Conformational Sampling for Transition State Searches on a Computational Budget

Qiyuan Zhao, Hsuan-Hao Hsu, and Brett M. Savoie*

Davidson School of Chemical Engineering, Purdue University, West Lafayette, IN, 47906

E-mail: bsavoie@purdue.edu

Abstract

Transition state searches are the basis for characterizing reaction mechanisms and activation energies, and are thus central to myriad chemical applications. Nevertheless, common search algorithms are sensitive to molecular conformation and the conformational space of even medium-sized reacting systems are too complex to explore with brute force. Here we show that it is possible to train a classifier to learn the features of conformers that conduce successful transition state searches, such that optimal conformers can be down-selected before incurring the cost of a high-level transition state search. To this end, we have benchmarked the use of a modern conformational generation algorithm with our reaction prediction methodology, Yet Another Reaction Program (YARP), for reaction prediction tasks. We demonstrate that neglecting conformer contributions leads to qualitatively incorrect activation energy estimations for a broad range of reactions, whereas a simple random forest classifier can be used to reliably down-select low-barrier conformers. We also compare the relative advantage of performing conformational sampling on reactant, product, and putative transition state geometries. The robust performance of this relatively simple machine learning classifier

mitigates cost as a factor when implementing conformational sampling into contemporary reaction prediction workflows.

1 Introduction

Computational transition state (TS) characterizations are a standard tool for differentiating between competing reaction mechanisms and predicting reaction kinetics, making these calculations essential in manifold applications.[1–7] Nevertheless, searching for the transition states of large systems and complex reactions is still a fragile process that is heavily dependent on user-expertise to guide search algorithms toward low barrier crossings (e.g., by initializing the search geometry based on an anticipated mechanism). For this reason, recent methods development efforts have been focused on automating the convergence of transition state searches, thus eliminating potential biases from user interventions and democratizing the availability of these calculations.[4,5,7,8] Among the challenges to full automation, is that the characteristics of discovered transition states, including the barrier height and identity of intermediates, can be profoundly affected by the conformational details of the reaction configuration. Currently, there is no inexpensive means of automatically in incorporating conformational sampling into transition state characterization workflows, which limits its adoption in high-throughput and reaction discovery applications.

Over the past several decades many algorithms have been developed to localize the transition states of chemical reactions. Among the most efficient are so-called doubled-ended search (DES) algorithms that make use of reactant and product information to locate the TS. Two of the most commonly used DES algorithms are the nudged elastic band (NEB) and string methods. NEB methods pre-define the discrete points on the reaction path as images and optimize the images towards the minimum energy pathway (MEP),[9,10] while string methods, including the growing string method (GSM)[11,12] and the freezing string method (FSM),[13] add new images after each

optimization step. All of these methods are pseudo one-dimensional searches to locate TSs under the geometric constraints of the reactants and products, which leads to a lower computational cost and wider application range compared with single-ended searching (SES) methods. In this sense, DES algorithms accelerate transition state localization by using the mutual information of the starting (reactant) and ending points (products).

Despite the speed-up enabled by DES algorithms, the details of a discovered transition state are strongly impacted by the conformation that is used to set up the transition state search. For example, misaligned reactant and product structures can stymie TS convergence.[12,14] Additionally, even when a TS converges, the barrier height and whether the TS corresponds to the intended reaction (i.e., saddle points in the potential energy surface that correspond to the putative reaction) strongly depend on the reaction conformation.[7] For a DES method, the possible conformational space of input structures is defined by the direct product of the reactant conformational space and the product conformational space. In this sense, M reactant conformers and N product conformers will lead to $M \times N$ possible conformational inputs to a DES method. Notably, the cost of sampling conformers for reasonably-sized molecules is small (approximately linear in the number of dihedrals and with a small prefactor due to the semi-empirical nature of available potentials) when compared with the computational cost of a transition state search at chemically-accurate levels of theory.[15,16] However, there is currently no means of down-selecting physically relevant conformers prior to performing a DES, which makes it common practice to neglect conformational sampling,[4,17–19] or to only perform partial sampling of one or more discovered TSs.[16,20] The prospect of selecting optimal reaction conformers in advance is complicated by the fact that the energetically favored reactant and/or product conformer may not be germane to finding the lowest-energy barrier of the intended reaction (e.g., consider the relatively unfavorable cis diene conformation that is necessary for a Diels-Alder reaction). Thus, the cost of conformational sampling in a transition state characterization workflow comes from the fact that putative reaction

conformers must be subjected to expensive TS searches to determine their relevance, and not from the conformer generation step itself.

The rationale for the current work is that if an indicator function could be developed that ranked conformers prior to performing a TS characterization, then the cost of incorporating conformational sampling into an automated DES work-flow would be decimated. This strategy is motivated by the hypotheses that (i) although the conformational search space is formally $M \times N$, the number of reaction conformations leading to non-generate transition states is much smaller in practice, (ii) neglecting conformational sampling leads to errors that are consistently many times kT and this confounds quantitative kinetic work, and (iii) since human intuition can often facilitate choosing a "good" conformer, the underlying fitness function is also amenable to being learned by a sufficiently flexible machine learning approach. Here these hypotheses are interrogated by training the proposed indicator function and testing its performance on four reaction prediction benchmarks. These benchmarks are taken from several application areas and cover both simple and complex organic reactions, including γ-ketohydroperoxide decomposition, Ireland–Claisen rearrangement, competing Diels-Alder reactions of a large ketothioester, and tetrapeptide cyclization. For all of these benchmarks, the trade-off between the completeness of conformational sampling and computational cost represents a major challenge for achieving automated and efficient localization of intended transition states. To perform these benchmarks, a modern conformational sampling algorithm was combined with our reaction prediction methodology, Yet Another Reaction Program (YARP),[7] to generate training data for developing and testing the transferability of the indicator function. The outcomes of these benchmarks demonstrate that conformational down-selection can be effectively performed across several reaction domains, thus representing a practical means of economically introducing conformational sampling into automated reaction prediction workflows.

2 Methods

The reaction characterization performed in each benchmark consisted of three components: conformer generation, conformer classification, and transition state characterization (Fig. 1). Each component is described in the subsequent sections. In brief, conformational sampling of the reactant and product species were performed first, followed by the generation of conformationally aligned reactant-product pairs. After this step, inexpensive geometry-based indicators for each reactant-product alignment were calculated that served as input features for a random forest (RF)[21] classification model. Finally, the conformations were ranked by the RF model based on their probability of yielding an intended reaction channel, then down-selected for transition state characterization using YARP.

