Review Article

Neil R. McFarlane and Jeremy N. Harvey*

Minimum energy path methods and reactivity for enzyme reaction mechanisms: a perspective

https://doi.org/10.1515/pac-2025-0484
Received April 16, 2025; accepted June 24, 2025

Abstract: In this perspective article, we discuss the link between minimum energy paths and activation parameters for reactions on complex potential energy surfaces involving many possible local minima, as are typically found for enzyme-substrate complexes. Such systems are frequently tackled with hybrid QM/MM methods in order to characterize reactivity. The link between local energy barriers along a minimum energy path, the so-called exponential average of such local energy barriers, and multiconformational transition state theory is discussed. Also, it is shown that in case of positive skewness of the distribution of barrier heights across sets of minimum energy paths, exponential averaging converges relatively quickly with the number of paths used.

Keywords: Enzyme catalysis; quantum chemistry; quantum science and technology; reaction mechanisms.

Introduction

Quantum chemical methods have proved highly successful in characterizing chemical reaction mechanisms, ranging from qualitative insights into the structure of intermediates and of transition states to quantitative description of reaction energetics and kinetics. In order to be able to study chemical reaction mechanisms in this way, one must obviously be able to perform quantum-chemical calculations of sufficient accuracy for a reasonably realistic model of the reacting system. The development of accurate density functional theory (DFT) functionals[1] and of efficient quantum-chemical codes have made the routine study of chemical reaction mechanisms possible. Efficient methods to systematically calculate the correlation energy, such as the domain-based local pair-natural orbital-based coupled-cluster theory (in particular the variant with perturbative treatment of triples excitations,[2] DLPNO-CCSD(T)), provide refined energies that are essential for obtaining quantitative insight.

Aside from the need to be able to calculate the energy of a given molecular system (or assembly of multiple molecules) for a given set of atomic coordinates, understanding reaction mechanisms also relies heavily on the ability to perform structural optimization, using the energy gradient.[3] The ability to locate saddle-points as excellent structural models for transition states[4] is of particular importance, as is the ability to identify the so-called "intrinsic reaction coordinate",[5] a reaction path leading downwards on each side of a saddle-point towards reactants and products. The properties of the saddle-point can be used with various statistical rate theories to

Article note: A collection of invited papers to celebrate the UN's proclamation of 2025 as the International Year of Quantum Science and Technology.

***Corresponding author: Jeremy N. Harvey**, Department of Chemistry and Division of Quantum Chemistry and Physical Chemistry, KU Leuven, Celestijnenlaan 200F, B-3001, Leuven, Belgium, e-mail: Jeremy.harvey@kuleuven.be. https://orcid.org/0000-0002-1728-1596
Neil R. McFarlane, Department of Chemistry and Division of Quantum Chemistry and Physical Chemistry, KU Leuven, Celestijnenlaan 200F, B-3001, Leuven, Belgium. https://orcid.org/0000-0003-2414-6461

predict rate coefficients, with more advanced such theories using properties all along the reaction path[6] to compute the rate coefficient.

As molecular systems become larger, determining reactivity by locating stationary points on the potential energy surface becomes more difficult for several reasons. First and most obvious, quantum-chemical calculations become more demanding in terms of processing power and in terms of storage of intermediate quantities (in memory or in storage) as the number of atoms increase. The scaling is effectively always at least linear with respect to the number of atoms N in the system modelled, and is often quadratic, cubic or more N^m (with $m = 2, 3, \ldots$) and for exact solutions is even exponential $\exp(\alpha N)$. Efficient quantum-chemical methods mean that it is now possible to obtain accurate energies (with expected errors close to or approaching "chemical accuracy" or 1 kcal/mol) even for quite large systems with many tens or even hundreds of atoms in a reasonable amount of time and using reasonable computational infrastructure.[2]

The issue that concerns us here is that the complexity of the potential energy surface also increases steeply with the number of atoms in the modelled system. The number of distinct local minima on the potential energy surface also increases exponentially, as $\exp(\alpha' N)$,[7] so that locating the overall lowest energy point or global minimum – something that is more or less trivial for small systems with only a handful of atoms – becomes a major challenge. The problem of describing the "mechanism" of reaction on such a complicated energy landscape also becomes very challenging. In principle, one needs to calculate the relative free energy ΔG_i of each of the $i = 1, \ldots n_c$ local minima (with $n_c = \exp(\alpha' N)$) and then also locate transition states for all relevant pathways leading from one local minimum i to another j, together with the associated activation free energies $\Delta G^{\ddagger}_{ij}$ and hence the rate coefficients k_{ij}. Then the behaviour of the system – assuming that one starts from some initial minimum i_{init} (or some distribution of initial minima) – can be modelled by solving the relevant set of kinetic equations. The emerging time-series of abundances for each of the species can then be plotted and examined in order to compare to already performed experiments or to make predictions about the behaviour of a system that is yet to be studied in the lab.

In many cases, the above procedure can be simplified by noting that the set of minima on the potential energy surface can be classified *a priori* as corresponding either to "reactant" states $i \in \mathbf{R}$, or "intermediates" $i \in \mathbf{I}$, or "products" $i \in \mathbf{P}$. Furthermore, it can often be shown that the pathways converting one particular reactant state to another have low barriers hence large rate coefficients, such that the ensemble of reactant states **R** is in quasi-equilibrium, whereby it is possible to predict the relative populations of the different reactant states throughout the whole reaction process simply by considering their relative free energies ΔG_i (taken here to be expressed relative to the most stable form of the reactants, which hence has $\Delta G_i = 0$). The same applies to intermediates and products. Furthermore, one can often also assume that each saddle point or transition state can be classified as "connecting" just one particular reactant minimum i to a given product state j. A rate coefficient k_{ij} and a free energy of activation $\Delta G^{\ddagger}_{ij}$ can be associated to each such process. In fact, in this picture where the reactant minima are in quasi-equilibrium with each other, it is no longer strictly necessary to "tag" each TS with the corresponding reactant and product minima, and one can simply speak of individual rate coefficients k_j for reaction over a given TS j, with associated free energies of activation $\Delta G^{\ddagger}_{j}$, expressed as a difference in free energy with respect to a single reference point, *e.g.* the lowest relative free energy of all of the reactant minima. We will assume that each rate coefficient is given by canonical transition state theory (TST):

$$k_j = \frac{k_B T}{h} \exp\left(\frac{-\Delta G^{\ddagger}_{j}}{RT}\right) \tag{1}$$

