

**Manuscript version: Author's Accepted Manuscript**

The version presented in WRAP is the author's accepted manuscript and may differ from the published version or Version of Record.

**Persistent WRAP URL:**

<http://wrap.warwick.ac.uk/129842>

**How to cite:**

Please refer to published version for the most recent bibliographic citation information. If a published version is known of, the repository item page linked to above, will contain details on accessing it.

**Copyright and reuse:**

The Warwick Research Archive Portal (WRAP) makes this work by researchers of the University of Warwick available open access under the following conditions.

Copyright © and all moral rights to the version of the paper presented here belong to the individual author(s) and/or other copyright owners. To the extent reasonable and practicable the material made available in WRAP has been checked for eligibility before being made available.

Copies of full items can be used for personal research or study, educational, or not-for-profit purposes without prior permission or charge. Provided that the authors, title and full bibliographic details are credited, a hyperlink and/or URL is given for the original metadata page and the content is not changed in any way.

**Publisher's statement:**

Please refer to the repository item page, publisher's statement section, for further information.

For more information, please contact the WRAP Team at: [wrap@warwick.ac.uk](mailto:wrap@warwick.ac.uk).

Cite this: DOI: 00.0000/xxxxxxxxxx

# Fast screening of homogeneous catalysis mechanisms using graph-driven searches and approximate quantum chemistry

Christopher Robertson and Scott Habershon\*

Received Date

Accepted Date

DOI: 00.0000/xxxxxxxxxx

Computational methods for predicting multistep reaction mechanisms, such as those found in homogeneous catalysis by organometallic complexes, are rapidly emerging as powerful tools to support experimental mechanistic insight. We have recently shown how a graph-driven sampling scheme can be successfully used to propose a series candidate reaction mechanisms for nanoparticle catalysis; however, identifying the most-likely reaction mechanism amongst this candidate set in an efficient scheme remains a challenge. Here, we show how simple descriptors for each reaction path, calculated using quick semi-empirical quantum chemistry, enable identification of the mechanism, but only if one considers both thermodynamic and kinetic parameters of proposed reaction mechanisms. Successful application to cobalt-catalysed alkene hydroformylation is used to benchmark this strategy, and provides insight into remaining algorithmic challenges.

The last few years have seen rapid development of a number of computational methods which are aimed at automated discovery of complex chemical reaction networks.<sup>1–12</sup> The broad concept of this field is to integrate novel algorithms which can: (i) propose (at least in principle) the full set of chemical reactions for a given set of molecular reactants, (ii) calculate thermodynamic and kinetic properties for each reaction, and (iii) analyse the emergent chemical kinetic network using, for example, direct kinetic simulations<sup>13–16</sup> or methods such as network pruning<sup>5,17</sup> in order to predict the macroscopic reaction outcomes of experiments. These reaction-discovery-based simulations are therefore highly appealing in enabling a direct connection between mechanistic chemistry, *ab initio* quantum chemistry, and macroscopic observables such as rate laws and product selectivities.

Emerging directions in this field include the development of novel graph-based schemes in order to quickly postulate and categorize chemical reactions in order to build-up large-scale chemical reaction networks,<sup>3,6,11,17–20</sup> development of accelerated sampling schemes to drive chemical reactions to populate reaction networks,<sup>7</sup> and the integration of such reaction-sampling schemes with efficient and accurate strategies for finding transition-state structures in order to underpin reaction-rate calculations.<sup>1,21,22</sup> In parallel with these atomistic simulation developments, the last decade or so has seen a large explosion in the application of artificial intelligence and machine-learning techniques, which aim to mine large experimental datasets for chem-

ical reactivity patterns in order to predict new synthesis routes for organic molecules; for example, neural networks have been shown capable of predicting reaction outcomes, typically based on training against the USPTO database.<sup>23–25</sup> All together, the rapidly-expanding field of reaction discovery and prediction is set to become an increasingly powerful supplement to traditional experimental synthesis in the coming years.