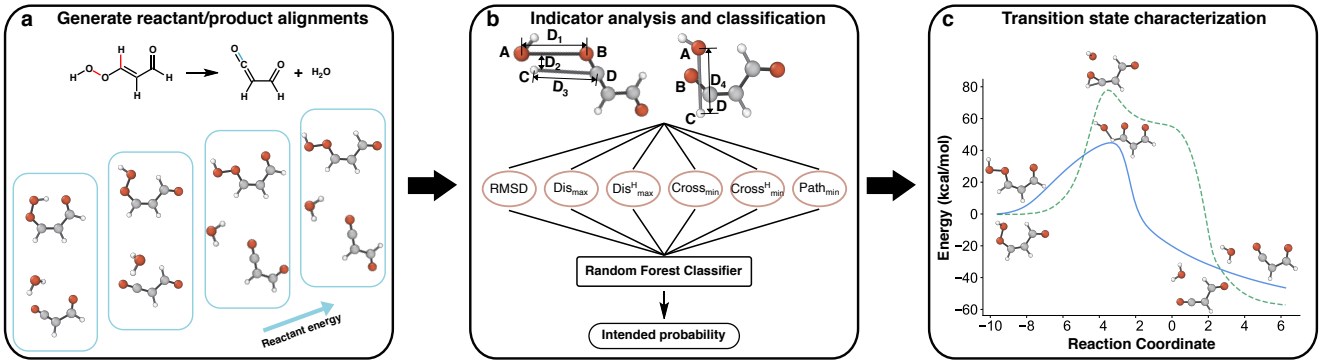

Figure 1: Overview of the presented approach for incorporating conformational sampling into an automated transition state search. (a) Conformational sampling is performed on the reactants and/or products to yield M and N conformers, respectively. These conformers are then used to generate conformationally aligned reactant-product geometries (N+M, total) that are candidates for double-ended transition state searches. (b) The random forest classification model is employed to rank reactant-product geometries based on inexpensive geometric features. An example of one conformationally aligned reactant-product pair is shown. (c) Transition states are characterized using growing string localization followed by Berny optimization and intrinsic reaction coordinate (IRC) calculations. An example is shown of a reaction with two qualitatively different transition states discovered from distinct reaction conformers.

2.1 Conformational Sampling of Reaction Geometries

The conformers of both the reactant and product influence a DES transition state search. Thus the space of potential reaction conformers (i.e., the unique pairs of reactant and product conformers) is the product of these two conformational spaces. Nevertheless, many such pairs are poorly conditioned for convergence since they may involve multiple dihedral rearrangements between reactants and products. Motivated by this, we recently developed an algorithm for generating conformationally-aligned reactant-product pairs, based on the conformer of either one or the other.[7] For instance, given a conformer of the reactant (product), this algorithm generates a conformationally-aligned product (reactant) geometry by performing a gradient decent optimization starting from the reactant (product) geometry but using a force-field potential and the bonding matrix of the product (reactant). Given the demonstrated robustness of this algorithm in earlier work, it is retained here for generating pairs of conformers after conformationally sampling the reactants or products. With this approach, the number of potential reaction conformations is reduced from $M \times N$ to $M + N$, where M and N are the number of conformers generated separately for the reactants and products, respectively.

Despite the reduction from $M \cdot N$ to $M + N$, there are still scenarios where conformational sampling is prohibitively costly with contemporary algorithms. For example, in an application like network exploration, there may be one set of reactants but hundreds of potential products. Likewise, many stereoisomers can result from the same set of reactants. In such cases, reactant-side conformational sampling can be several orders of magnitude less expensive than product-side sampling ($M \ll \Sigma_i N_i$, where i refers to each product) due to the many candidate products. Additionally, constrained conformational sampling can be performed on already converged transition states (TSs) to potentially discover more stable conformations .[8,16] Thus, conformational sampling of the reactant, product, and discovered TSs are all potentially relevant for finding the lowest bar-

rier corresponding to the intended reaction. Unless stated otherwise, reactant-side conformational sampling is performed in the benchmarks reported here to keep the computational costs tractable, but in two of the benchmarks the benefits of product-side sampling and TS conformational sampling are also compared.

All conformational sampling was performed using the CREST methodology,[16] which is a metadynamics-based algorithm for sampling the dihedral degrees of freedom (iMTD-GC algorithm) at the GFN2-xTB[22] level of theory. The thresholds for root-mean-squared displacement (RMSD) and energy used by CREST to determine distinct conformers were 0.125 Åand 0.05 kcal/mol, respectively. For multi-molecular reactants, conformers were discarded if they exhibited a centroid-centroid intermolecular separation greater than twice the sum of the molecular radii. After conformational sampling and reactant-product alignment, an RMSD minimized geometry was generated by rotation and center-of-mass translation to align the product with the reactant.[23] In our experience, RMSD minimization often helps, but it sometimes leads to geometries that fail to localize transition states compared with the merely conformationally-aligned geometries. Here, whichever geometry exhibited a higher rank based on the RF model (*vide infra*) was retained and the other was discarded.

2.2 Ranking Reaction Conformations

Table 1: Indicators for selecting reactant-product alignments

Indicator	Definition	Threshold
RMSD	Mass-weighted root mean square displacement.	1.8Å
$\mathrm{Dis}_{\mathrm{max}}$	Maximum displacement over all bonded heavy atom pairs.	4.8Å
$\mathrm{Dis}^{\mathrm{H}}_{\mathrm{max}}$	Maximum displacement over all bonded atom pairs involving at least one hydrogen.	/
$\mathrm{Cross}_{\mathrm{min}}$	Minimum separation between segments connecting bonded heavy atoms and persistent bonds.	0.03Å
$\mathrm{Cross}^{\mathrm{H}}_{\mathrm{min}}$	Minimum separation between segments connecting bonded atoms involving hydrogen and persistent bonds.	/
$\mathrm{Path}_{\mathrm{min}}$	Distance between the reactive heavy atom segments.	0.20Å

To rank reaction conformations, we developed a set of geometric features that would be in-

expensive to calculate while also being sufficiently informative to train an accurate classifier. In particular, we anticipated that physically-motivated features would enable us to train simpler ML models, like random forests, that can be applied with better transferability and reduced training data requirements than, say, a neural network.

In total six geometric features were developed (Table 1). First, the mass-weighted root mean square displacement (RMSD) between the reactant and product geometries was computed to represent the overall structural change. Second, the maximum separation over all pairs of bonded atoms was defined to indicate the likelihood of fragment roaming. For this feature, a pair of atoms qualifies as bonded if a bond exists in either the reactant or in the product, not necessarily both. The separations of bonded pairs are calculated in both the reactant and product geometries and the maximum over all of these separations comprises the feature. Since proton roaming is much more common and facile than other fragments roaming (e.g. methyl and hydroxy), this feature was calculated separately with respect to bonded atoms involving hydrogen ($\mathrm{Dis}^{\mathrm{H}}_{\mathrm{max}}$) and bonded atoms involving heavy atoms ($\mathrm{Dis}_{\mathrm{max}}$). For instance, $\mathrm{Dis}_{\mathrm{max}}$ of the reaction conformation shown in Figure 1b is the distance between oxygen atoms A and B in the product geometry (denoted as D_1 in Fig. 1b); whereas $\mathrm{Dis}^{\mathrm{H}}_{\mathrm{max}}$ is taken from the larger value of the distance between the hydrogen atom, C, and the carbon atom, D, in the product geometry (denoted as D_3) and the distance between oxygen atom, A, and hydrogen atom, C, in the reactant geometry (denoted as D_4). To indicate the likelihood of a steric clash during the transition state search, the minimum distance between segments connecting bonded atoms and segments connecting persistent bonds (i.e. unchanged bonds in the reaction) was calculated. If any atoms were shared between these two segments then the combination was omitted since it would trivially yield zero. Similar to the previous case, $\mathrm{Cross}_{\mathrm{min}}$ and $\mathrm{Cross}^{\mathrm{H}}_{\mathrm{min}}$ were both calculated to distinguish between potential clashes involving heavy atoms and those involving at least one hydrogen, respectively. An illustrative example of this feature is shown in Figure 1b where the distance between D_4 (i.e., the segment