Where k_B is the Boltzmann constant, T is the absolute temperature, h is Planck's constant, and R is the gas constant. In principle, we should perhaps include a standard state concentration term to allow for differences in molecularity between reactants and TS but we will assume all reactions are unimolecular so this term vanishes. For bimolecular reactions, the large entropy change that arises from the loss of relative translational degrees of freedom makes a significant contribution to the free energy of activation, but again, in the present case this will be neglected. Also, canonical TST is of course not exact, and we could account for deviations by including a

transmission coefficient, κ_j, but here we will instead assume for simplicity that TST is exact, so that this term also can be omitted. This can also be understood as taking equation (1) as being the *definition* of the free energy of activation, given a presumed exact value for k_j.

From now on, in this discussion, we will focus on the simplest possible case of an irreversible one-step reaction A → B leading from an ensemble of "reactant" minima A_i to an ensemble of "product" minima B_i, through a set of saddle points TS_j. As the reaction is assumed to be irreversible, we do not need to take account of the relative free energies of the different product conformers. This simplified model can be generalized to describing multi-step reactions but this will not be done here. Note also that in practice, when studying detailed potential energy surfaces, one often finds that the simplistic picture above whereby one reactant state minimum corresponds to one minimum energy reaction path, to one saddlepoint, to one product state minimum needs to be refined. Reaction paths can merge or split through valley-ridge inflexion points (VRI),[8] and when taking dynamical behaviour into account rather than assuming that minimum energy paths are sufficient to describe motion on the potential energy surface, an even more complicated picture emerges. Again, we will not discuss these issues in detail here.

With the above framework, it can be shown that the overall reaction kinetics for conversion of reactant A to product B can be described using a single rate coefficient k_{AB} which is given by the following *multi-conformer transition state theory* (MCTST) expression:

$$k_{AB} = \frac{\sum_j k_j}{q_R} \tag{2}$$

In this expression, each rate coefficient k_j is given by equation (1), with an activation free energy $\Delta G_j^{\ddagger}$, expressed with respect to the most stable minimum of the reactants, and the reactant partition function q_R is given in terms of the relative free energies ΔG_i of each member of the ensemble **R** of reactant-state minima (also expressed relative to the most stable minimum of the reactants) by:

$$q_R = \sum_{i \in \mathbf{R}} \exp\left(\frac{-\Delta G_i}{RT}\right) \tag{3}$$

These equations of multiconformer TST can be rewritten in a large number of different ways[9,10] with various notational conventions and different choices concerning the reference energy that is taken for each of the reactant minimum relative free energies ΔG_i and for each of the activation free energies $\Delta G_j^{\ddagger}$. Note that here we have taken the free energies to be expressed relative to the most stable form within the ensemble of reactant states, but this is not a requirement, and any other single reference state could be used, and would yield the same value of k_{AB} provided that it was consistently applied to both the reactant state ensemble free energies and the activation free energies. Using the suggested convention has the advantage that in case there are relatively few reactant state minima, they are relatively spread out in free energy, there are relatively few TSs, and they are quite spread out in free energy, then $q_R \cong 1$ and $k_{AB} \cong k_1$, the rate coefficient for passing over the lowest TS starting from the most stable reactant minimum. In other words, in this limit, equation (2) simply reduces to simple TST.

One different way of expressing equation (2) adopts a convention that will be more convenient when describing the biochemical reaction examples discussed below. It takes the same assumption as above for the relative free energies ΔG_i of the reactant state minima but for the TSs, instead assumes that each TS is associated to a given reactant minimum *i* (and a given product minimum *j*) and works with rate coefficients k'_{ij} which describe the hypothetical rate at which the reactant, *in case it was confined to exist only as minimum* A_i, would convert to product over the corresponding TS_j. These modified rate coefficients are given by:

$$k'_{ij} = \frac{k_B T}{h} \exp\left(\frac{-\Delta G_{ij}^{\ddagger}}{RT}\right) \tag{4}$$

Note that k'_{ij} is not equal to k_j, because $\Delta G_{ij}^{\ddagger}$ is not equal to $\Delta G_j^{\ddagger}$. The former is the difference in free energy between reactant minimum A_i and TS_j, while the latter is the difference in free energy between the globally most stable form of the reactants (or some other fixed reference state) and TS_j. Hence $\Delta G_j^{\ddagger} = \Delta G_{ij}^{\ddagger} + \Delta G_i$, and k_j is equal to

$k'_{ij} \times \exp(-\Delta G_i/RT)$. When expressing the overall MCTST rate coefficient in terms of the k'_{ij}, one should use the following modified form of equation (2):

$$k_{\mathrm{AB}} = \frac{\sum_i k'_{ij} \exp(-\Delta G_i/RT)}{q_{\mathrm{R}}} \tag{5}$$

It can also be noted that using the rate coefficient from equation (2) or (5) and reversing the Eyring equation of TST makes it possible to define an overall effective activation free energy $\Delta G^{\ddagger}$:

$$\Delta G^{\ddagger} = -RT \ln\left\{k_{\mathrm{AB}} \times \frac{h}{k_B T}\right\} \tag{6}$$

This can be rewritten with reference to equations (3)–(5), yielding:

$$\Delta G^{\ddagger} = -RT \ln\left\{\frac{\sum_i \exp\left(\frac{-\Delta G^{\ddagger}_{ij}}{RT}\right)\exp(-\Delta G_i/RT)}{\sum_i \exp\left(\frac{-\Delta G_i}{RT}\right)}\right\} \tag{7}$$

This multiconformer model for TST can also be represented graphically as *e.g.* shown in Fig. 1.

