciPost Chem.* 1.1 (Oct. 2021).
- [12] Ruben Van de Vijver and Judit Zádor. “KinBot: Automated stationary point search on potential energy surfaces”. In: *Computer Physics Communications* 248 (Mar. 2020), p. 106947. DOI: 10.1016/j.cpc.2019.106947. URL: <https://doi.org/10.1016/j.cpc.2019.106947>.
- [13] Tom A Young, Joseph J Silcock, Alistair J Sterling, and Fernanda Duarte. “autodE: Automated Calculation of Reaction Energy Profiles- Application to Organic and Organometallic Reactions”. en. In: *Angew. Chem. Int. Ed Engl.* 60.8 (Feb. 2021), pp. 4266–4274.
- [14] Paul M Zimmerman. “Single-ended transition state finding with the growing string method”. en. In: *J. Comput. Chem.* 36.9 (Apr. 2015), pp. 601–611.
- [15] Satoshi Maeda, Tetsuya Taketsugu, and Keiji Morokuma. “Exploring transition state structures for intramolecular pathways by the artificial force induced reaction method”. en. In: *J. Comput. Chem.* 35.2 (Jan. 2014), pp. 166–173. DOI: 10.1002/jcc.23481. URL: <https://doi.org/10.1002/jcc.23481>.
- [16] Stefan Grimme. “Exploration of Chemical Compound, Conformer, and Reaction Space with Meta-Dynamics Simulations Based on Tight-Binding Quantum Chemical Calculations”. en. In: *J. Chem. Theory Comput.* 15.5 (May 2019), pp. 2847–2862. DOI: 10.1021/acs.jctc.9b00143. URL: <https://doi.org/10.1021/acs.jctc.9b00143>.

- [17] Christoph Bannwarth, Sebastian Ehlert, and Stefan Grimme. “GFN2-xTB-An Accurate and Broadly Parametrized Self-Consistent Tight-Binding Quantum Chemical Method with Multipole Electrostatics and Density-Dependent Dispersion Contributions”. en. In: *J. Chem. Theory Comput.* 15.3 (Mar. 2019), pp. 1652–1671. DOI: 10.1021/acs.jctc.8b01176. URL: <https://doi.org/10.1021/acs.jctc.8b01176>.
- [18] Colin A Grambow, Lagnajit Pattanaik, and William H Green. “Reactants, products, and transition states of elementary chemical reactions based on quantum chemistry”. en. In: *Sci Data* 7.1 (May 2020), p. 137.
- [19] Greg Landrum. *RDKit: Open-source cheminformatics*. 2020. URL: <http://www.rdkit.org>.
- [20] Maria H Rasmussen and Jan H Jensen. “Fast and automatic estimation of transition state structures using tight binding quantum chemical calculations”. en. In: *PeerJ Phy. Chem.* 2 (Sept. 2020), e15.
- [21] M. J. Frisch, G. W. Trucks, H. B. Schlegel, G. E. Scuseria, M. A. Robb, J. R. Cheeseman, G. Scalmani, V. Barone, G. A. Petersson, H. Nakatsuji, X. Li, M. Caricato, A. V. Marenich, J. Bloino, B. G. Janesko, R. Gomperts, B. Mennucci, H. P. Hratchian, J. V. Ortiz, A. F. Izmaylov, J. L. Sonnenberg, D. Williams-Young, F. Ding, F. Lipparini, F. Egidi, J. Goings, B. Peng, A. Petrone, T. Henderson, D. Ranasinghe, V. G. Zakrzewski, J. Gao, N. Rega, G. Zheng, W. Liang, M. Hada, M. Ehara, K. Toyota, R. Fukuda, J. Hasegawa, M. Ishida, T. Nakajima, Y. Honda, O. Kitao, H. Nakai, T. Vreven, K. Throssell, J. A. Montgomery Jr., J. E. Peralta, F. Ogliaro, M. J. Bearpark, J. J. Heyd, E. N. Brothers, K. N. Kudin, V. N. Staroverov, T. A. Keith, R. Kobayashi, J. Normand, K. Raghavachari, A. P. Rendell, J. C. Burant, S. S. Iyengar, J. Tomasi, M. Cossi, J. M. Millam, M. Klene, C. Adamo, R. Cammi, J. W. Ochterski, R. L. Martin, K. Morokuma, O. Farkas, J. B. Foresman, and D. J. Fox. *Gaussian 16 Revision A.03*. Gaussian Inc. Wallingford CT. 2016.
- [22] Jan H. Jensen. *xyz2mol*. <https://github.com/jensengroup/xyz2mol>. 2021.
- [23] Lars Ruddigkeit, Ruud van Deursen, Lorenz C Blum, and Jean-Louis Reymond. “Enumeration of 166 billion organic small molecules in the chemical universe database GDB-17”. en. In: *J. Chem. Inf. Model.* 52.11 (Nov. 2012), pp. 2864–2875.
- [24] Arkajyoti Sengupta and Krishnan Raghavachari. “Solving the Density Functional Conundrum: Elimination of Systematic Errors To Derive Accurate Reaction Enthalpies of Complex Organic Reactions”. en. In: *Org. Lett.* 19.10 (May 2017), pp. 2576–2579.
- [25] Jimmy C Kromann, Alexander Welford, Anders S Christensen, and Jan H Jensen. “Random versus Systematic Errors in Reaction Enthalpies Computed Using Semiempirical and Minimal Basis Set Methods”. en. In: *ACS Omega* 3.4 (Apr. 2018), pp. 4372–4377.
- [26] Masha Elkin, Suzanne M Szewczyk, Anthony C Scruse, and Timothy R Newhouse. “Total Synthesis of (±)-Berkeleyone A”. en. In: *J. Am. Chem. Soc.* 139.5 (Feb. 2017), pp. 1790–1793.
- [27] Varinder K Aggarwal, Paul A Bethel, and Robert Giles. “A formal synthesis of (+)-pyripyropene A using a biomimetic epoxy-olefin cyclisation”. en. In: *Chem. Commun.* 4 (Jan. 1999), pp. 325–326.
- [28] Varinder K Aggarwal, Paul A Bethel, and Robert Giles. “A formal synthesis of (+)-pyripyropene A using a biomimetic epoxy-olefin cyclisation: effect of epoxy alcohol/ether on cyclisation efficiency”. en. In: *J. Chem. Soc. Perkin 1* 22 (Jan. 1999), pp. 3315–3321.
- [29] Montague Alexandra Phillips. “CCCXVII.—The formation of 2-substituted benzimidazoles”. en. In: *J. Chem. Soc.* 0 (Jan. 1928), pp. 2393–2399.