defining the eventual bond between hydrogen C and oxygen A) and the persistent bond between atoms B and D is close to zero. All else being equal, this is unfavorable since it implies that the atoms will have to bypass this chemical bond to complete the reaction. Finally, to indicate the likelihood of a steric clash between the reacting atoms (i.e., atoms involved in a bond that is broken or formed in a reaction), the minimum separations were calculated between the segments connecting each pair of heavy atoms that form a new bond during the reaction. The minimum separation over all such segments calculated in the reactant and product geometries comprises the feature, Path_{min}. Hydrogens were omitted from this feature, since we found that clashes for protons are only weak indicators for a failure to localize a TS. For example, the distance D_2 shown in Figure. 1b represents the minimum separation between the segments D_1 and D_3 that correspond to the bonds being formed during reaction. However, since bond segment D_3 involves a hydrogen, this separation is omitted from the calculation of Path_{min}. For any reactions where $\text{Cross}_{\text{min}}$ and Path_{min} could not be calculated because all reacting bonds involved hydrogen, an above average value of 2.0 Å was used to indicate a low probability of a steric clash.

Before training or applying the RF model, reaction geometries were discarded if they exhibited features in excess of any of the thresholds listed in Table 1. These threshold values were chosen to be very conservative such that only very poorly conditioned geometries are excluded by these criteria (e.g., less than 1% of structures are discarded in this fashion). Examples of poorly conditioned geometries that fail these criteria are discussed in the supporting information (Fig. S1).

Two RF models were trained that we will refer to as the "conformation-rich" and "conformation-poor" models. These models were trained on the same dataset, with the same indicators and hyper-parameters, but they are distinguished by using true (intended) and false (unintended) class weight ratios of 1:1.5 and 1.5:1, respectively, during training. The reweighting of the class labels increases the penalty for false-positives and false-negatives in the conformation-rich and conformation-poor models, respectively. The rationale for this is that the two types of errors become problematic for

different reasons in these data regimes and that both regimes can be encountered during conformational sampling. In the conformation-rich scenario, the surplus of conformers makes false-positives a concern. In contrast, being too conservative in excluding false negatives can lead to too few conformers being sampled in the conformation-poor scenario. The model that was applied in each case was determined by the ratio of the total number of reaction conformations generated from the conformational sampling algorithm and the targeted number of conformations to be used for TS characterization (N_{conf}). Here, we used the conformation-rich model when this ratio was greater than three, otherwise the conformation-poor model was used. If the number of conformations classified as "intended" was greater than N_{conf}, all of these conformations were ranked by a scoring function–here both the RF predicted intended probability and the GFN2-xTB energy were considered as options–and the highest N_{conf} conformations were selected; otherwise, all conformations that were classified as "intended" were retained for TS characterization. The training details and the description of the reaction dataset can be found in the SI.

2.3 Characterizing transition states.

The performance of the reported classifier was measured by how well it predicted promising conformations for localizing intended transition states for reactions that were not used in training. YARP was used to localize and characterize the transition states of the benchmarked reactions. These calculations consisted of performing a double-ended transition state search using the growing string method (GSM)[12,24] at the GFN2-xTB level of theory, followed by DFT-level Berny optimizations to refine the TS, and IRC calculations to characterize whether the resulting TS corresponded to the intended reaction. Several levels of DFT were used here, since some of the benchmarks involved comparisons with previous studies. The corresponding functional and basis set are reported where each benchmark system is discussed. Gaussian 16 was used as the reference quantum chemistry engine for the DFT calculations associated with the Berny optimizations and

IRC characterizations.[25] The GSM calculations were performed by interfacing YARP with the pyGSM package[26] using most of the default convergence hyperparameters (e.g., climbing image and the translation-rotation-internal coordinate system) with 9 nodes in the case of the relatively simple organic reaction benchmark, and 11 nodes for all other cases. All GFN2-xTB calculations were performed with the xTB program (version 6.2.3) maintained by the Grimme group.[22] Universal Force Field (UFF) based geometry optimizations were performed with Open Babel (version 2.4.1).[27] Atomic Simulation Environment (ASE)[28] was called to apply the RMSD minimization. All simulations were run on a 448 node commodity cluster composed of two AMD Rome CPUs (2.0 GHz), 128 effective cores, and 256 GB of memory per node. DFT calculations were performed with 32-core parallelization, while all other calculations were performed as bundled single-core jobs.

3 Results and discussion

3.1 Benchmark on simple organic reaction predictions

The first benchmark was chosen to establish the impact of conformational sampling on relatively simple systems (e.g. less than 20 atoms).[29] For this purpose we revisited 284 reactions from the original YARP study that failed to converge or localized to unintended TSs when characterized without conformational sampling. These reactions were generated by performing comprehensive graph-based reaction enumeration (i.e., all possible break two bonds form two bonds reactions for each reactant) on 20 reactants from the Zimmerman dataset and γ-ketohydroperoxide.[7,30] Since these reactions are generated using graphical rules rather than established reaction templates, it is expected that many of them are intrinsically high-barrier or may not possess any intended TS even were a comprehensive search of the potential energy surface possible. Among the 656 attempted reactions, YARP without conformational sampling failed to converge at either the GSM,

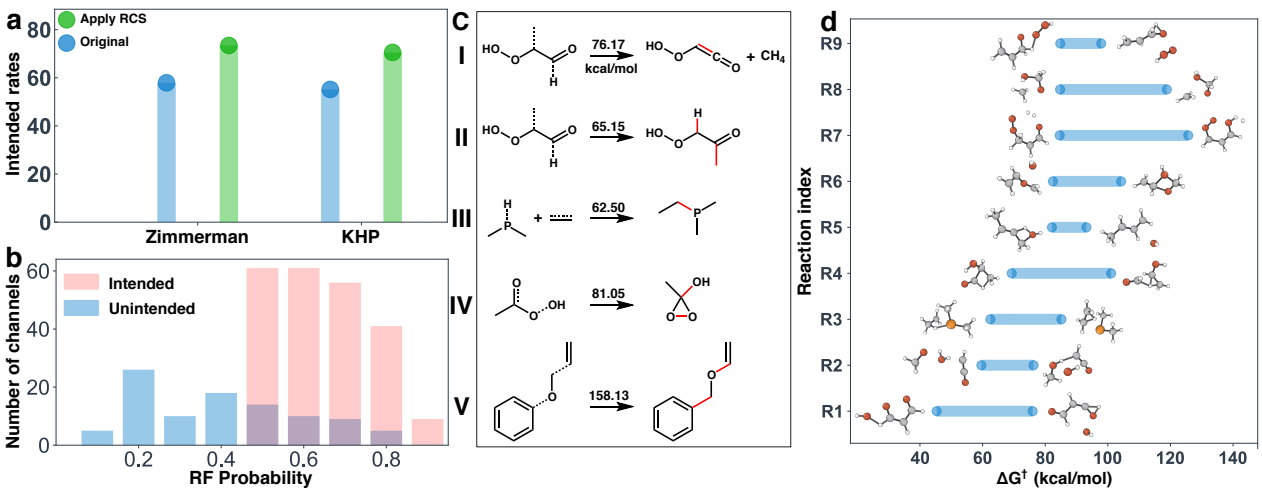

Figure 2: Overview of the performance of YARP with reaction conformational sampling on simple organic reactions. (a) The intended rate of reaction prediction for the Zimmerman and KHP datasets. (b) Distribution of RF intended probabilities using the conformation-poor model applied to intended and unintended reaction conformations. (c) Five types reactions where intended TSs were only discovered after conformational sampling. I: E1 and E2 elimination reactions, appear 45 times out of 101 reactions, II: small fragments exchange reaction, appear 15 times, III: addition and insertion reactions, appear 12 times, IV: 3- and 4-membered ring closure reactions, appear 9 times, V: inner rotation reactions, appear 5 times. (d) 9 reactions with the range of activation energies distribution larger than 10 kcal/mol, the lowest and highest energy transition states geometries are provided (grey: Carbon, red: Oxygen, yellow: Phosphorus, white: Hydrogen).