Multi-conformer forms of TST are increasingly used when studying reactivity of "larger" molecules for which equation (2) (or its modified form (5)) do **not** reduce to simple TST. This occurs whenever more than one minimum or conformer of the reactants makes a significant contribution to q_{R}, and/or whenever more than one reaction path makes a non-negligible contribution to the overall rate of the reaction. In fact, it can be argued that even 'simple' TST often uses a disguised form of MCTST by accounting for "symmetry" factors taking into account the different permutation symmetry of the TS with respect to the reactant. In any case, non-trivial forms of MCTST are used nowadays in the literature in the context of many quantum-chemical studies of reactions for systems with as few as 10 to 20 atoms, showing that the already-mentioned exponential scaling of the number of distinct minima (and saddlepoints) on a potential energy surface already has very noticeable impact for quite small systems.

For reactions of *very* large systems, such as enzyme:substrate complexes, it is thereby obvious that some form of MCTST will need to be used in order to perform reactivity studies. In this perspective, we discuss different ways in which MCTST can in practice be implemented for such cases, and discuss in particular the accuracy of an approach based very closely on equation (5).

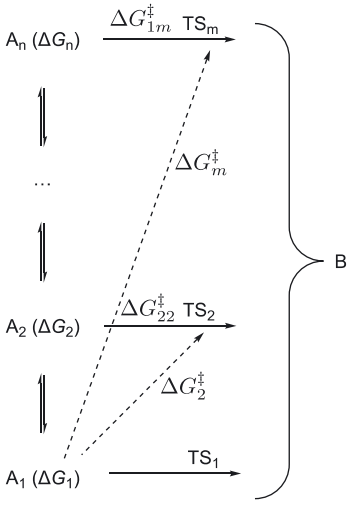

Fig. 1: Multiconformer TST (MCTST) model for a reaction A → B, taking into account multiple conformers A_i and TS_j of the reactant and TS, with relative free energies ΔG_i and $\Delta G^{\ddagger}_j$. A given TS_j has a free energy $\Delta G^{\ddagger}_{ij}$ relative to a given conformer A_i.

Rate theory for enzyme reactions

In this study, we consider how quantum-chemical methods can be used to study reaction mechanisms in enzyme-catalyzed reactions. We focus our discussion on the problem of applying TST, whether in the form of simple TST, MCTST, or more complex forms. As can perhaps be anticipated from the discussion in the previous section, moving from molecular systems to enzymes, and the associated increase in the number of atoms within the systems treated, carries with it two challenges: the scaling of the computational cost of electronic structure theory methods with the number of atoms treated, and the scaling in the number of relevant minima and saddle points on the potential energy surface. We will discuss ways to deal with each of these problems.

Focussing first on the quantum-chemical scaling problem, two main techniques have emerged for solving the problem of calculating quantum-chemical wavefunctions for enzyme systems with their many thousands of atoms. The first approach, analogous to the solution found in the early days of quantum chemical studies of molecular chemistry when dealing with larger molecules, involves using a **model** of the enzyme:substrate system, comprising most of or all the substrate and/or cofactor atoms, together with atoms from the sidechains (and perhaps backbone) of important amino-acid residues surrounding the active site and perhaps important water molecules that participate in hydrogen-bonding to the substrate or other important functional groups. Even when using a quite broad definition of "important", this approach rarely leads to models comprising much more than a few hundred atoms, which is well within reach for contemporary quantum-chemical techniques like density functional theory (DFT) or local versions of correlated *ab initio* methods like DLPNO-CCSD(T). This approach is often referred to as the quantum-chemical (or QM) cluster approach.[11]

The other approach is to use some form of embedding, whereby the "important" atoms within the active site, essentially the same set of atoms as mentioned above, are treated with a quantum-chemical method, while the less important "environment" of the system is described using a simpler level of theory, such as molecular mechanics, or a lower level of quantum-chemical treatment. We will refer here to all such methods as "QM/MM" methods and will not attempt to discuss the various advantages or disadvantages of the multiple variants of this family of methods.[12] We will simply note that all of these techniques offer the ability to calculate the potential energy for a given configuration of the atoms in very large models of the enzyme system, comprising typically all (or at least most) the protein, substrate and cofactor atoms as well as a significant number of surrounding solvent molecules. The computational cost is much lower than for a single-level quantum-chemical calculation on the same system. Especially when the environment portion of the system is treated with molecular mechanics, the computational effort is similar to that required for a quantum-chemical calculation only on the core "important" atoms, *i.e.* the QM region. As the "environment" atoms are nevertheless treated, at least approximately, this sort of QM/MM approach also offers, in principle, the advantage that the computational expense can be *lower* than in the quantum-chemical cluster approach. This is because an accurate or "converged" description can be obtained when including in the QM region a smaller number of atoms than would be needed in the cluster model.

Here we do not propose to relitigate the question of the relative accuracy of QM cluster and QM/MM methods from the point of view of the quantum-chemical treatment. We will note, though, that the description above, which could be taken to suggest that the QM/MM approach is always preferable, is overly simplistic and neglects the fact that the cluster approach may be preferable in practical terms or even in principle in many cases. Our focus here is instead on the different implications in terms of TST.

One huge advantage of the QM cluster approach over QM/MM approaches is that it is conceivable to use it while applying "simple" versions of TST. For medium-sized cluster models of an enzyme:substrate complex, there may be only a fairly limited number of local minima on the potential energy surface, and with experience and chemical insight, it may even be possible to locate the global minimum structure (or the global minimum structure within the reactant space). The full number of degrees of freedom may be smaller than is implied by the number of atoms in the model, as QM cluster approaches often involve "freezing" or fixing the coordinates of several atoms in the amino acid environment of the active site,[11] based on coordinates obtained from experimental studies such as X-ray crystallography. This freezing of atom coordinates is a reflection of the fact that the ensemble **R** of reactant-state structures must respect the boundary conditions of the full system.