In our own recent work in this area,<sup>3,6,11</sup> we have begun to make progress in addressing a slightly different, albeit related, challenge. Initially, we developed a Hamiltonian-based reaction discovery scheme which treats a given reaction-path as a dynamic object, allowing conformation sampling of reaction-path-space. A novel aspect of this Hamiltonian sampling scheme was the introduction of a graph-restraining potential (GRP) which enforced a well-defined bond connectivity matrix on the reaction-path end-point configurations; by introducing random changes to the reactant and product connectivity matrices, corresponding to possible chemical reactions, the natural dynamics of the Hamiltonian is then such that reaction-path configurations for the new set of reactant/product bonding matrices are sampled. In other words, by combining Hamiltonian reaction-path sampling with GRP-enforced stochastic changes to the end-point bonding, this scheme enables generation of complex chemical reaction networks, while simultaneously providing the configurational information required for further analysis by *ab initio* thermodynamics or rate calculations.

Building on this scheme, we have recently demonstrated that similar ideas can be adapted to enable generation of multi-step mechanisms connecting user-defined reactant and product con-

figurations;<sup>6</sup> this double-ended graph-driven sampling (GDS) scheme has been successfully demonstrated in predicting reaction mechanisms for the water-gas shift reaction, carbon monoxide oxidation and hexane aromatization, all occurring on a platinum nanoparticle. Here, the key distinction between the reaction discovery methods noted above and the approach developed herein is the fact that our GDS scheme is specifically aimed at searching for reaction mechanisms which definitively connect a user-defined set of reactants and products, rather than aiming at unguided (or open-ended) reaction discovery. As described below, the search for a mechanism connecting well-defined end-points is treated as a problem in optimization: one seeks the sequence of chemically-allowed bonding changes which transform the connectivity matrix (CM) of the reactants into that of the products. The sequence of CM updates can then be transformed into corresponding molecular structures using the concept of the GRP, after which *ab initio* quantum chemical calculations can be used to calculate thermodynamic and kinetic parameters for each individual reaction-step comprising the entire reaction mechanism; as a result, our scheme provides a direct way of discovering, and assessing the suitability of, reaction mechanisms in an *ab initio* manner.

Our GDS scheme for finding reaction mechanisms does not directly require *ab initio* quantum chemistry calculations, instead relying on manipulation of CMs followed by optimization under the GRP in order to generate atomic coordinates. However, even though our double-ended mechanism search scheme can successfully locate reaction mechanisms connecting input reactants and products, there are of course many such possible mechanisms: so how can one go about trying to filter out the “most likely” mechanism? This is the key problem addressed in this Article.

Clearly, in order to distinguish which of the (potentially many) GDS-generated reaction mechanisms are the most likely mechanism, *ab initio* electronic structure calculations become essential. In an ideal scenario, one would take each GDS-generated reaction mechanism and calculate the reaction free energy change and reaction free energy barrier for each elementary step comprising the multi-step mechanism. Calculating the reaction free energy change is (often) relatively straightforward, particularly if one adopts the usual rigid-rotor/harmonic oscillator approximations for the molecular partition function;<sup>26–29</sup> however, this approach requires geometry optimization and evaluation of the Hessian matrix, which may both be time-consuming if a high-level of *ab initio* theory is demanded. However, the real bottle-neck in these hypothesized screening calculations would be the evaluation of the activation free energy barriers; in general, this requires a multi-stage process which might encompass a nudged elastic band (NEB) calculation,<sup>30–33</sup> followed by a transition-state (TS) search. Despite ongoing development of TS structure-finding algorithms,<sup>1,22,34,35</sup> such calculations remain a challenge to any automated scheme; in addition, if one must perform such TS-finding calculations for *all* elementary steps for all proposed reaction mechanisms, the computational burden would be enormous.