Supporting Information

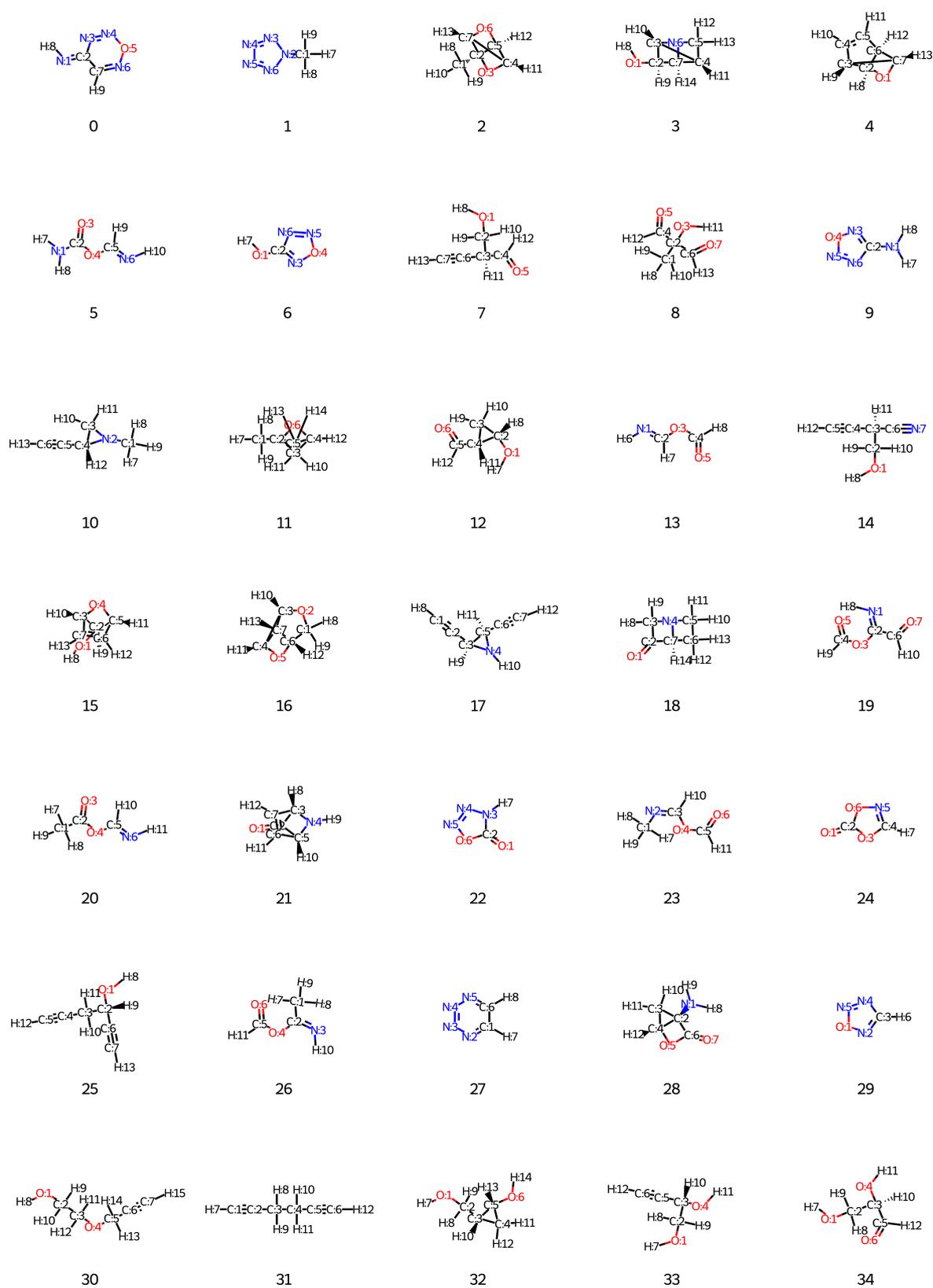


Figure S1: 163 reactants tested - part 1

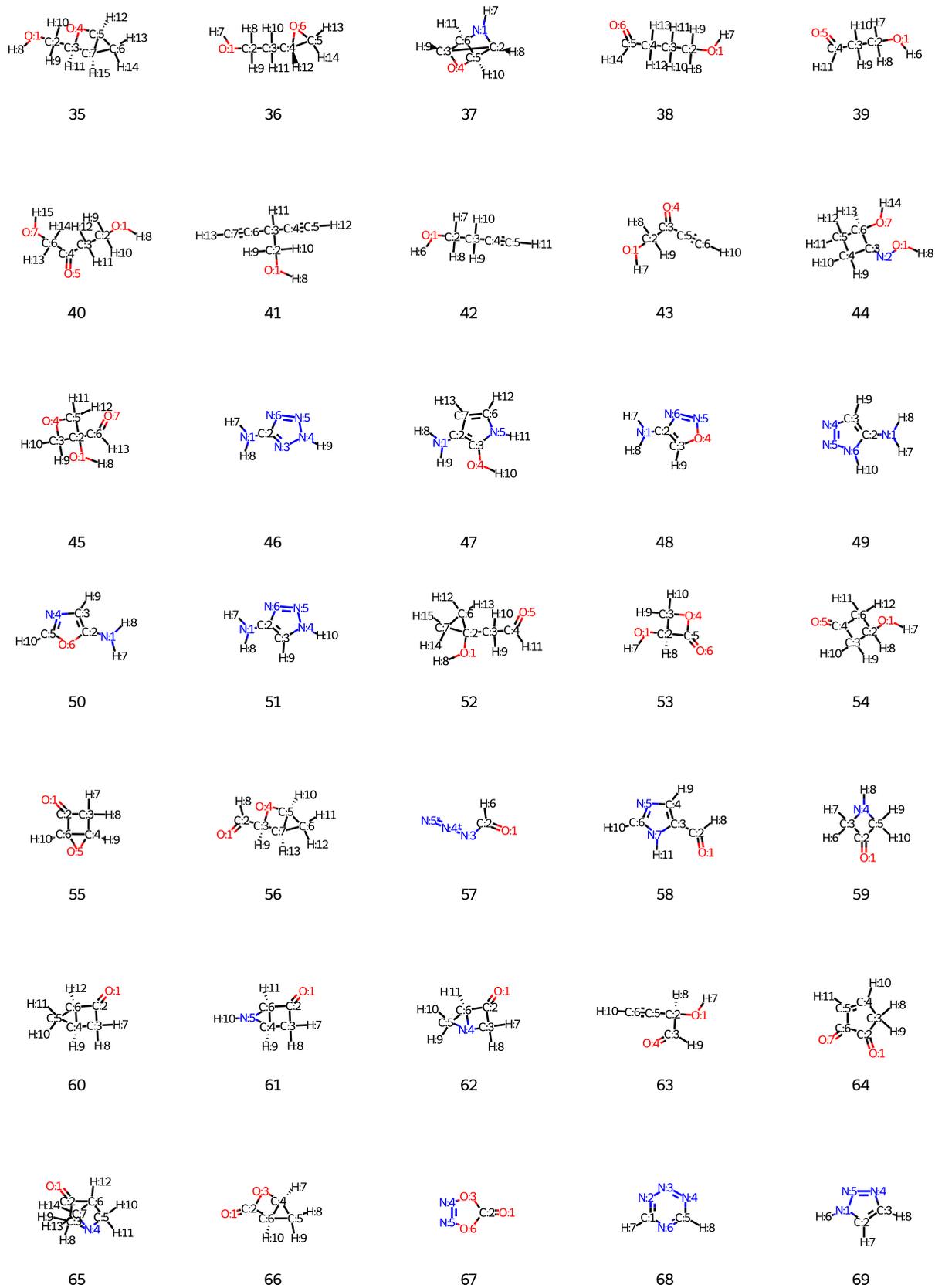


Figure S1: 163 reactants tested - part 2