Berny optimization, or IRC calculation steps for 54 reactions and located unintended TSs for 230 reactions. Together, these 284 previously failed or unintended reactions were selected as a test for whether conformational sampling would yield intended TSs. Additionally, 35 reactions from the Zimmerman dataset that previously localized intended transition states were also included to ensure that the addition of conformational sampling did not inhibit localization of previously established TSs. Conformational sampling of all reactants was performed followed by down-selection of up to eight reaction conformers ($N_{\text{conf}} = 8$). The reactions from the Zimmerman dataset and KHP systems were characterized at the B3LYP/6-31G** level to be consistent with earlier work.

For 270 out of the 284 previously failed or unintended reactions, reactant conformational sampling generated at least one conformation predicted by the RF model to yield an intended TS.

For the remaining 14 reactions, no reaction conformation passed the indicator criteria, suggesting that these are fundamentally unphysical reactions (Fig. S6). Among the 270 attempted reactions, YARP located at least one TS and one intended TS for 257 and 101 reactions, respectively, corresponding to a success and intended rate of 95.2% and 37.4%. Considering that 0% of this subset of reactions converged to an intended transition state previously, this is a compelling demonstration that many reaction channels, even for small systems, can easily be neglected without conformational sampling. Out of the 35 previously intended reactions, all exhibited conformations that yielded intended transition states, demonstrating that the RF-based down-selection of conformers has no discernible negative impact compared with the previous algorithm. Accounting for the discovery of these intended TSs, the intended rate for the reactions in the Zimmerman dataset and KHP decomposition network are increased from 57.9% and 54.8% to 73.4% and 70.2%, respectively, compared with the earlier work (Fig. 2a).

To illustrate the performance of the RF classifier, the predicted intended probability for unintended and intended reaction conformations from this benchmark are provided Figure 2b. The results for the conformation-poor RF model, which is biased against false-negatives, are presented. The unintended conformers all come from the earlier study, whereas the intended conformers were those generated from conformational sampling and confirmed by IRC calculations. The intended conformers are all classified as such by the conformation-poor RF model (i.e., perfect recall) and the mean classification of the unintended conformers is unintended. This separation is favorable given the small number of features being used for the classification; nevertheless, a small number of unintended conformers still exhibit a high intended probability, which has motivated us to always choose $N_{\text{conf}} > 1$ in all presented benchmarks. The corresponding RF predictions for the conformation-rich RF model are presented in Fig. S4.

The 101 reactions that changed from unintended to intended were classified by mechanism (Fig. 2c) to investigate why some of the reactions benefited from conformational sampling. By

frequency, the five types of reactions are elimination (I, 45 out of 101 reactions), fragment exchanges (II, 15/101), addition and insertion (III, 12/101), 3- and 4-membered cyclizations (IV, 9/101), and inner rotations (V, 5/101), with the remainder composed of other mechanisms. I includes both E1 and E2 (i.e. unimolecular and bimolecular) eliminations; reaction II corresponds to the exchange in the position of two fragments, each consisting of no more than one heavy atom (e.g. in Fig. 2c, a methyl group and an hydrogen atom exchange); reaction V refers to a head-to-tail rotation of an internal segment of a reactant. Reactions I and II often occur as competing reactions involving the same bond breaks (i.e., they could easily discover an unintended TS that corresponds to the alternative). Notably, all of these mechanisms involve a large rotation or translation of atoms between their reactant and product geometries, such that conformational sampling is consequential despite the relatively small system size. Additional comparisons of intended and unintended TS geometries and for these classes of reactions are provided in Figure S5.

Conformational sampling also leads to the discovery of multiple TSs for many of the reactions that exhibit distinct mechanisms and a broad range of activation energies (Fig. 2d). Out of 136 intended reactions (101 from previous unintended reactions and 35 from previous intended reactions), 9 exhibit multiple intended TSs with activation energies spanning more than 10 kcal/mol. These large differences in activation energies correspond to qualitatively different mechanisms in several cases. Considering reaction R1, for example, the higher energy TS corresponds to the involvement of a hydrogen-roaming intermediate compared with the lower activation energy pathway that bypasses roaming. In another example, reaction R7, the TS corresponding to a concerted reaction mechanism is much more stable than the TS corresponding to a sequential mechanism. In summary, even for this benchmark of reactions on small molecules, applying conformational sampling results in improved success rates, improved intended rates, and the discovery of lower activation energy pathways.

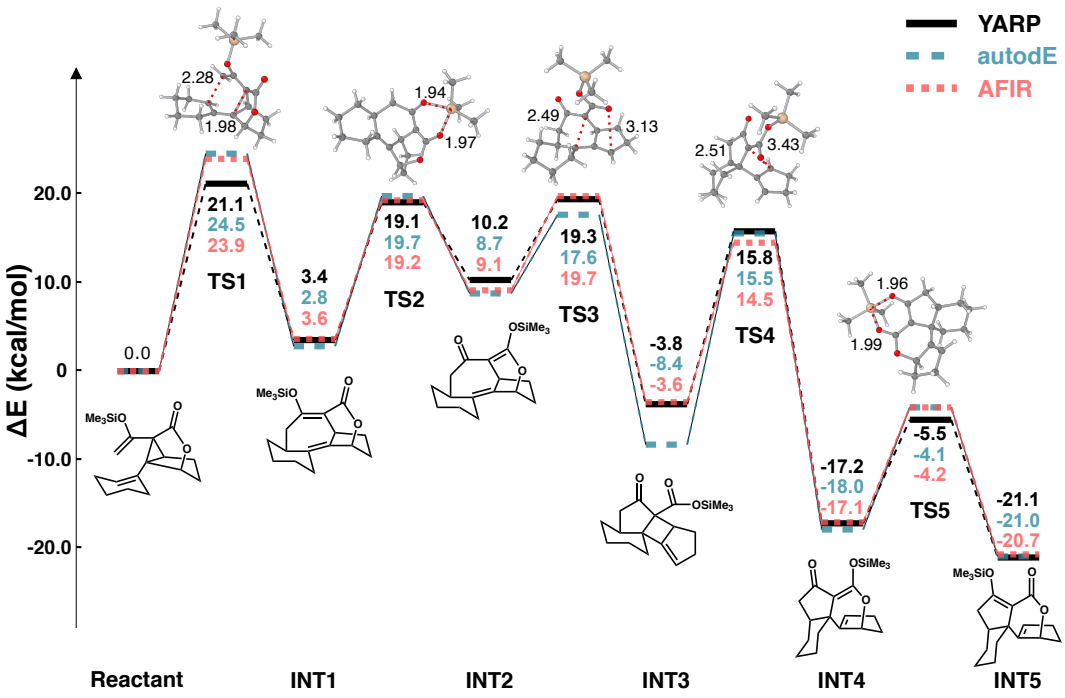

Figure 3: Ireland–Claisen rearrangement intermediate (INT) and transition state (TS) energies calculated using YARP with conformational sampling (black), autodE (green) reported by Young et al.[8] and AFIR (red) reported by Lee et al.[31] All curves are calculated at the B3LYP-D3BJ/6–311++G(2d,2p) and B3LYP-D3BJ/6-31G(d) levels of theory for the single point calculations and geometry optimizations, respectively. The CPCM(hexane) solvent model was used in autodE, whereas the IEF-PCM(hexane) solvent model was used in this work and AFIR.