Instead, the aim of this paper is to investigate the extent to which one can screen for the “most likely” reaction mechanism amongst a large number of possible mechanisms using only ap-

proximate quantum chemical methods and seeking to avoid direct TS searches. In particular, we provide the first proof-of-concept results showing that our GDS-based reaction discovery scheme, when combined with approximate quantum chemical calculations, enables direct identification of the reaction mechanism of a homogeneous catalytic cycle. We show that by generating a large number of candidate multi-step reaction mechanisms, and subsequently screening these based only on calculations of relative energetics of reaction intermediates, one can quickly identify a small number of plausible mechanism candidates. We then show that further calculations of kinetic properties, namely approximate activation barriers for each elementary reaction step in the candidate mechanism, enable unique identification of the most plausible reaction mechanism. In particular, we find that this highly-automated reaction-mechanism-finding scheme can auto-discover the well-known Heck-Breslow mechanism for hydroformylation of ethene by  $\text{HCo}(\text{CO})_4$ . The methodology outlined here is therefore demonstrated as an exciting computational tool for mechanism discovery and, ultimately, catalyst design, as discussed below.

## 1 Theory

### 1.1 Double-ended graph-driven search

We have recently reported an automated computational scheme for proposing candidate reaction mechanisms for multi-step chemical reaction mechanisms which takes the viewpoint of chemical reactions as updates of CMs. Our GDS scheme has been described previously,<sup>6</sup> so we only give the key details here.

First, the input coordinates are required for the reactants  $\mathbf{r}^R$  and products  $\mathbf{r}^P$ , noting that each system must contain the same number of atoms  $n$ . These coordinates are then used to generate the reactant and product CMs,  $\mathbf{G}^R$  and  $\mathbf{G}^P$  respectively. The CMs  $\mathbf{G}$  are  $n \times n$  matrices with elements given by

$$G_{ij} = \begin{cases} 1 & \text{if } r_{ij} < r_{ij}^{cut}, \\ 0 & \text{otherwise.} \end{cases} \quad (1)$$

We note that the definition of Eq. 1 does not rely on the type of bonding (e.g. single, double, triple). In Eq. 1,  $r_{ij}^{cut}$  is a distance cut-off value which defines whether or not two atoms  $i$  and  $j$  are bonded; this cut-off is typically defined as

$$r_{ij}^{cut} = \gamma(R_i + R_j), \quad (2)$$

where  $R_i$  and  $R_j$  are approximate covalent radii for the element-types of atoms  $i$  and  $j$ , and  $\gamma$  is a parameter which allows for some chemical variation in bonding definitions, with a typical value  $\gamma = 1.1$ . In the simulations reported below, we use  $R_{\text{Co}} = 1.52 \text{ \AA}$ ,  $R_{\text{C}} = 0.72 \text{ \AA}$ ,  $R_{\text{O}} = 0.62 \text{ \AA}$ , and  $R_{\text{H}} = 0.4 \text{ \AA}$ , noting that our experience to date suggests that the precise values of these constants is not too important as long as they fall within typical estimated covalent radii. At this point, we note that, in keeping with many other graph-based reaction discovery tools, our focus is in identifying the set of bond-changes (or elementary reaction steps) which connect reactants and products. However, we note that the approach outlined here can also be modified to account

for the formation of van der Waals intermediates too, using the same method developed by Peláez and coworkers,<sup>36</sup> which re-cast the CM in a block-diagonal form separating out bonding for discrete molecular species and their intermolecular interactions.