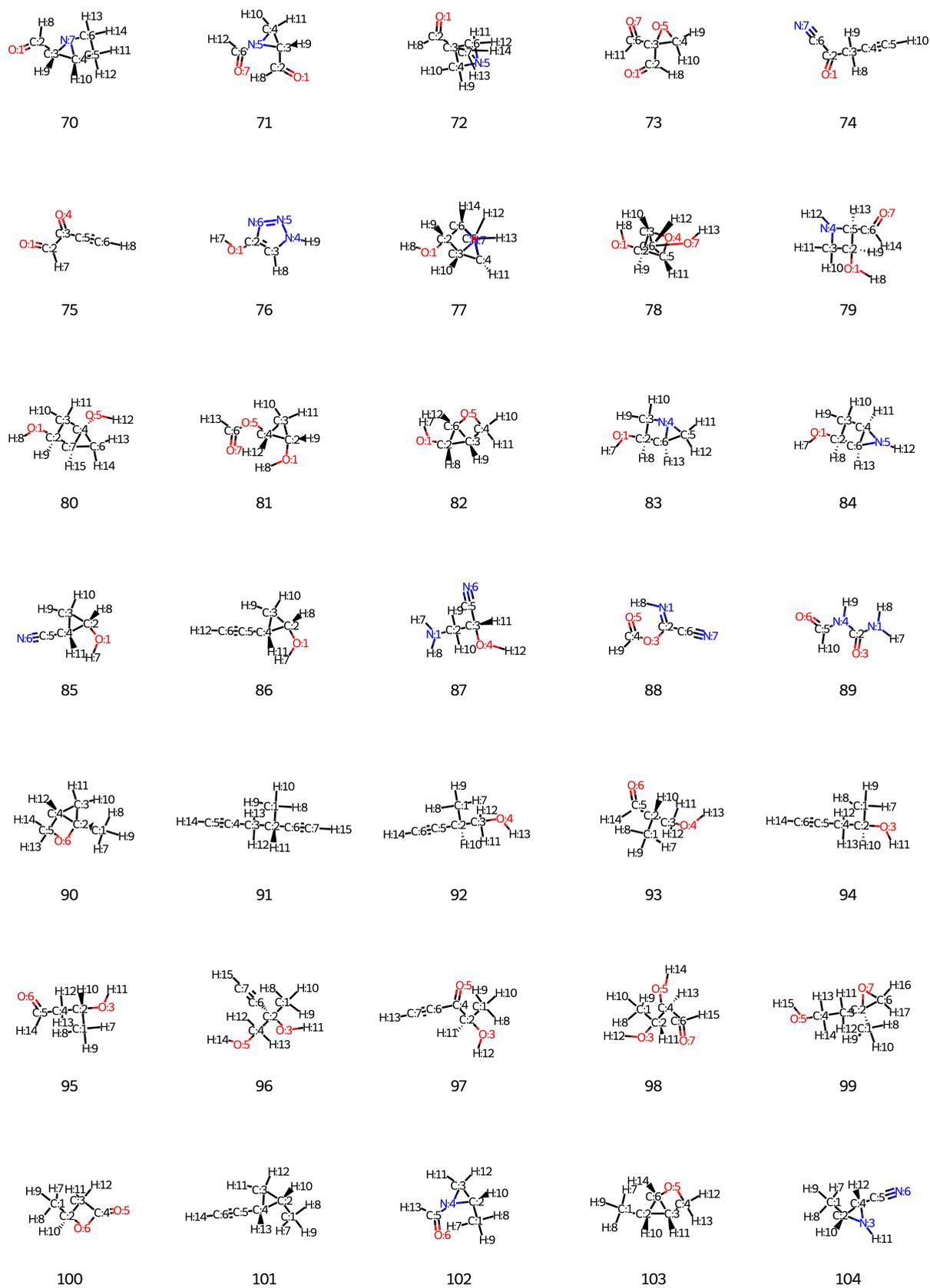


Figure S1: 163 reactants tested - part 3

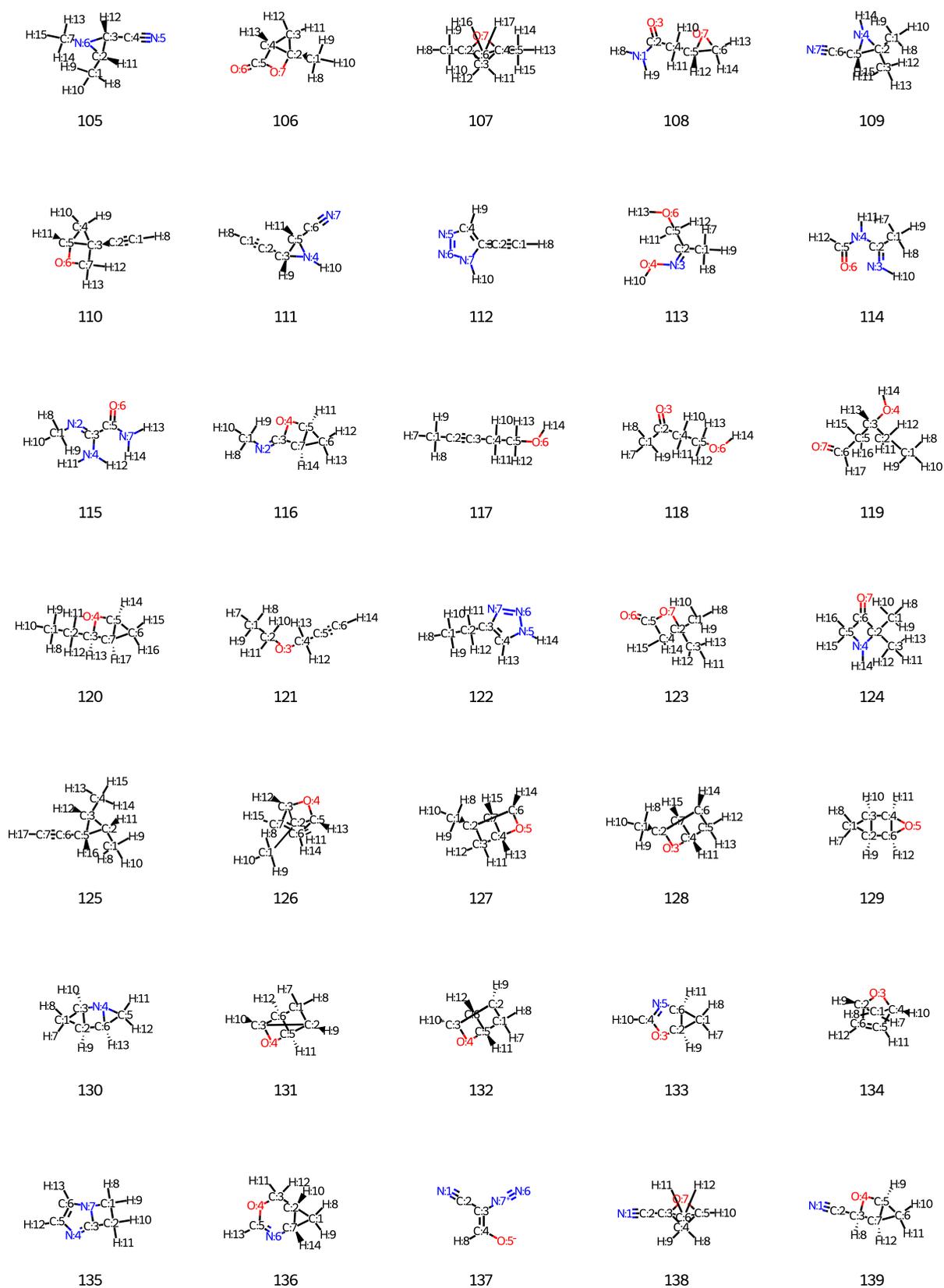


Figure S1: 163 reactants tested - part 4