3.2 Ireland–Claisen rearrangement

The multi-step Ireland-Claisen rearrangement shown in Figure 3 was selected for the second benchmark, since it exhibits increased conformational complexity and has recently been characterized using other automated transition state localization methods. Lee et al.[31] proposed the illustrated Ireland–Claisen rearrangement pathway consisting of 5 sequential rearrangements as discovered by the Artificial Force-Induced Reaction (AFIR) method.[32] Young et al. performed a follow-up study using the autoDE method and a heuristic for sampling conformations to localize distinct transition states.[8] Here, conformational sampling was performed on all five reactants to generate reaction conformers that were then ranked using the RF models to down-select up to 8 reaction conformations for each step ($N_{conf} = 8$). The transition state localization was performed at the B3LYP-D3BJ/6–31G(d) level of theory and further evaluated at the B3LYP-D3BJ/6–311++G(2d,2p) level of theory with the IEF-PCM(hexane) solvent model to be consistent with the earlier AFIR study. Among the 40 total attempted reactions, all successfully localized a TS and 37 were confirmed to be intended TSs after IRC characterization. In total only 465 DFT gradient calls were used (11.6 per reaction) to locate these TSs. These two statistics indicate the high quality of the reaction conformations generated by RF ranking.

The three methods predict nearly identical single-point energies for the intermediates and TSs (Fig. 3). The deviations among three methods are typically within 2 kcal/mol, indicating similar reaction mechanisms are being described. Notably, YARP discovers a low energy reaction conformation leading to a ~ 3 kcal/mol reduction of **TS1**. Another large deviation of ~ 5 kcal/mol is observed for **INT3** (i.e., not for a transition state but for an intermediate). A more stable conformation is reported by autodE, whereas the energies reported by AFIR and YARP are similar. This may be due to the distinct solvation model used by autodE in comparison with AFIR and YARP, or differences in sampling the intermediate geometries. Nevertheless, the Ireland–Claisen

rearrangement example demonstrates the transferability to more complex systems of the indicator function for down-selecting reaction conformations and the reproduction of TS energies achieved by contemporary approaches.

3.3 Intramolecular Diels-Alder reactions of a ketothioester

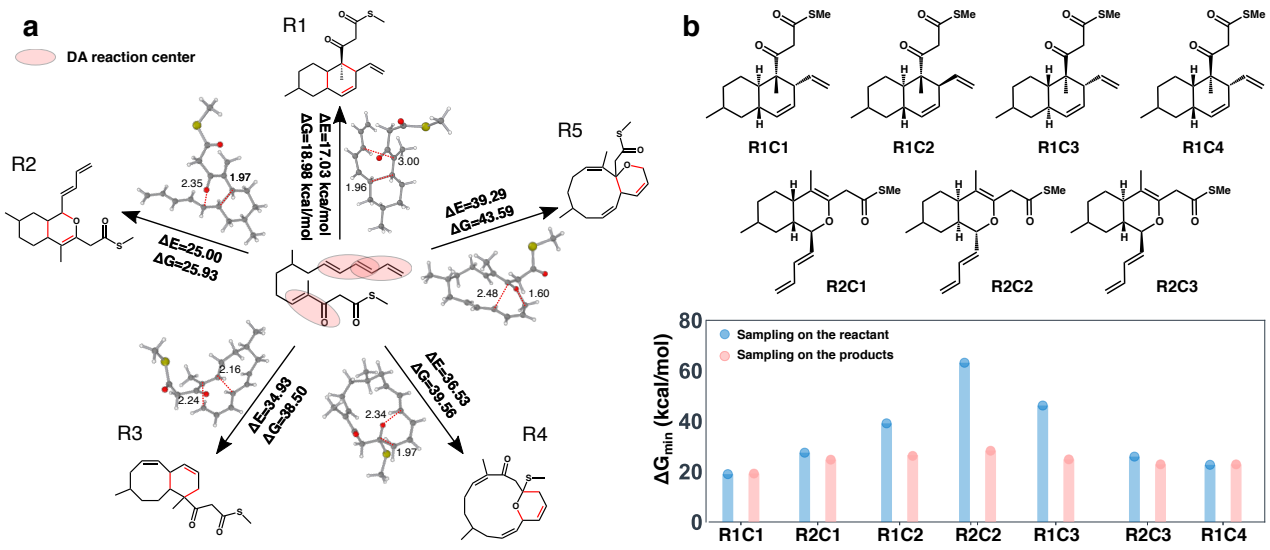

Figure 4: Five competing Diels Alder (DA) reactions of ketothioester with activation energy lower than 40 kcal/mol. a. 5 Diels Alder products with corresponding activation energies (ΔE), free energies of activation (ΔG) and transition state structures. The bonds denoted as red refer to bonds formed during the DA reaction. b. Product conformations identified by YARP in the most favorable reaction 1 and 2. Corresponding free energies of activation are provided in the bottom.

For a third benchmark system we investigated competing Diels-Alder (DA) ring-closures for the ketothioester shown in Figure 4. Although some cyclizations were present in the previous benchmarks, this case presents additional isomeric complexity, with 18 possible DA ring-closures yielding up to 304 possible stereoisomers. In the context of a double-ended TS search, the product stereochemistry can be fixed by the user; but more generally, the relative likelihood of different product stereochemistries is something that a reaction prediction workflow would need to predict. With respect to conformational sampling, this poses an additional challenge since reactant-side

conformational sampling neglects the conformational constraints that are unique to each stereoisomer. However, performing product-side conformational sampling on all 304 distinct stereoisomers is wasteful since many, though not all, of the products can be ruled out using heuristics. To investigate this asymmetry, conformational sampling was first performed on the ketothioester reactant to generate 10 reaction conformations for each of the 18 possible Diels-Alder reactions. These reaction conformations were ranked by the RF models, thus the resulting stereoisomers were selected solely based on the likelihood that they would be connected by an intended transition state rather than any chemical intuition. For a direct comparison with previous results, all TSs were optimized and IRC calculations were performed at the B3LYP/6-31G level of theory (Fig. S3). We also reperformed all single-point calculations at the more accurate B3LYP-D3BJ/def2-TZVP level of theory (Fig. 4).

The reactant-side conformational search yielded at least one intended TS for all 18 attempted DA ring-closures. Among them, 5 reactions exhibit activation energies less than 40 kcal/mol (Fig. 4a). The two lowest activation barrier reactions, denoted as R1 and R2, are two reactions reported previously by Yang et al.,[20] in an automated reaction prediction study focusing on a subset of reactive atoms of this ketothioester. In contrast, R3, R4 and R5 have not been previously reported. These reactions display larger barriers but are still potentially kinetically relevant. It is encouraging that YARP, with reactant-side conformational sampling, predicts the same lowest barrier DA closures as in the earlier study despite not using a restricted subset of atoms.