In cases where it is possible in this way to locate the global minimum, and the lowest TS, then simple TST in the form of equation (1) can be used to correlate the calculated potential energy surface with expected reactivity. The free energy of the reactant state and the TS can be computed with simple statistical mechanical expressions based on the harmonic oscillator approximation or slightly modified versions thereof. The many successes of the QM cluster approach demonstrate that this approach can be quite accurate. Note that when applying TST to smaller molecular systems an analogous definition of a small "model" is also performed, certainly for reactions in solution or on liquid or solid surfaces, but even for gas-phase reactions (which do not take place in a perfectly isolated environment – it is simply the case that the "environment" atoms are in that case clearly much further away from the "core" region). With reference to equation (2), in both cases one is not neglecting the fact that the true ensemble of reactant states is vastly larger than implied by the harmonic oscillator expression for the reacting core. Instead, one assumes that integration or summation over the vast number of additional degrees of freedom can be expected to contribute an almost identical multiplicative correction factor to both the numerator and the denominator. This will be (approximately) the case whenever the interatomic interactions between the reacting core region and the (implicit, neglected) environment are (roughly) equal in strength over the ensembles of reactant and transition state structures.

If needed, the QM cluster approach can be modified to include some element of multi-configuration nature, by applying equations (2) and (3) and taking into account multiple close-lying minima for the reactant and multiple close-lying TSs. This approach is not often performed in practice, perhaps because the perceived benefit is small, with the resulting correction factor being close enough to one that it can be assumed to make a smaller contribution than other error terms, *e.g.* to the electronic structure energies.

Turning to QM/MM methods, the very first study of a biochemical reaction, focusing on the mechanism of the lysozyme enzyme,[13] used an energy minimization and simple TST picture to model reactivity. A small number of degrees of freedom of the reacting subsystem were optimized, to yield a potential energy surface in the reactant and TS regions. Many QM/MM studies continue to use similar energy minimization approaches and (explicitly or implicitly) a simple TST model for reactivity.

When optimizing multiple degrees of freedom, however, this is a hazardous procedure as there are many possible minimum energy structures for reactants, and many possible TS structures. It becomes essentially impossible to locate the global minimum. Usually, the starting coordinates will be taken from experiment, so the starting structure should be fairly close to the global minimum in some sense at least. However, given the many ways in which experimental structures are approximate,[14] and the many ways in which a structure needs to be modified prior to QM/MM calculations[15] (addition of hydrogen atoms, addition of solvent, resolution of "bad contacts", …), the starting structure will usually **not** correspond to the global minimum in the strong sense that is necessary for obtaining reliable relative energies that can be used in conjunction with simple TST. Several studies have discussed the way in which "naïve" simple TST analysis of QM/MM calculations based on energy minimization of reactant and TS strcutures can lead to highly variable results depending on the starting structure,[16] or even to nonsensical results.[17]

For these reasons, many contemporary QM/MM studies of enzyme reactivity have turned to using molecular dynamics-based free energy methods[18] to compute minimum *free energy* pathways along collective reaction coordinates.[19] To describe this approach in terms of MCTST and equations (2), (3) and (5), these can be re-expressed as follows: the net emerging rate constant k_{AB} is given as a weighted average of "individual" rate constants, with the weighting depending on the Boltzmann factors associated with the different reactant states. The individual rate constants themselves depend on "local" free energy differences between individual reactant states and individual TSs, which are "local" in the sense that they are calculated using partition functions that consider only ensembles of structures close to individual minima and saddlepoints. In QM/MM free energy methods, simulations (usually using some or other biasing method) are used to compute the mean free energy difference between the reactant state and the TS, allowing for averaging over large ensembles of structures in both states, *i.e.* going well beyond the "local" treatment just mentioned. The free energy difference is then translated into a rate constant if needed. This whole approach can be seen as an extension of eq. (2) or (5) to replace the summation over distinct minima into an integral over possible structures, that accounts for the local curvature of the potential energy surface as well as its broader multiple-minimum nature.

Free energy methods of this type are in many ways much more rigorous when applied to reactions in complex media but they also have challenges: they require extensive sampling of the potential energy surface along the

collective variables corresponding to the reaction coordinate and possible other coupled coordinates, which in turn require a very large number of evaluations of the energy and gradient of the system. In practice, when the potential energy is computed with an accurate quantum-chemical method, this means that very powerful computer systems are needed and sampling is limited to less than would be desirable. More approximate semi-empirical quantum-chemical methods are often used, which are perhaps less accurate. Also, these methods do not yield unique key structures for which the bonding properties and electronic structure can be analyzed in order to yield qualitative insight into the origin of reactivity.

For this reason, minimum energy reaction path techniques are also valuable for studying enzymatic reactivity and receive continuing attention.[20] As it is not possible in most cases to locate the global minimum or the lowest energy TS, as noted above, some care is needed in order to obtain meaningful results. In particular, it is important to check that one *does* obtain a connected pair for the TS structure and the corresponding reactant structure, *i.e.* that the two structures are connected by a minimum energy path that does not involve very extensive rearrangement of the coordinates of the environment atoms surrounding the QM region and its reactive centre. This is possible by using techniques such as the adiabatic mapping (or coordinate scan) procedure,[21] the nudged elastic band method,[22] or the growing string method.[23]

Starting from the crystal structure and any necessary system preparation steps such as molecular dynamics equilibration, a given minimum energy path leading from a given reactant minimum to its connected TS will provide a given value of the activation potential energy $\Delta E_{ij}^{\ddagger}$. This will not necessarily correlate well with the overall mean activation free energy of equation (7) for several reasons. One of these is that changes in bonding between the reactant state and the TS may mean that there is a substantial entropic contribution to the free energy of activation coming from changes in stiffness of the vibrational frequencies of the modes orthogonal to the reaction coordinate. Such an effect can in principle be treated by computing a vibrational entropy contribution to the activation free energy from a QM cluster model or using vibrational frequencies computed using a QM/MM model.