We next define the allowed library of chemical reactions which *might* occur in our system. In our approach, building on previous graph-based reaction discovery systems,<sup>5,37</sup> we define a series of *reaction classes*; each reaction class is defined as a pair of  $m \times m$  matrices which define the bonding pattern for  $m$  atoms before and after reaction. Noting that two-atom, three-atom and four-atom reactions are by far the most common in typical chemical systems, the corresponding size of each reaction class is typically  $2 \leq m \leq 4$ . We note that the set of  $m$ -atom indices  $\mathbf{I}$  which participate in a reaction does not have to be defined individually for every possible reaction; instead, in our reaction discovery approach,  $\mathbf{I}$  is treated as set of parameters to vary, as described below. Throughout this Article, we use  $\mathbf{R}^i(\mathbf{I})$  to indicate reaction class  $i$  operating on the set of atomic indices  $\mathbf{I}$ .

To clarify the concept of reaction classes, Fig. 1 shows two examples of reaction classes which are used in our simulations below. In Fig. 1(A), we show a 2-atom dissociation reaction applied to the atomic indices  $\mathbf{I} = (1, 2)$ , and also to the indices  $\mathbf{I} = (1, 3)$ ; note that, in these two cases, the reaction class is the same but the atomic indices  $\mathbf{I}$  are different. However, by using reaction classes, we do not need to individually define each possible reaction independently. Furthermore, we note that the definition of reaction classes is such that the bonding pattern for the atomic indices  $\mathbf{I}$  is defined both before and after reaction. As a result, we note that our definition of reaction classes places constraints on which atoms in a given molecular system can react; for example, the dissociation reaction in Fig. 1(A) can only occur for initially-bonded atoms. This aspect seems like a triviality now, but will be important when searching for reaction mechanisms, as described below. As a further example, Fig. 1(B) shows a 3-atom insertion reaction applied to example indices  $\mathbf{I} = (1, 2, 3)$ ; again, we note that only the generic reaction class must be initially defined here, with the indices  $\mathbf{I}$  appearing as optimization parameters below.

Given the definition of the input reactant and product CMs,  $\mathbf{G}^R$  and  $\mathbf{G}^P$ , as well as definitions of the reaction classes  $\mathbf{R}$ , our aim is then to search for the sequence of chemical reactions which transform  $\mathbf{G}^R$  into  $\mathbf{G}^P$ , while simultaneously also constraining this search to only those "chemically sensible" transformations. In other words, we want to identify the sequence of reactions and associated atomic indices such that

$$\mathbf{G}^P = \mathbf{G}^R + \sum_{i=1}^{n_r} \mathbf{R}^{k(i)}(\mathbf{I}_i). \quad (3)$$

Here,  $n_r$  is the total number of allowed reaction steps in a given proposed mechanism,  $k(i)$  is the reaction class of the  $i$ -th reaction step, and  $\mathbf{I}_i$  is the corresponding set of atomic indices.

On the basis of Eq. 3, the search for a mechanism connecting  $\mathbf{G}^R$  to  $\mathbf{G}^P$  can now be viewed as an optimization problem. If we have a trial sequence comprising  $n_r$  reaction classes and atomic indices, namely  $[\mathbf{R}^{k(1)}(\mathbf{I}_1), \mathbf{R}^{k(2)}(\mathbf{I}_2), \dots, \mathbf{R}^{k(n_r)}(\mathbf{I}_{n_r})]$ , then application of Eq. 3 gives a corresponding product graph  $\tilde{\mathbf{G}}^P$ , and the

error associated with the trial reaction mechanism can be quantified by the error function  $F$  given by