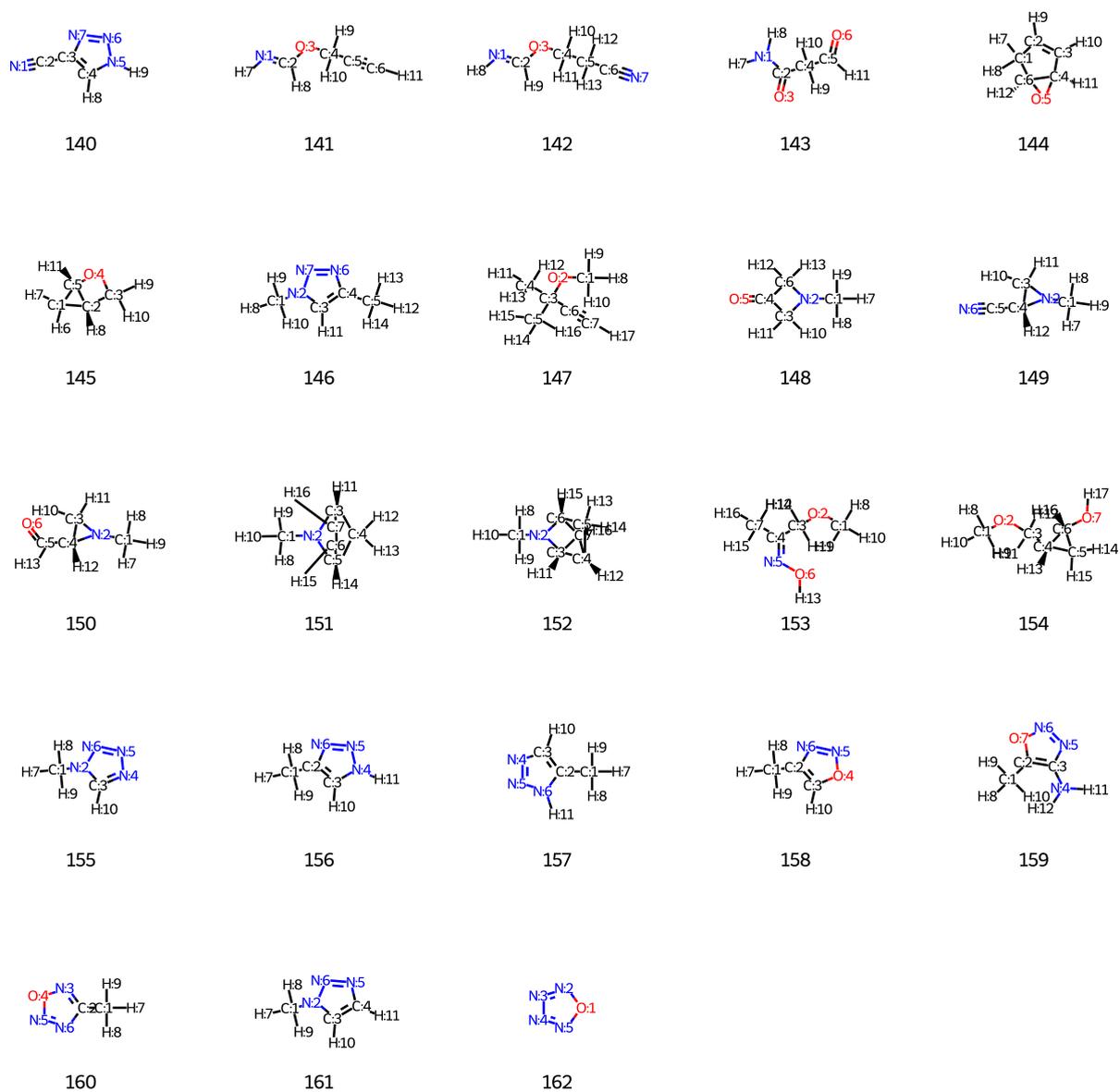
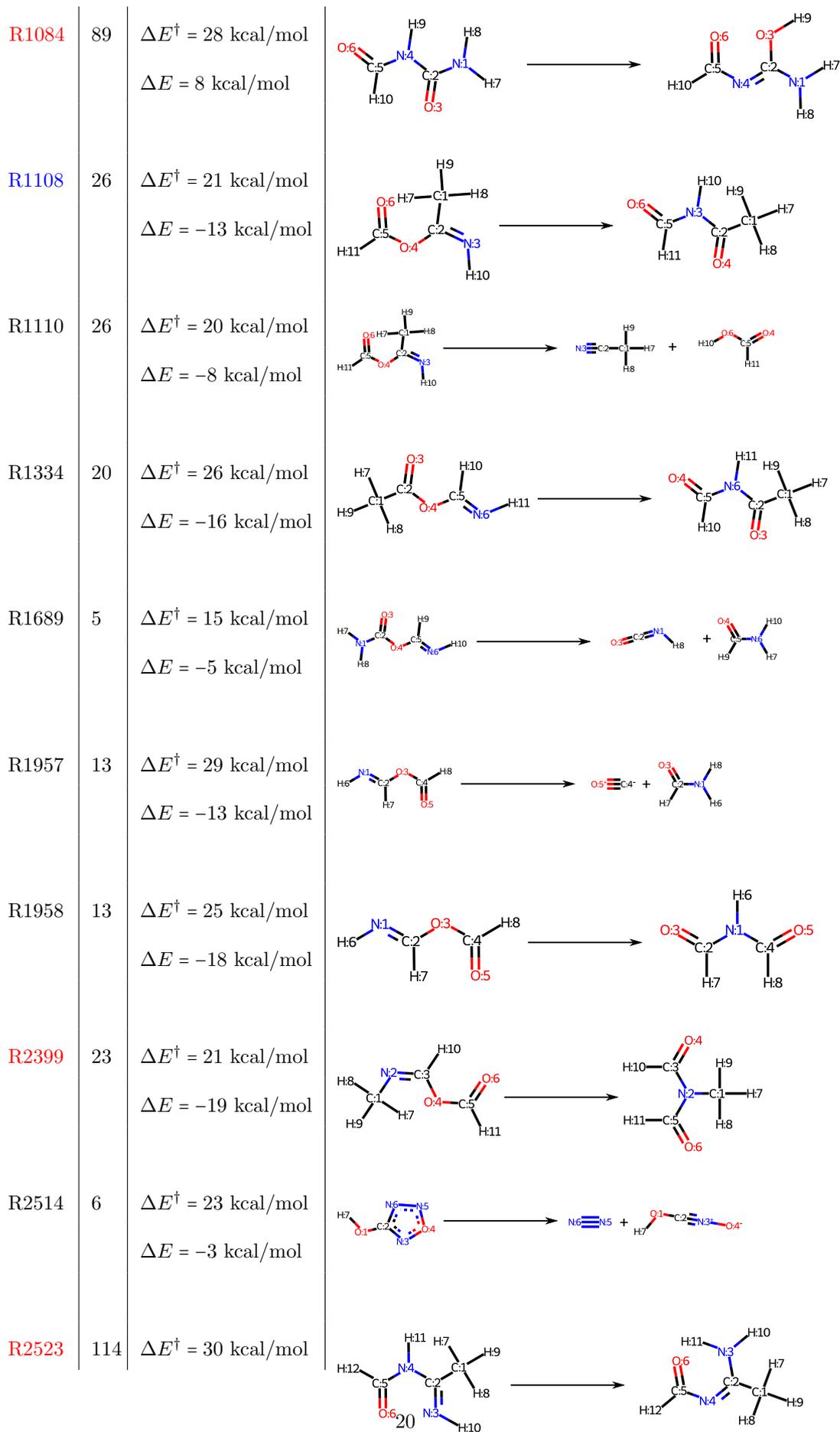
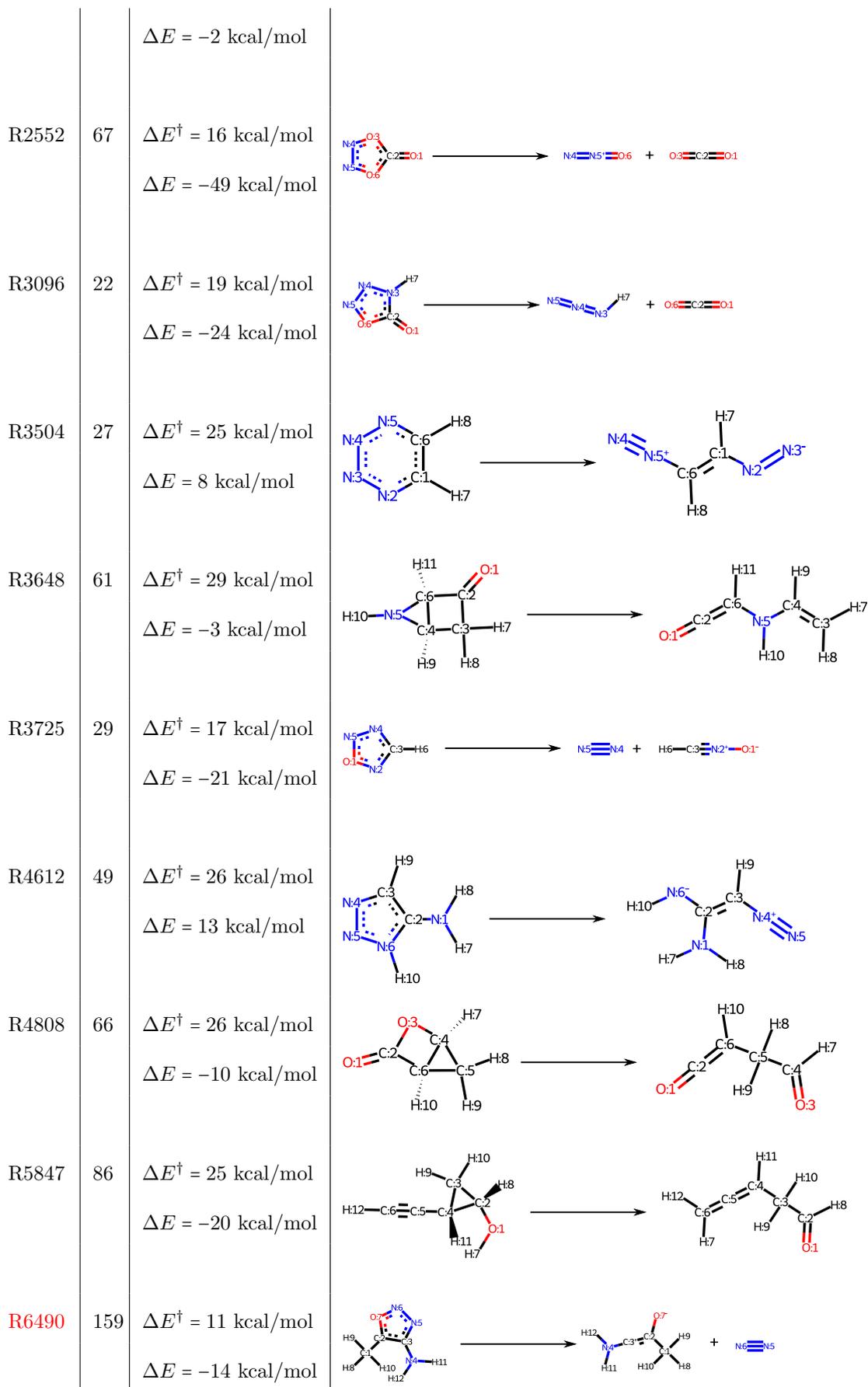
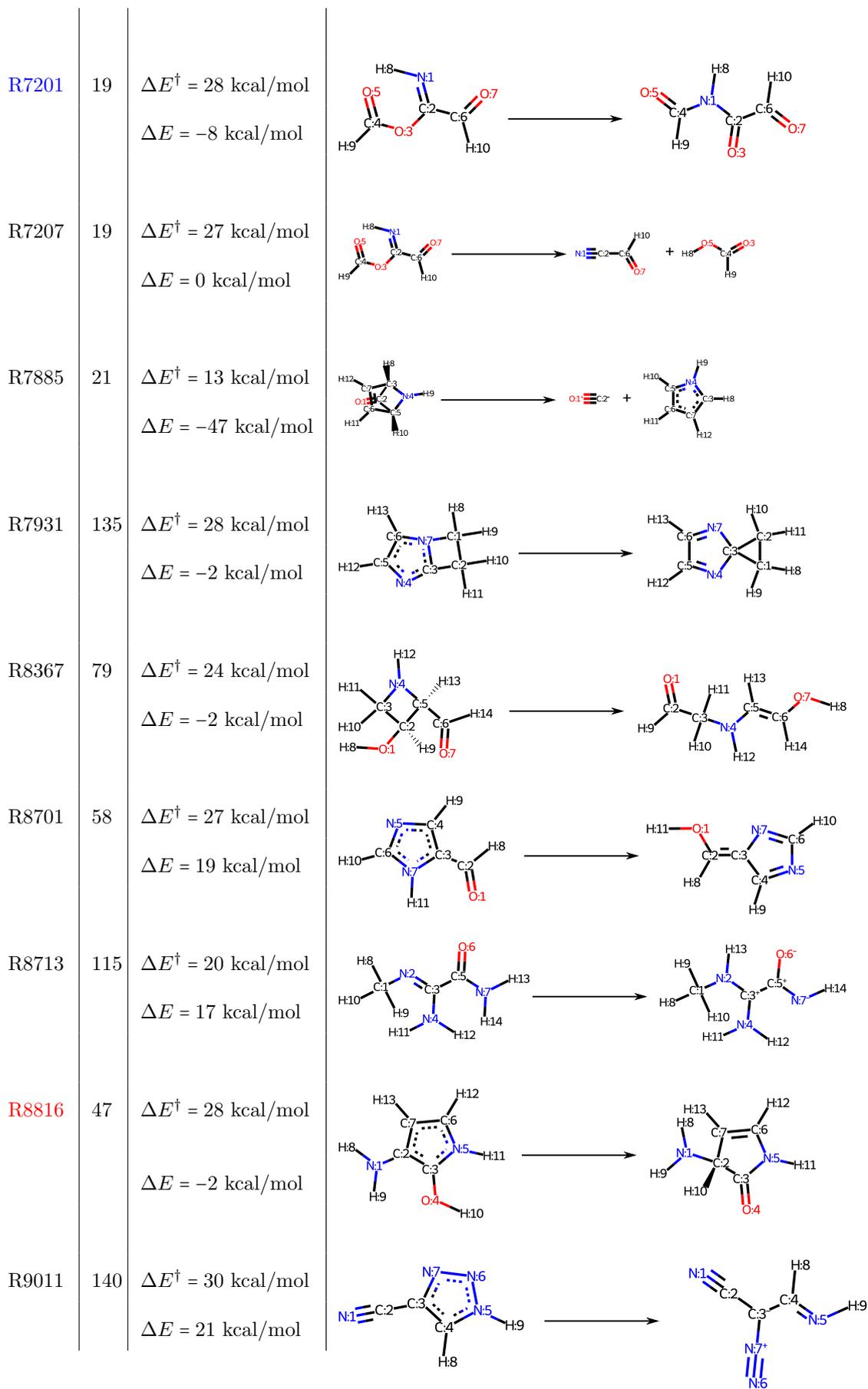