Another effect is that the obtained value of $\Delta E_{ij}^{\ddagger}$ is just one of many possible values, while $\Delta G^{\ddagger}$ depends on an average. For this reason, many computational studies involve calculation of *multiple* minimum energy paths starting from multiple starting structures, thereby yielding a set of values $\left\{\Delta E_{ij}^{\ddagger}\right\}$. A better estimate of the activation free energy can be obtained by taking some type of average over this set. A simple approach would be to take an arithmetic mean, but this gives undue importance to individual large values corresponding to less reactive pathways. In the spirit of equation (7), it would make more sense to use the analogous expression, eq. (8):

$$\Delta G^{\ddagger} \cong -RT \ln\left\{\frac{\sum_i \exp\left(\frac{-\Delta E_{ij}^{\ddagger}}{RT}\right)\exp\left(-\Delta E_i/RT\right)}{\sum_i \exp\left(\frac{-\Delta E_i}{RT}\right)}\right\} \tag{8}$$

Again, this assumes that the "local" pathways do not involve a substantial entropic component to the activation free energy. If needed, a correction can be computed by using vibrational frequencies.

In many cases, the different starting points used for studying individual reaction paths hence obtaining different $\Delta E_{ij}^{\ddagger}$ values are drawn more or less randomly from molecular dynamics simulations of the reactant state. In that case, the structures can be assumed to be equally likely in terms of their contributions to the overall reactant state partition function, which implies that the Boltzmann factors $\exp(-\Delta E_i/RT)$ can be assumed to be equally large. In that case, equation (8) can be simplified to the following form (where n is the number of sampled activation energy values):

$$\Delta G^{\ddagger} \cong -RT \ln\left\{\frac{1}{n}\sum_i^n \exp\left(\frac{-\Delta E_{ij}^{\ddagger}}{RT}\right)\right\} \tag{9}$$

This equation – illustrated in Fig. 2 – is the expression for "exponential averaging" that has been used in multiple QM/MM studies of reactivity. This expression has been used in an *ad hoc* way in many studies. It has also been pointed out by Cooper and Kästner[24] that this formula is reasonable in the framework of the Jarzynski equality, taking the values of $\Delta E_{ij}^{\ddagger}$ as being samples of the work needed to reach the TS from the reactant state.

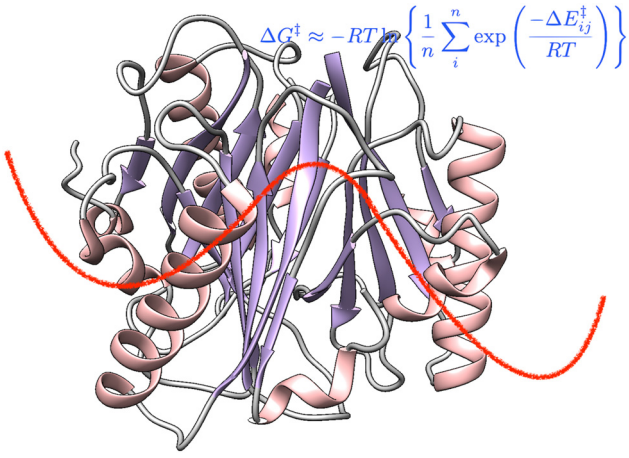

Fig. 2: The "exponential averaging" method for computing average barrier heights from multiple individual QM/MM energy pathways (eq. (9)).

Despite its approximate nature, equation (9) is widely used. Apart from the use of potential energies rather than free energies, it has been suggested that this approach to generating average barrier heights and approximate activation free energies is problematic also because it shows poor convergence with respect to the number of reaction paths that need to be sampled. Ryde stated[25] that a very large number of independent pathways will typically be needed to obtain a reasonably converged overall barrier height. This was based on a model whereby the individual energy barriers $\Delta E_{ij}^{\ddagger}$ obtained along each of the pathways are normally distributed.

This assumption of a normal distribution of the individual barrier heights seems a reasonable one, given the central limit theorem.[26] The study by Ryde[16] showed that the set of barrier heights obtained in various QM/MM studies are indeed roughly normally distributed. On the other hand, given the way individual TS structures are optimized – as maxima along a reaction path initiated from a local minimum, the barrier height for each pathway must be a positive number. This sets a minimum bound on the barrier height, such that the distribution of individual values cannot be exactly described by a gaussian curve. It is not clear, though, whether the minimum bound leads to a detectable difference from a normal distribution.

Here we reinvestigate the datasets from ref. 16 and attempt to assess how closely they conform to a normal distribution. We also perform the same analysis for a new set of barrier heights, drawn from our previous work,[27] in which we generated many QM/MM reaction pathways for the classic test case of the Claisen rearrangement catalysed by chorismate mutase.[6] For this purpose, we use standard statistical analyses to examine the way in which the individual barrier heights $\Delta E_{ij}^{\ddagger}$ are distributed.

The first such statistical analysis of the QM/MM barrier heights that can be performed assesses the extent of the deviation from a normal distribution using the Fisher-Pearson skewness (g, which provides insight into left- versus right-symmetry: a value of zero indicates a symmetric curve, while positive values indicate a longer tail to the right, and negative values indicate a tail to the left) and kurtosis (k, which is related to the importance of the tails of the distribution compared to the centre; it is equal to three for a Gaussian, larger than 3 for curves with more importance to the tails, and smaller than 3 for curves with smaller tails).[28] These may be calculated using Equations (10) and (11), respectively, where the x_i are the individual barrier heights ($\Delta E_{ij}^{\ddagger}$ from eq. (8)), μ is the arithmetic mean of the barrier heights for the given system, σ is the sample standard deviation, and N is the size of the sample.

$$g = \sum_{i=1}^{N} \frac{(x_i - \mu)^3}{N} \Big/ \sigma^3 \tag{10}$$

$$k = \sum_{i=1}^{N} \frac{(x_i - \mu)^4}{N} \Big/ \sigma^4 \tag{11}$$

Using first the datasets from ref. 16, we find that it is not possible to make a definite judgement on whether they correspond best to a symmetric Gaussian or whether they have a significant non-normal character. We find that both

metrics vary rather significantly, with g indicating both positive- and negative-skewed datasets. Test calculations using subsets of each dataset show that the values of the g and k metrics are probably determined by outliers in the datasets, which do not contain many values for each reaction, and hence the metrics are poorly defined. For these reasons it is not possible to conclude whether a Gaussian gives a reliable description of the distribution of barrier heights.