The present experiment found intended pathways to 38 distinct stereroisomers out of the 304 possible that could result from the 18 DA reactions. In the cases of R1 and R2, intended TSs were localized for four and three distinct enantiomers out of 32 and 16 possible, respectively. Six of these correspond to those previously reported by Yang et al., R1C1 is new, and the earlier study reported intended TSs for two enantiomers that were not attempted here. Despite this large overlap in predicted enantiomers, several of the activation energies are overestimated using

reactant-side conformational sampling (Fig. 4b). In particular, we interpret this result as arising from the asymmetric conformational constraints of the reactant versus products in the case of these ring closures. To clarify this, we reperformed the transition state searches for these seven enantiomers using product-side conformational sampling. In all cases, the activation energies discovered after product-side sampling are lower than those discovered from reactant-side sampling, and the differences in the activation energies for distinct entantiomers largely vanishes. This is a rather dramatic illustration of the inequivalence of reactant-side versus product-side conformational sampling. This asymmetry is present to varying degrees in any reaction, but it is acute for cyclizations and reactions yielding stereoisomers. In contrast, for the previous benchmarks we observed relatively little distinction between performing conformational sampling on the reactants versus the products, then generating conformationally aligned structures.

3.4 Head-to-Tail cyclic tetrapeptide formation

For a final benchmark we selected the cyclic peptide formation of a series of large tetrapetides. Peptide cyclization is challenging for TS searches because of the potentially large conformational search space of both the linear chains (reactants) and relatively large rings (products). Ring contractions are also relevant to these systems such that the conformational space of reactants and products are associated with distinct rings. For this benchmark we selected a series of tetrapeptides that have been experimentally studied in the context of lowering the activation energy for accessing certain cyclic peptide products via tailored substitutions to the peptide backbone.[33] In particular, the SAAA tetrapeptide (Fig. 5a) undergoes a relatively high barrier cyclization via dehydration to form a 12-membered ring. SAAA-aux (Fig. 5b) is hypothesized to undergo a facile ring-contraction to the same 12-membered ring. Finally, SAAA-SAL (Fig. 5c) was previously predicted to undergo cyclization to a 16-membered precursor that later undergoes more facile ring contraction to the 12-membered product; whereas the direct formation of the 12-membered product from SAAA-

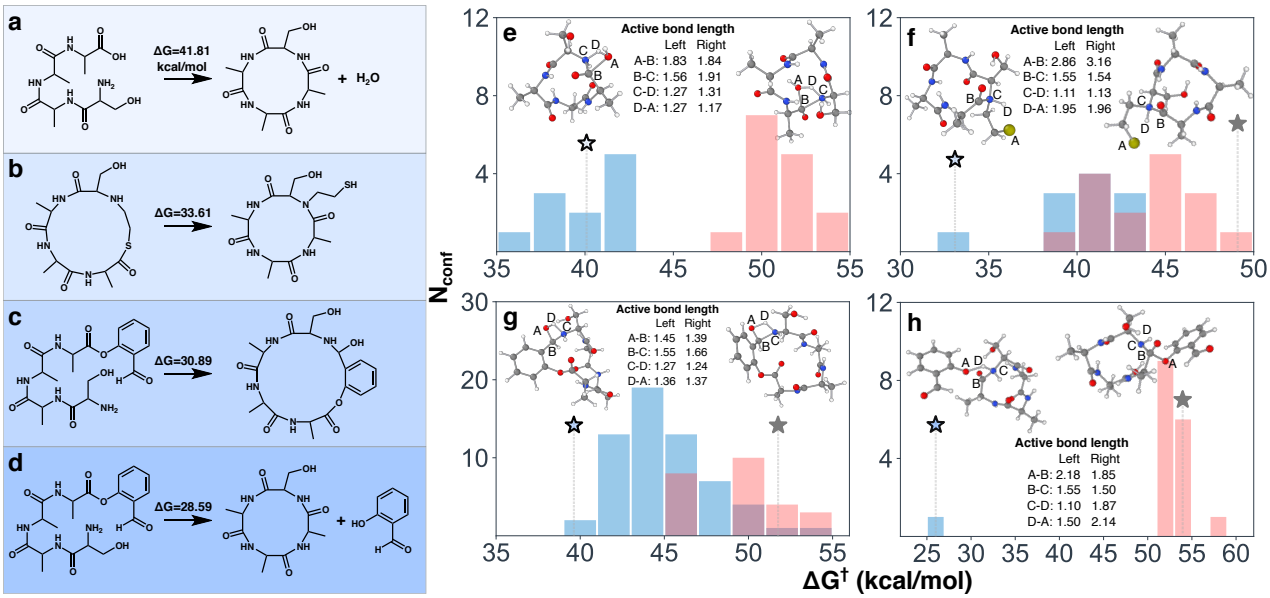

Figure 5: a-d: Four cyclic tetrapeptides formation reactions with reactant of a. D-seryl-D-alanyl-L-alanyl-D-alanine (denoted as SAAA), b. 12-(hydroxymethyl)-3,6,9-trimethyl-1-thia-4,7,10,13-tetraazacyclopentadecane-2,5,8,11-tetraone (Ethanethiol Auxiliary SAAA, denoted as SAAA-aux), c-d. 2-formylphenyl D-seryl-D-alanyl-L-alanyl-D-alaninate (SAL ester auxiliary SAAA, denoted as SAAA-SAL). The activation energies were computed at the M062X/6-311+G(d,p) level of theory on the geometries of the most stable transition state (TS) sampled at B3LYP/6-31G level. The SMD solvation model for acetic acid was applied. e-f: Examples of TS conformational sampling starting from the lowest and highest transition states. The position of each star refers to the activation energy of the initial TS. Constrained distances of all TSs are shown with unit Å. The lowest and highest free energies of activation (ΔG) of (a) SAAA, (b) SAAA-aux (c) and (d) SAAA-SAL are 41.2/65.0; 32.5/49.5; 39.1/53.0; 26.1/53.7 (unit is kcal/mol, computed at B3LYP/6-31G level, the highest energy TS in case a is 65 kcal/mol and is not shown in the figure)

SAL (Fig. 5d) was ruled out as being unfavorable. Some computational work has been done to rationalize these transformations,[33,34] but no DFT-level transition states have been calculated for these large and conformationally challenging systems.