Figure S1: 163 reactants tested - part 5







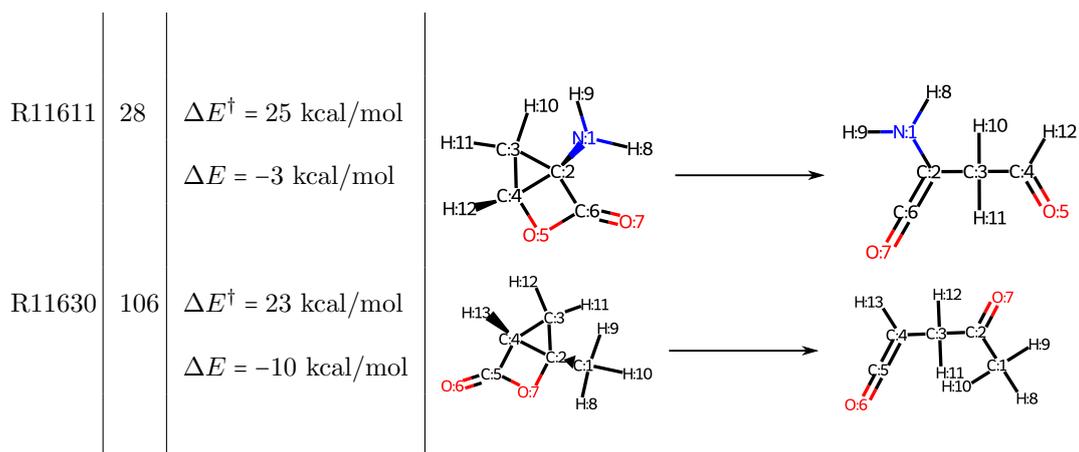
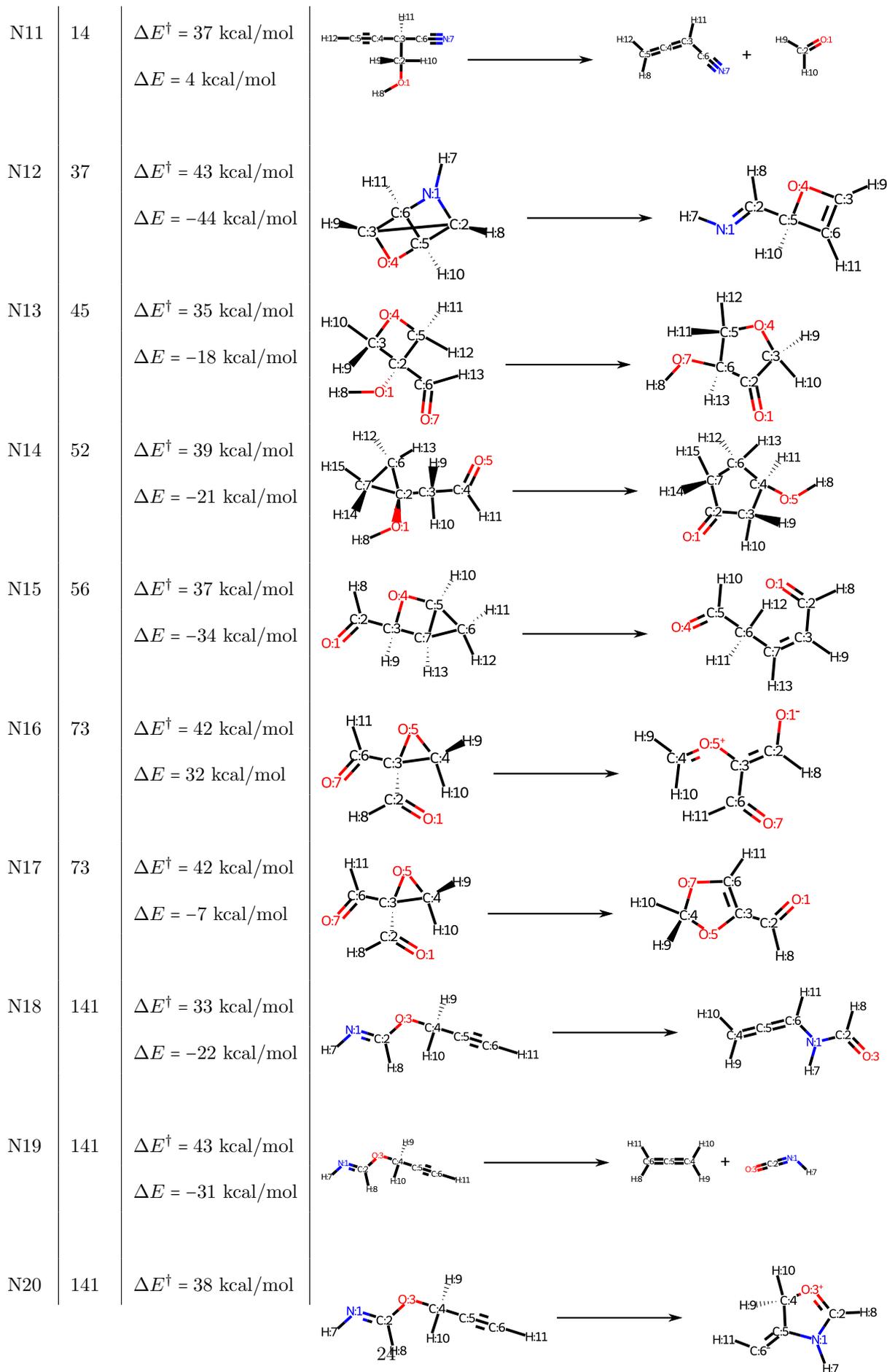