For the chorismate mutase case from our previous work (ref. 18), on the other hand, the dataset is larger and a somewhat firmer conclusion can be reached. We had obtained 117 barrier heights, by running MD simulations of the reactant state of the chorismate mutase:substrate complex (built based on the experimental structure with pdb-id:2CHT[29]) to equilibrate the system, and on the extracted MD snapshots, QM/MM reaction pathways (and therefore barrier heights) were calculated using our newly-implemented QM/MM growing string method together with DFT. Figure 3 shows a histogram of the barrier height values.

The obtained distribution is close to normal, as assumed by Ryde,[16] but both the Fisher-Pearson skewness and the kurtosis indicate some deviation from a Gaussian distribution. The skewness is slightly positive ($g > 0.0$), indicating a longer tail for higher barrier heights. This provides preliminary evidence that indeed the lower bound on individual barrier heights means that the distribution of QM/MM energy barriers obtained by careful sampling has a smaller tail at smaller values of the barrier height than at larger values. We checked whether the sample size used here (117 MD snapshots) is large enough to establish that there is indeed a positive skewness to the distribution. Taking 1000 random half-size sub-samples of the dataset, 85 % are found to yield positive skewness, while 15 % are negatively skewed, though these all have a rather small negative skewness. This suggests that the minimum energy paths for chorismate mutase are indeed likely to have a positively skewed distribution, though given the small magnitude of the skewness, further data would be needed to firmly conclude this.

To further investigate the relevance of this departure from a Gaussian distribution, the dataset has been fitted to various probability distribution functions (PDFs), using the residual sum-of-squares method as implemented in the distfit library to score the fit.[30] The PDFs tested included the Gaussian, t-distribution, gamma, beta, log-normal, generalised extreme value, Weibull, continuous uniform, Pareto, and exponential, all of which except the uniform, Pareto, and exponential distributions yielded a reasonable fit of the data. The best accuracy was obtained with a gamma distribution, as the fitted curve yielded the closest match to the μ, σ, g, and k obtained for the actual dataset. Accordingly, further analysis was done with this distribution model, as well as the Gaussian model. Plots of the best-fit gamma and Gaussian distributions are shown in Fig. 4. As can be seen, the deviation from normal distribution is quite modest, though as noted above, it is robust with respect to leaving out a portion of the data.

As was already shown in the study by Ryde,[16] with a model for the distribution of QM/MM barrier heights, it is possible to generate and compute average properties for random datasets containing many more barrier heights than are available from the raw QM/MM dataset. These can be used to test how well the averaging procedure of eq. (9) converges, *i.e.* one can compute the standard error on the mean of eq. (9) when generating random sets of barrier heights that follow the gamma or Gaussian distributions fitted to the actual QM/MM barrier heights. The results of this Monte Carlo procedure are shown in Table 1. We consider sample sizes of 10^n values of $\Delta E^{\ddagger}_{ij}$ (n = 1–7)

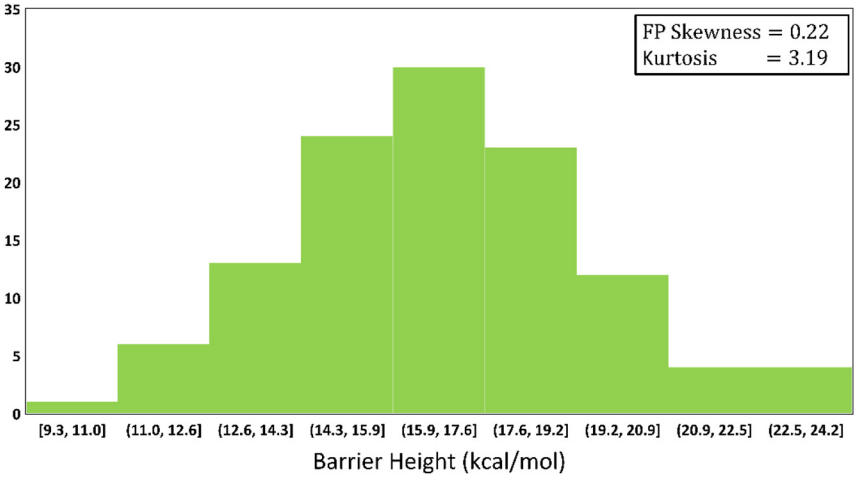

Fig. 3: Histogram of the distribution of barrier heights for the 117 snapshots. Also given is the Fisher-Pearson skewness and kurtosis for the dataset, showing a slight positive skewness and slightly leptokurtic distribution.

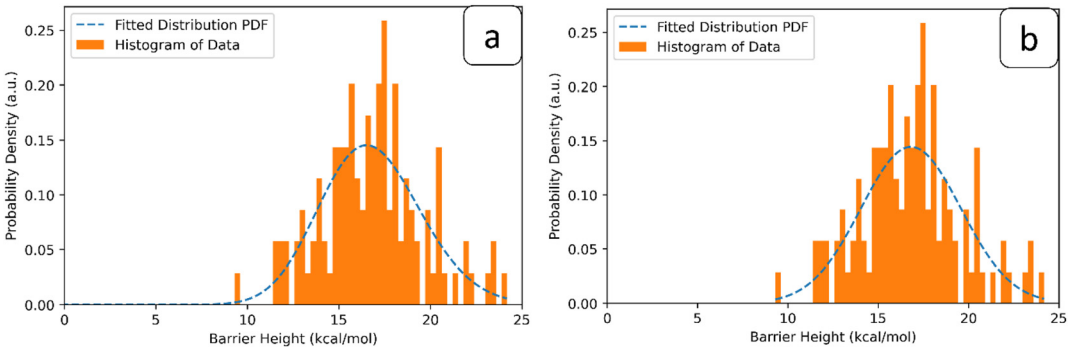

Fig. 4: Histograms of 117 calculated barrier heights for chorismate mutase with comparative analytical fits. (a) Gamma distribution fit; (b) Gaussian distribution fit.