Conformational sampling was performed on three different cyclic tetrapeptide precursors (Fig. 5) to generate reaction conformers that were then ranked using the RF models to down-select up to 20 reaction conformations for each reaction ($N_{conf} = 20$). YARP was first applied at the B3LYP/6-31G level to locate and characterize intended TSs. The most stable intended TSs for

each reaction were further characterized at the M062X/6-311+G(d,p) level of theory (geometries taken from the B3LYP/6-31G results) with the SMD solvation model[35] for acetic acid, to match the level of earlier computational work.[33]

For all four reactions, the RF-ranked conformations converged to at least one intended TS (Fig. 5a-d), which represents a first for all of these reactions. Based on the experimental data, it is expected that the introduction of the ethanethiol (Fig. 5b) and SAL-ester (Fig. 5c) auxiliaries will promote 12-membered cyclization in comparison with the SAAA cyclization (Fig. 5a). The lowest barrier intended TS discovered for each reaction likewise confirms this trend. Interestingly, YARP predicts that the direct cyclization of the SAAA-SAL species is more favorable than proceeding through the 16-membered precursor that was previously invoked.[33]

Since these tetrapeptides are the largest systems investigated here, we decided to also use them as a case study of the potential benefits of performing conformational sampling on already localized TSs to search for a lower barrier crossing. In particular, several recent studies have used this approach,[8,16] but no direct comparison has been performed between TS conformational sampling and reactant/product-side conformational sampling. Conformational sampling of TSs is more difficult than that of reactants or products, because a TS is a saddle point and thus conformational sampling can easily lead to an unintended reaction channel. To prevent this, auxiliary restraints are used for TS sampling that try to maintain the bond lengths of atoms involved in the imaginary frequency mode, but there is still no guarantee that the resulting TS conformer will remain an intended channel, and thus it must be verified by further IRC calculations.

Constrained TS conformational sampling was applied to both the lowest and highest barrier intended TSs obtained for the four tetrapeptide reactions. CREST was used to perform the TS conformational sampling while restraining the active bond (i.e. bonds broken or formed in each reaction) lengths to their distances in the corresponding intended TS (these distances are shown in the inset of Fig.5e-h). The TS conformations generated in this way were then subjected to Berny

optimization, IRC calculations, and the distribution of barrier heights for the intended TSs are reported in Fig. 5e-h. TS conformational sampling on the lowest barrier intended TSs (stroked star marker) had only a marginal benefit. For three of the cases, all of the discovered TSs are of equal or higher barrier to the original intended TS. For SAAA, the majority of new TSs were also of equal or higher barrier, but four lower barrier TSs exhibiting up to a 5kcal/mol reduction were discovered. In contrast, performing TS sampling on the highest barrier intended TSs from the original benchmark (filled star marker) yielded new lower barrier TSs in all cases. Yet critically, none of these new TSs exhibit lower barriers than the lowest barrier TS discovered from the original reactant-side conformational sampling search (i.e. the stroked star marker in Fig. 5e-h).

Our interpretation of this case study on TS conformational sampling is that it can lead to barrier reduction in some cases, but it is strongly dependent on the TS used to seed the search and it is not a substitute for reactant and/or product-side conformational sampling. Moreover, computational cost and efficiency are important factors in deciding whether to use TS conformational sampling. On average, 26 conformers were generated by CREST for each of the 8 TSs used to seed the TS conformational search. This is more than N_{conf} used for reactant-side sampling. Additionally, the intended rate of the generated TS conformers varies greatly between 100% and 3% (Fig. S2). For these reasons, TS conformational sampling as currently implemented remains a costly procedure of marginal benefit.

4 Conclusions and outlook

The maturation of conformational sampling algorithms has created new opportunities to incorporate conformational exploration into transition state searches. Nevertheless, the cost of performing parallel TS searches starting from all possible reaction conformers remains prohibitive for most applications. Here we have shown how this limitation can be side-stepped using a relatively simple

random forest classifier to rank reaction conformations before performing costly TS searches. The performance of this approach was investigated in four distinct benchmarks. First, we demonstrated that intended TSs could be localized for over a hundred reactions (˜39% of those attempted) that were previously discarded due to failed TS searches. Likewise, conformational sampling revealed many competing barriers for these reactions that would have been otherwise missed. Second, in head-to-head comparisons with other algorithms, we observed no loss in fidelity despite the several-fold reduction in computational cost associated with the present approach. In particular, conformational down-selection was able to reproduce all TSs of a complex multistep Ireland–Claisen rearrangement, as well as (re)discover the lowest barrier stereoisomeric product published to date (out of over three hundred possible) for competing DA ring closures of a model ketothioester. These demonstrations establish the versatility of using ML classifiers to downselect promising reaction conformers and thus mitigate the additional cost of incorporating conformational sampling into TS localization workflows. Conversely, these benchmarks illustrate the hazards, including overlooked reaction pathways and inaccurate barriers, of neglecting conformational sampling in TS searches.

There are still several avenues for improving the current approach. Conformer down-selection reduces costs, but it is still N_{conf} times more costly than a single TS search. One possibility is to use a low-level semi-empirical TS search on the down-selected configurations as a preliminary step to further reduce the number of reaction conformers subjected to high-level characterization. Likewise, many reaction conformers end up localizing redundant TSs, which it may be possible to anticipate in advance with a more sophisticated ranking procedure. There are also opportunities for developing an optimal conformer sampling policy. For example, in the third and fourth benchmarks we deliberately highlighted scenarios where there is asymmetry between reactant-side versus product-side versus TS conformational sampling. We concluded that TS conformational sampling alone is insufficient in all cases, but the importance of each is highly contextual and naïve inclusion of all cases can dramatically increase sampling costs. Finally, relatively simple RF models were

utilized here due to our desire for transferabililty and the scarcity of intended TS data for training. In the presented benchmarks, these models were able to profitably rank conformers, but they nevertheless still exhibit low confidence in classifying many conformations. As more reaction data becomes available, we envision more sophisticated models potentially displacing the RF models while keeping the overall workflow relatively unchanged. Similarly, the classification models were only trained to predict the likelihood that a reaction conformation would converge to an intended reaction and did not utilize or consider the reaction activation energy. Future models might also select conformations based on their relative likelihood of yielding a low barrier TS. Finally, we have focused on conformer selection for double-ended TS searches due to their much lower computational costs. Such down-selection is also compatible with single-ended models, although the featurization presented here would have to be modified.

Data and Code Availability

The authors declare that the data supporting the findings of this study are available within the paper and its supplementary information files. The packages used in this study and a guide to reproducing the results is available through GitHub under the GNU GPL-3.0 License [https://github.com/zhaoqy19
(will be updated when this paper is accepted).

Author Contributions

Q.Z. and B.M.S conceived and designed the study. Q.Z and H.H developed the tool, performed the data analysis, and wrote the paper. B.M.S. oversaw the project and wrote the paper.

Conflicts of interest

The authors declare no conflict of interest.

Acknowledgements

The work performed by Q.Z., H.H, and B.M.S was made possible by the Office of Naval Research (ONR) through support provided by the Energetic Materials Program (MURI grant number: N00014-21-1-2476, Program Manager: Dr. Chad Stoltz). B.M.S also acknowledges partial support for this work from the Dreyfus Program for Machine Learning in the Chemical Sciences and Engineering and the Purdue Process Safety and Assurance Center.

References

(1) Rodrigo, G.; Carrera, J.; Prather, K. J.; Jaramillo, A. DESHARKY: automatic design of metabolic pathways for optimal cell growth. *Bioinformatics* **2008**, *24*, 2554–2556.

(2) Wu, D.; Wang, Q.; Assary, R. S.; Broadbelt, L. J.; Krilov, G. A computational approach to design and evaluate enzymatic reaction pathways: application to 1-butanol production from pyruvate. *J. Chem. Inf. Model.* **2011**, *51*, 1634–1647.

(3) Stine, A.; Zhang, M.; Ro, S.; Clendennen, S.; Shelton, M. C.; Tyo, K. E.; Broadbelt, L. J. Exploring De Novo metabolic pathways from pyruvate to propionic acid. *Biotechnol. Prog.* **2016**, *32*, 303–311.