Table S1: 30 target reactions with one-fragment reactants and barriers below 30 kcal/mol
Reactions colored in red: IRC following re-optimization of the TS led to different reactant/product pair than the one stated.

Reactions colored in blue: IRC following re-optimization of the TS led to a different stereoisomer of the reactant (rotation around C=N bond).



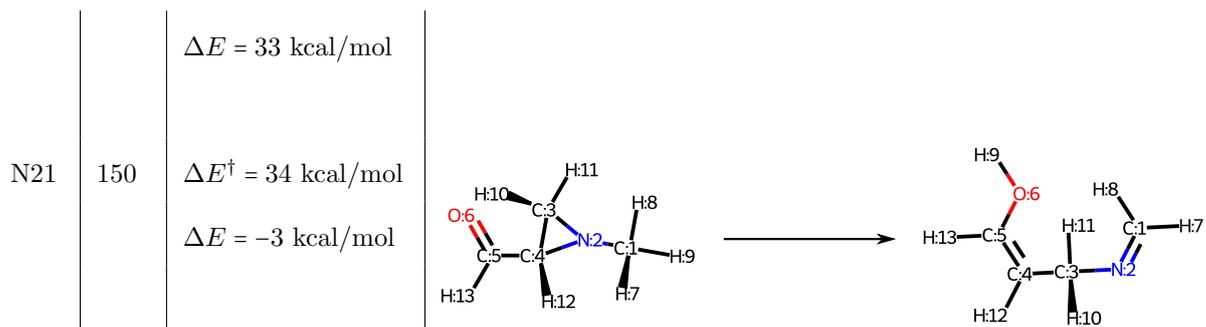
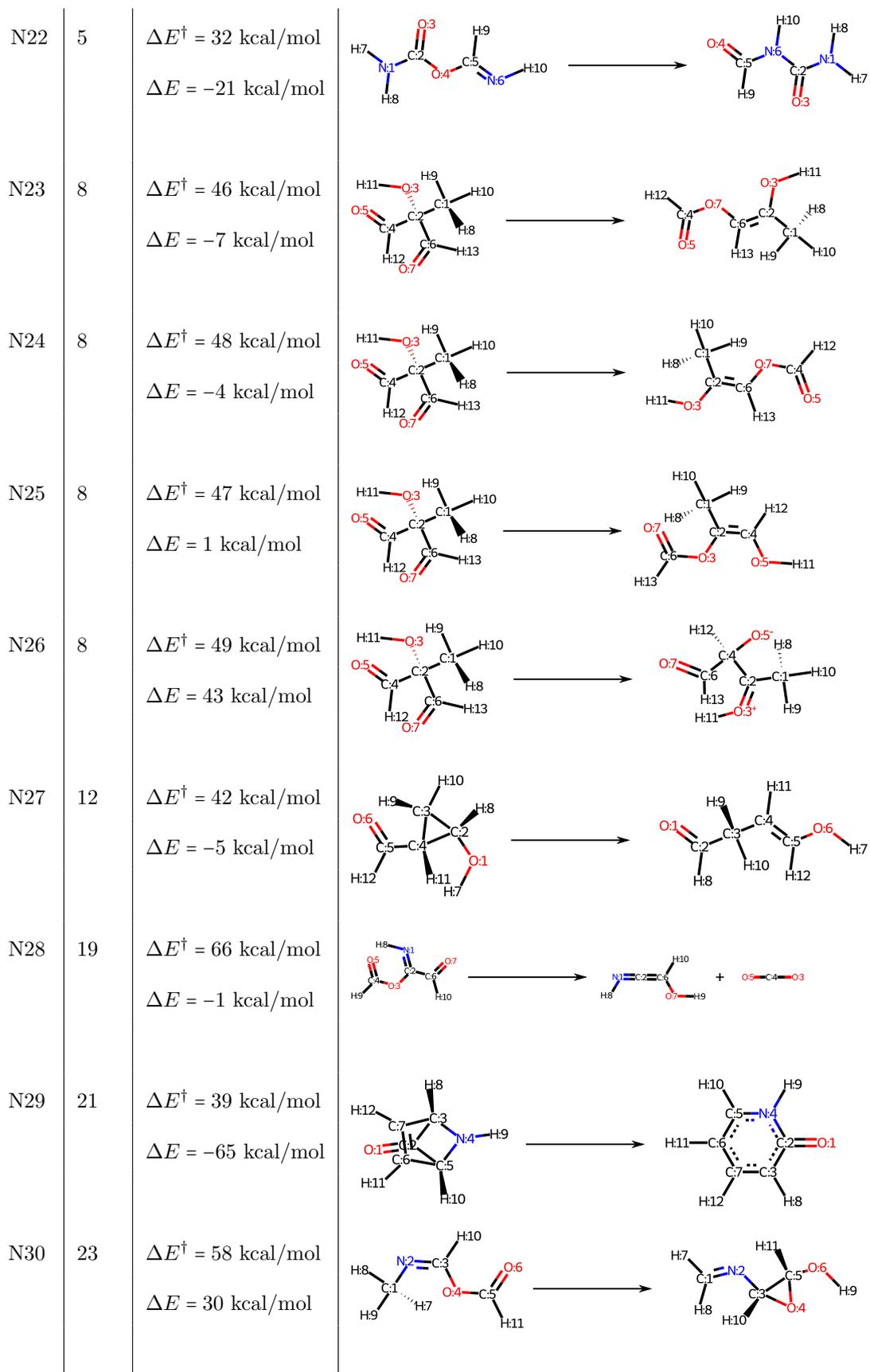
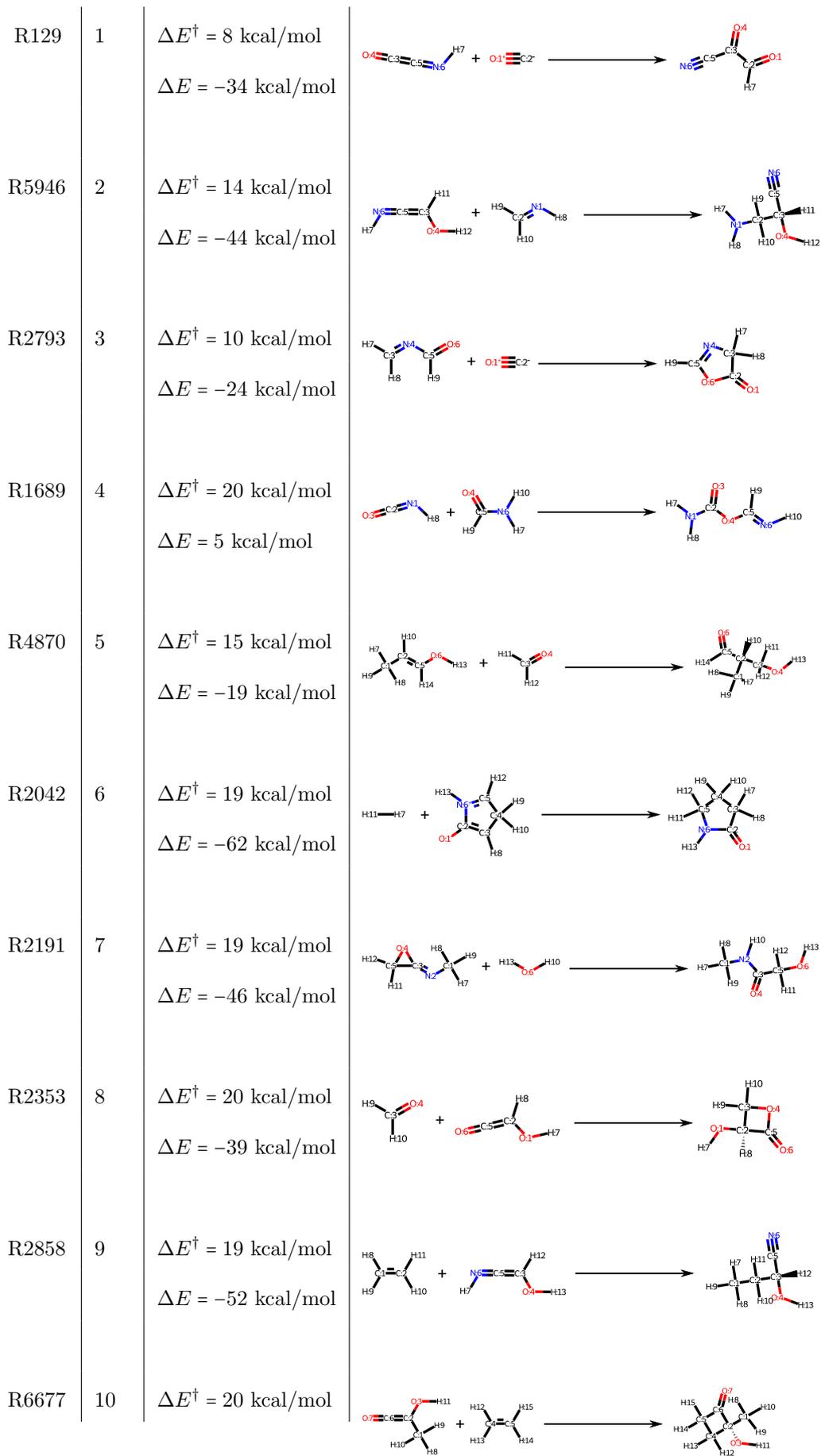