Table 1: The standard deviation of the exponentially averaged barrier (eq. (9)) with logarithmically increasing sample size. A Monte Carlo sampling along the gamma and Gaussian distributions obtained 1000 datasets for sample sizes 10^n (n = 1–7).

Sample size *n*	**Exponentially averaged barrier height standard deviation (kcal/mol)**	
	Gamma Distribution	**Gaussian Distribution**
1	1.927	2.172
2	0.556	1.099
3	0.231	0.537
4	0.080	0.296
5	0.021	0.171
6	0.005	0.117
7	0.001	0.039

for both the gamma and Gaussian distribution, and calculate the error on the mean of eq. (9) obtained when taking 1000 random datasets of each size.

As can be seen, for the smallest sample size of just 10 values (n = 1), there is a significant error on the mean barrier for both a Gaussian and gamma distribution. For larger datasets, though, the error tends towards zero, and is smaller than the expected error due to the quantum-chemical method already with n = 2. It can be seen that the standard error is in each case slightly smaller when assuming the skewed distribution *i.e.* when using the gamma distribution. This indicates that the mean barrier height converges more rapidly when sampling from a positively skewed distribution, even in a case such as this one where the degree of the skewness is very modest. This suggests that QM/MM minimum energy pathways can indeed be used to generate estimates of the effective barrier for a reaction when using eq. (9), provided that there is some skewness and that one is able to generate sufficient QM/MM energy pathways.

Conclusions

In this perspective, we discuss the way in which transition state theory can be applied to reactivity on complicated potential energy surfaces where there are multiple minima corresponding to the reactants, and multiple different sadlle points corresponding to the transition state or bottleneck to reaction. In particular, we try to sketch how TST-based methods can in principle be applied in the context of calculations of minimum energy paths for chemical reactions taking place in biomolecular systems, while using hybrid quantum-mechanical/molecular-mechanical (QM/MM) techniques to compute the potential energy surface. For such reactions, one faces the problem that there exist so many minima on the potential energy surface that it is not possible to locate the global minimum, and one

instead obtains samples of energies for the reactant minima, and samples of energy pathways and TSs leading to products.

The paper starts by showing that the problem of multiple local minima is not unique to QM/MM studies of biomolecular reactions. Indeed, applications of TST to even quite modest-sized systems involve modifications taking multiple conformers of the reactant state into account, leading to so-called multi-conformer TST (MCTST).[9] Quantum-chemical calculations of rate coefficients for reactions using large atomistic models will typically need to use some form of MCTST in order to take into account the many conformers. The fact that many conformers need to be considered when performing QM/MM biomolecular reactivity studies is thereby not linked to the hybrid QM and MM nature of the approach, but instead simply to the large number of atoms involved in typical QM/MM studies.

QM/MM studies of reactivity based on minimum energy paths typically leads to *sets* of energy barriers $\left\{\Delta E_{ij}^{\ddagger}\right\}$ each corresponding to a different initial structure. These energy barriers can typically vary substantially, and as discussed here, MCTST provides a useful framework for obtaining an overall estimated free energy of activation for a system where many minimum energy paths have been located. The exponential averaging approach[24] of eq. (9) is shown here to be a simple adaptation of the core equation of MCTST, with some additional assumptions concerning the impact of entropy effects for minimum energy paths. This averaging approach has been criticized[25] as it may in principle require a very large number of pathways to be calculated before a converged average barrier is obtained. This problem becomes less severe if the distribution of individual barrier heights is not normal as was previously assumed in ref. 25, but is indeed skewed to have a larger tail of high-energy reaction pathways and a smaller number of low-energy barriers. Such a skewed distribution is *a priori* likely given the fact that barrier heights along minimum energy paths have a natural lower bound of zero, and in practice, a small skew in the barrier height distribution is found in a large set of minimum energy paths that we obtained in a previous study.[27] Monte Carlo study of the convergence of eq. (9) with synthetic sets of barrier heights created with the same skewness as that of the actual barrier heights of ref. 27 confirm that convergence to within an error smaller than that of the electronic structure method used in QM/MM is in practice quite rapid.

Acknowledgments: The authors thank Profs. Julianna Olah and Ulf Ryde for helpful discussions. They also acknowledge funding from KU Leuven research through grant C14/22/087.
Research ethics: Not applicable.
Informed consent: Not applicable.
Author contributions: All authors have accepted responsibility for the entire content of this manuscript and approved its submission. The text was written by both authors. The Monte Carlo analysis was performed by N.M.
Use of Large Language Models, AI and Machine Learning Tools: Non declared.
Conflict of interest: The authors state no conflict of interest.
Research funding: KULeuven research funding, grant ref. C14/22/087.
Data availability: Not applicable.