(4) Suleimanov, Y. V.; Green, W. H. Automated discovery of elementary chemical reaction steps

using freezing string and Berny optimization methods. *J. Chem. Theory Comput.* **2015**, *11*, 4248–4259.

(5) Zimmerman, P. M. Navigating molecular space for reaction mechanisms: an efficient, automated procedure. *Mol. Simul.* **2015**, *41*, 43–54.

(6) Kim, Y.; Kim, J. W.; Kim, Z.; Kim, W. Y. Efficient prediction of reaction paths through molecular graph and reaction network analysis. *Chem. Sci.* **2018**, *9*, 825–835.

(7) Zhao, Q.; Savoie, B. M. Simultaneously improving reaction coverage and computational cost in automated reaction prediction tasks. *Nat. Comput. Sci.* **2021**, *1*, 479–490.

(8) Young, T. A.; Silcock, J. J.; Sterling, A. J.; Duarte, F. autodE: Automated Calculation of Reaction Energy Profiles—Application to Organic and Organometallic Reactions. *Angew. Chem., Int. Ed.* **2021**, *60*, 4266–4274.

(9) Henkelman, G.; Jónsson, H. Improved tangent estimate in the nudged elastic band method for finding minimum energy paths and saddle points. *J. Chem. Phys.* **2000**, *113*, 9978–9985.

(10) Henkelman, G.; Uberuaga, B. P.; Jónsson, H. A climbing image nudged elastic band method for finding saddle points and minimum energy paths. *J. Chem. Phys.* **2000**, *113*, 9901–9904.

(11) Peters, B.; Heyden, A.; Bell, A. T.; Chakraborty, A. A growing string method for determining transition states: Comparison to the nudged elastic band and string methods. *J. Chem. Phys.* **2004**, *120*, 7877–7886.

(12) Zimmerman, P. M. Growing string method with interpolation and optimization in internal coordinates: Method and examples. *J. Chem. Phys.* **2013**, *138*, 184102.

(13) Behn, A.; Zimmerman, P. M.; Bell, A. T.; Head-Gordon, M. Efficient exploration of reaction paths via a freezing string method. *J. Chem. Phys.* **2011**, *135*, 224108.

(14) Baker, J.; Kessi, A.; Delley, B. The generation and use of delocalized internal coordinates in geometry optimization. *J. Chem. Phys.* **1996**, *105*, 192–212.

(15) O'Boyle, N. M.; Vandermeersch, T.; Flynn, C. J.; Maguire, A. R.; Hutchison, G. R. Confab-Systematic generation of diverse low-energy conformers. *J. Cheminf.* **2011**, *3*, 1–9.

(16) Pracht, P.; Bohle, F.; Grimme, S. Automated exploration of the low-energy chemical space with fast quantum chemical methods. *Phys. Chem. Chem. Phys.* **2020**, *22*, 7169–7192.

(17) Zimmerman, P. M. Automated discovery of chemically reasonable elementary reaction steps. *J. Comput. Chem.* **2013**, *34*, 1385–1392.

(18) Grambow, C. A.; Pattanaik, L.; Green, W. H. Deep learning of activation energies. *J. Phys. Chem. Lett.* **2020**, *11*, 2992–2997.

(19) Xie, X.; Clark Spotte-Smith, E. W.; Wen, M.; Patel, H. D.; Blau, S. M.; Persson, K. A. Data-Driven Prediction of Formation Mechanisms of Lithium Ethylene Monocarbonate with an Automated Reaction Network. *J. Am. Chem. Soc.* **2021**, *143*, 13245–13258.

(20) Yang, M.; Zou, J.; Wang, G.; Li, S. Automatic reaction pathway search via combined molecular dynamics and coordinate driving method. *J. Phys. Chem. A* **2017**, *121*, 1351–1361.

(21) Breiman, L. Random forests. *Mach. Learn.* **2001**, *45*, 5–32.

(22) Bannwarth, C.; Ehlert, S.; Grimme, S. 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.* **2019**, *15*, 1652–1671.

(23) Melander, M.; Laasonen, K.; Jonsson, H. Removing external degrees of freedom from transition-state search methods using quaternions. *J. Chem. Theory Comput.* **2015**, *11*, 1055–1062.

(24) Zimmerman, P. M. Reliable transition state searches integrated with the growing string method. *J. Chem. Theory Comput.* **2013**, *9*, 3043–3050.

(25) Frisch, M. J. et al. Gaussian 16 Revision C.01. 2016; Gaussian Inc. Wallingford CT.

(26) Aldaz, C.; Kammeraad, J. A.; Zimmerman, P. M. Discovery of conical intersection mediated photochemistry with growing string methods. *Phys. Chem. Chem. Phys.* **2018**, *20*, 27394–27405.

(27) O'Boyle, N. M.; Banck, M.; James, C. A.; Morley, C.; Vandermeersch, T.; Hutchison, G. R. Open Babel: An open chemical toolbox. *J. Cheminf.* **2011**, *3*, 33.

(28) Larsen, A.; Mortensen, J.; Blomqvist, J.; Castelli, I.; Christensen, R.; Dułak, M.; Friis, J.; Groves, M.; Hammer, B.; Hargus, C.; Hermes, E. a. The atomic simulation environment—a Python library for working with atoms. *J. Phys.: Condens. Matter* **2017**, *29*, 273002.

(29) Zhao, Q.; Savoie, B. YARP dataset. 2021; `https://doi.org/10.6084/m9.figshare.14766624`.

(30) Grambow, C. A.; Jamal, A.; Li, Y. P.; Green, W. H.; Zador, J.; Suleimanov, Y. V. Unimolecular reaction pathways of a γ-ketohydroperoxide from combined application of automated reaction discovery methods. *J. Am. Chem. Soc.* **2018**, *140*, 1035–1048.

(31) Lee, C. W.; Taylor, B. L.; Petrova, G. P.; Patel, A.; Morokuma, K.; Houk, K.; Stoltz, B. M. An Unexpected Ireland–Claisen Rearrangement Cascade During the Synthesis of the Tricyclic Core of Curcusone C: Mechanistic Elucidation by Trial-and-Error and Automatic Artificial Force-Induced Reaction (AFIR) Computations. *J. Am. Chem. Soc.* **2019**, *141*, 6995–7004.

(32) Maeda, S.; Taketsugu, T.; Morokuma, K. Exploring transition state structures for intramolec-

ular pathways by the artificial force induced reaction method. *J. Comput. Chem.* **2014**, *35*, 166–173.

(33) Wong, C. T.; Lam, H. Y.; Song, T.; Chen, G.; Li, X. Synthesis of constrained head-to-tail cyclic tetrapeptides by an imine-induced ring-closing/contraction strategy. *Angew. Chem., Int. Ed.* **2013**, *52*, 10212–10215.

(34) Meutermans, W. D.; Golding, S. W.; Bourne, G. T.; Miranda, L. P.; Dooley, M. J.; Alewood, P. F.; Smythe, M. L. Synthesis of difficult cyclic peptides by inclusion of a novel photolabile auxiliary in a ring contraction strategy. *J. Am. Chem. Soc.* **1999**, *121*, 9790–9796.

(35) Marenich, A. V.; Cramer, C. J.; Truhlar, D. G. Universal solvation model based on solute electron density and on a continuum model of the solvent defined by the bulk dielectric constant and atomic surface tensions. *J. Phys. Chem. B* **2009**, *113*, 6378–6396.