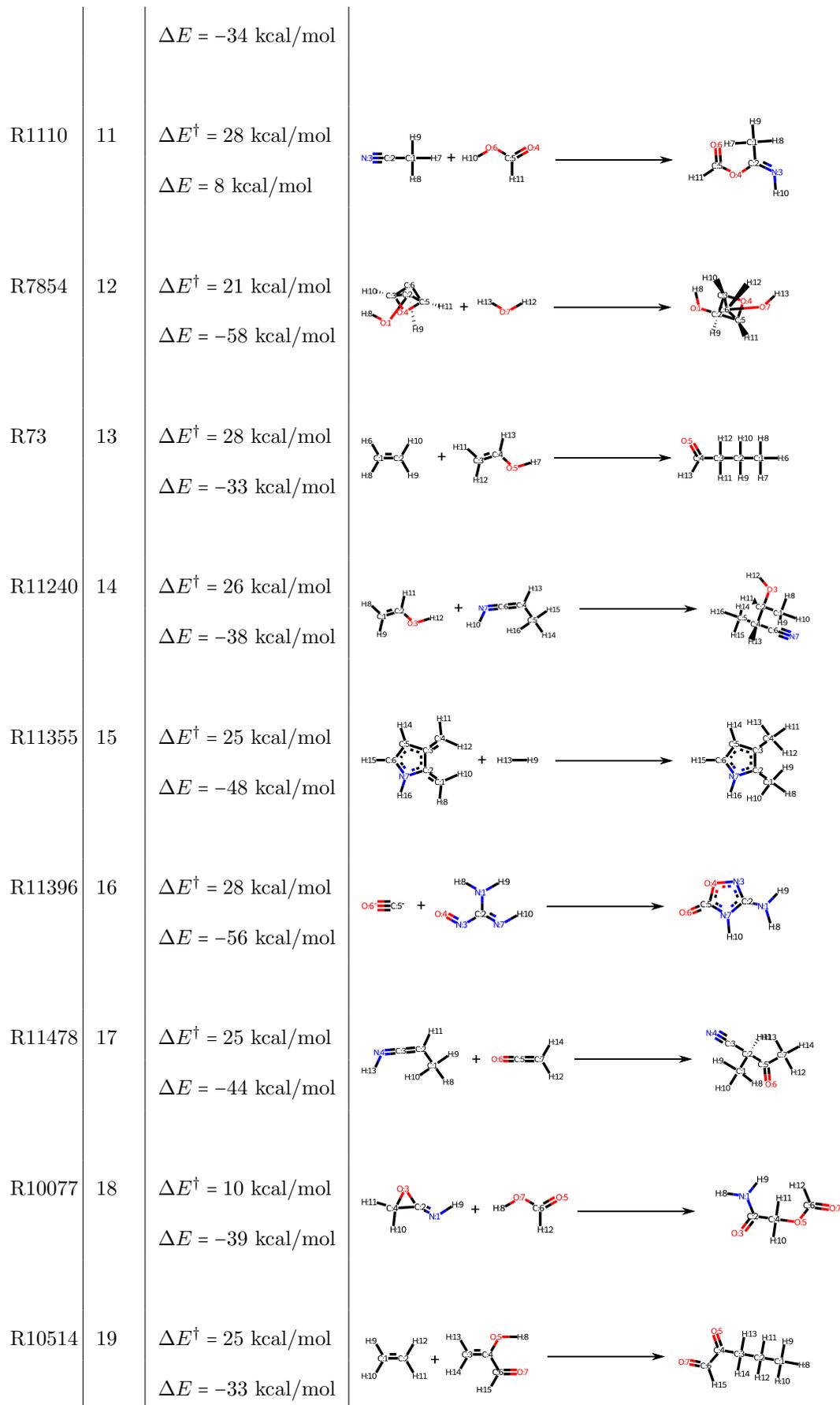
Table S2: 11 new reactions found with barriers above 30 kcal/mol - lower than existing reactions



N31	34	$\Delta E^\ddagger = 40$ kcal/mol $\Delta E = 14$ kcal/mol	
N32	63	$\Delta E^\ddagger = 47$ kcal/mol $\Delta E = 0$ kcal/mol	
N33	71	$\Delta E^\ddagger = 53$ kcal/mol $\Delta E = -21$ kcal/mol	
N34	79	$\Delta E^\ddagger = 36$ kcal/mol $\Delta E = -3$ kcal/mol	
N35	88	$\Delta E^\ddagger = 63$ kcal/mol $\Delta E = 13$ kcal/mol	
N36	115	$\Delta E^\ddagger = 37$ kcal/mol $\Delta E = 13$ kcal/mol	

Table S3: 15 new reactions found with barriers over 30 kcal/mol. Note that Grambow et al reported a reaction to another diastereomer of the product. We include it here because we saw a significantly lower barrier than the reported (55 kcal/mol)