References

1. Mardirossian, N.; Head-Gordon, M. *Mol. Phys.* **2017**, *115*, 2315–2372. https://doi.org/10.1080/00268976.2017.1333644.
2. Riplinger, C.; Sandhoefer, B.; Hansen, A.; Neese, F. *J. Chem. Phys.* **2013**, *139*, 134101. https://doi.org/10.1063/1.4821834.
3. Pulay, P. *Molec. Phys.* **1969**, *17*, 197–204. https://doi.org/10.1080/00268976900100941.
4. Schlegel, H. B. *J. Comput. Chem.* **1982**, *3*, 214–218. https://doi.org/10.1002/jcc.540030212.
5. Fukui, K. *J. Phys. Chem.* **1970**, *74*, 4161–4163. https://doi.org/10.1021/j100717a029.
6. Miller, W. H.; Handy, N. C.; Adams, J. E. *J. Chem. Phys.* **1980**, *72*, 99–112. https://doi.org/10.1063/1.438959.
7. Stillinger, F. H. *Phys. Rev. E* **1999**, *59*, 48–51. https://doi.org/10.1103/PhysRevE.59.48.
8. Ess, D. H.; Wheeler, S. E.; Iafe, R. G.; Xu, L.; Çelebi-Ölçüm, N.; Houk, K. N. *Angew. Chem. Int. Ed.* **2008**, *47*, 7592–7601. https://doi.org/10.1002/anie.200800918.
9. For some examples, see Vereecken, L.; Peeters, J. *J. Chem. Phys.* **2003**, *119*, 5159–5170. https://doi.org/10.1063/1.1597479, K. H. Møller, R. V. Otkjær, N. Hyttinen, T. Kurtén, H. G. Kjaergaard, *J. Phys. Chem. A* 120, 10072–10087 (2016), https://doi.org/10.1021/acs.jpca.6b09370.

10. A very similar formalism has also been introduced to account for the kinetics of enzyme-catalyzed reactions observed at the single-molecule level, see Kou, S. C.; Cherayil, B. J.; Min, W.; English, B. P.; Xie, X. S. *J. Phys. Chem. B* **2005**, *109*, 19068–19081. https://doi.org/10.1021/jp051490q.
11. Siegbahn, P. E. M.; Himo, F. *WIREs Comp. Mol. Sci.* **2011**, *1*, 323–336. https://doi.org/10.1002/wcms.13.
12. For some reviews, see Senn, H. M.; Thiel, W. *Angew. Chem. Int. Ed.* **2009**, *48*, 1198–1229. https://doi.org/10.1002/anie.200802019; L. W. Chung, W. C. Sameera, R. Ramozzi, A. J. Page, M. Hatanaka, G. P. Petrova, T. V. Harris, X. Li, Z. Ke, F. Liu, H.-B. Li, L. Ding, K. Morokuma, *Chem. Rev.* 115, 5678–5796 (2015), https://doi.org/10.1021/cr5004419.
13. Warshel, A.; Levitt, M. *J. Mol. Biol.* **1976**, *103*, 227–249. https://doi.org/10.1016/0022-2836(76)90311-9.
14. Wlodawer, A.; Minor, W.; Dauter, Z.; Jaskolski, M. *FEBS J.* **2013**, *280*, 5705–5736. https://doi.org/10.1111/febs.12495.
15. Lonsdale, R.; Harvey, J. N.; Mulholland, A. J. *Chem. Soc. Rev.* **2012**, *41*, 3025–3038. https://doi.org/10.1039/c2cs15297e.
16. Klähn, M.; Braun-Sand, S.; Rosta, E.; Warshel, A. *J. Phys. Chem. B* **2005**, *109*, 15645–15650. https://doi.org/10.1021/jp0521757.
17. Siegbahn, P. E. M.; Borowski, T. *Faraday Discuss*. **2011**, *148*, 109–117. https://doi.org/10.1039/C004378H.
18. This is a very broad field of research, with too many methods and variants to cover in detail in this study. We cite here a few landmark methods: Bash, P. A.; Field, M. J.; Karplus, M. *J. Am. Chem. Soc.* **1987**, *109*, 8092–8094. https://doi.org/10.1021/ja00260a028.
(b) Hu, H.; Lu, Z.; Yang, W. *J. Chem. Theory Comput.* **2007**, *3*, 390–406. https://doi.org/10.1021/ct600240y.
19. For a recent example from our group, see Čivić, J.; Tuñon, I.; Harvey, J. N. *ACS Catal.* **2025**, *15*, 1684–1692. https://doi.org/10.1021/acscatal.4c06972.
20. For an example showing how both types of method continue to be be developed, see Yagi, K.; Ito, S.; Sugita, Y. *J. Phys. Chem. B* **2021**, *125*, 4701–4713. https://doi.org/10.1021/acs.jpcb.1c01862.
21. Ranaghan, K. E.; Mulholland, A. J. *Int. Rev. Phys. Chem.* **2010**, *29*, 65–133. https://doi.org/10.1080/01442350903495417.
22. Henkelman, G.; Jónsson, H. *J. Chem. Phys.* **2000**, *113*, 9978–9985. https://doi.org/10.1063/1.1323224.
23. Zimmerman, P. M. *J. Chem. Phys.* **2013**, *138*, 184102. https://doi.org/10.1063/1.4804162.
24. Cooper, A. M.; Kästner, J. *Chem. Phys. Chem.* **2014**, *15*, 3264–3269. https://doi.org/10.1002/cphc.201402382.
25. Ryde, U. *J. Chem. Theory Comput.* **2017**, *13*, 5745–5752. https://doi.org/10.1021/acs.jctc.7b00826.
26. Kwak, S. G.; Kim, J. H. *Korean J. Anesthesiol.* **2017**, *70*, 144–156. https://doi.org/10.4097/kjae.2017.70.2.144.
27. McFarlane, N. R.; Harvey, J. N. *Phys. Chem. Chem. Phys.* **2024**, *26*, 5999–6007. https://doi.org/10.1039/d3cp05772k.
28. (a) Groeneveld, R. A.; Meeden, G. *The Statistician* **1984**, *33*, 391–399. https://doi.org/10.2307/2987742.
(b) Joanes, D. N.; Gill, C. A. *The Statistician* **1998**, *47*, 183–189. https://doi.org/10.1111/1467-9884.00122.
(c) Balanda, K. P.; MacGillivray, H. L. *Am. Statistician* **1988**, *42*, 111–119. https://doi.org/10.2307/2684482.
29. Chook, Y. M.; Gray, J. V.; Ke, H.; Lipscomb, W. N. *J. Mol. Biol.* **1994**, *240*, 476–500. https://doi.org/10.1006/jmbi.1994.1462.
30. Taskesen, E., 2020. distfit is a python library for probability density fitting. (version 1.4.0). https://erdogant.github.io/distfit.