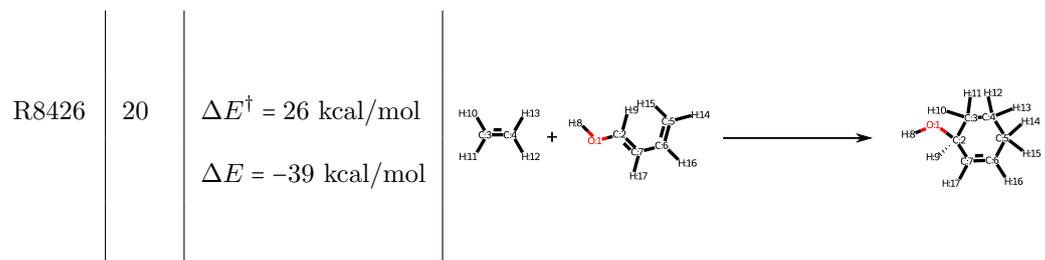


Table S4: 20 target reactions with two-fragment reactants and barriers below 30 kcal/mol

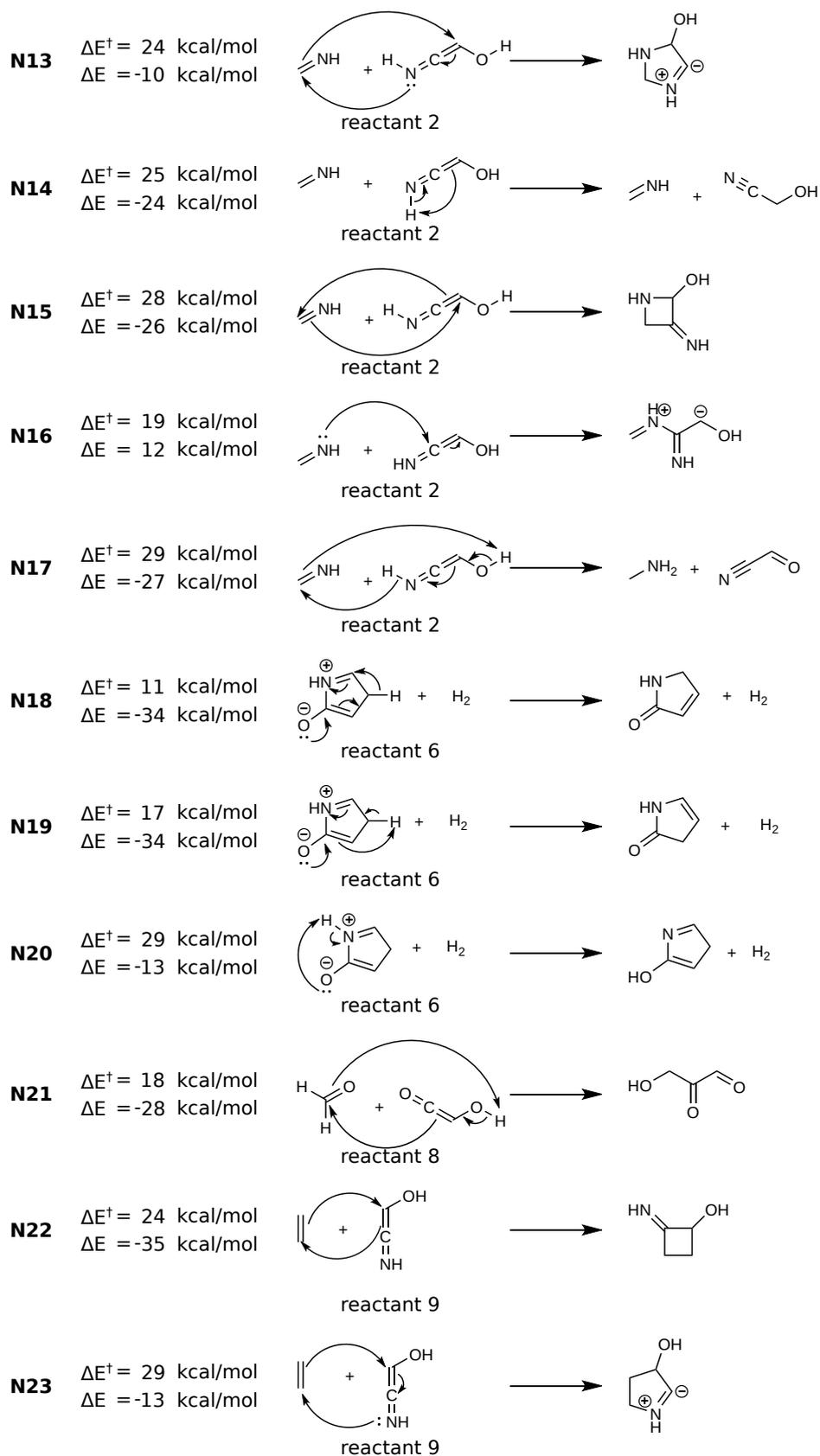


Figure S2

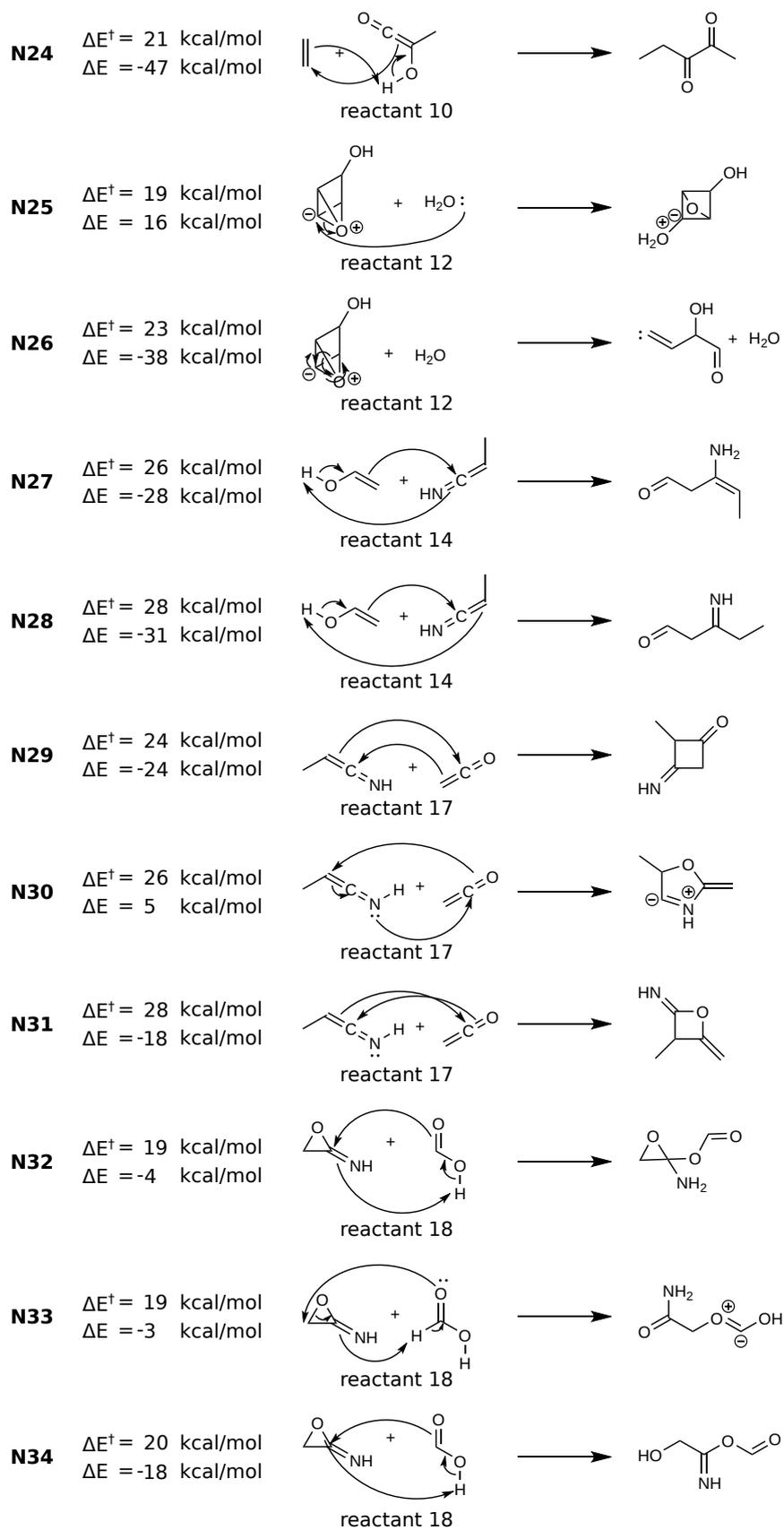


Figure S2: (continued)