$$F = \sum_{i < j}^n \left( \tilde{G}_{ij}^P - G_{ij}^P \right)^2, \quad (4)$$

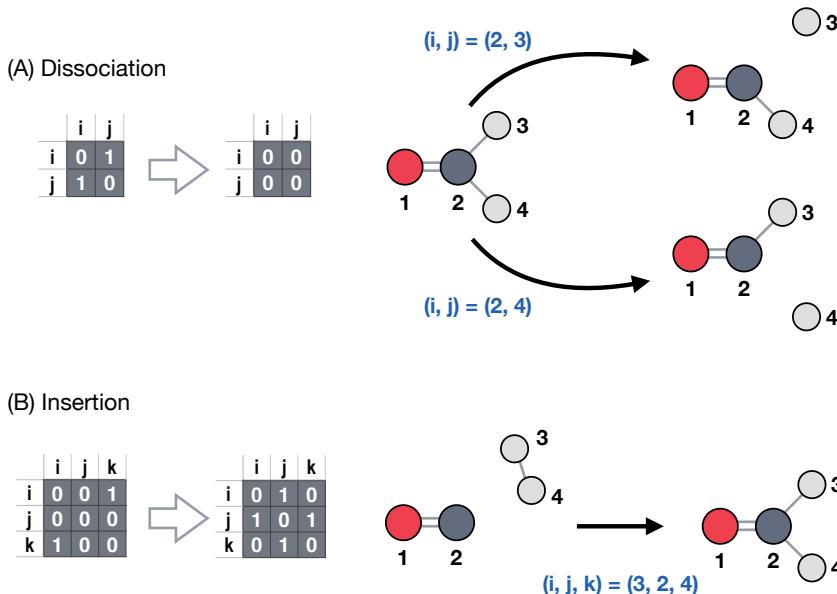
where  $G_{ij}$  is the  $(i, j)$  matrix element of  $\mathbf{G}$ , and similarly  $\tilde{G}_{ij}^P$  is the  $(i, j)$  matrix element of  $\tilde{\mathbf{G}}$ .

To perform the search for a mechanism which obeys Eq. 3, we use a simulated annealing (SA) procedure (linearly cooling from an arbitrary temperature  $T_{\text{init}}$ ), using  $F$  as an effective energy function, in order to find a candidate mechanism with  $F = 0$ . Our SA algorithm uses simple moves at each iteration which modify either the atomic indices  $\mathbf{I}_i$  at a randomly selected intermediate reaction  $i$ , or modifies both the reaction class  $\mathbf{R}^{k(i)}$  and indices  $\mathbf{I}_i$  for intermediate reaction  $i$ . After a trial update, the new error function  $F_{\text{new}}$  is evaluated using Eqs. 3 and 4, with the new candidate mechanism being accepted or rejected based on the standard Metropolis criterion using the current temperature. This SA search is continued until a mechanism with  $F = 0$  is obtained, or a maximum number of iterations is reached.

## 1.2 Chemical constraints

At this point, it is worth noting several factors which have an important impact on the success of our graph-based reaction mechanism search. First, we note that our reaction class definitions are such that they act only on sets of atoms which obey a given bonding pattern, as highlighted in Fig. 1. By ensuring that our reaction classes available to the SA search algorithm represent "chemically sensible" reactions, this constraint on the reactive atom set similarly ensures that we limit our search to only "chemically sensible" reactions. Second, we note that it is straightforward to include common chemical valence constraints in our SA search. Most importantly, the user can define allowed valence ranges for each atom type in the simulation, such as "carbon atom valence must be between one and four" or "hydrogen atom valence must be one". If any trial reaction mechanism violates one of these constraints during the SA search, the corresponding trial mechanism is assigned an arbitrarily-high value for  $F$  to ensure that the trial move is rejected. In this way, the SA search can be easily limited to consider only the sub-set of chemical reaction mechanisms in which atomic valences lie within well-known ranges (although, of course, one can always remove this constraint to explore more exotic mechanisms if desired).

As a final point, we also note that our SA scheme is consistent with the idea of defining *active* and *inactive* atoms as a way of accelerating the mechanism search; a similar idea has been exploited by Kim and others in previous graph-based schemes. For example, in the case of organometallic complexes, one might define the organic ligands of an organometallic complex as being *inactive*; in cases where such ligands control the steric and electronic properties of the reactive metal site, this might be an entirely valid assumption. However, in the calculations below, we do not explicitly define ligands as being inactive; instead, we impose the related constraint that all reaction steps must involve the metal centre of our considered system as one of the atomic indices



**Fig. 1** Illustrative demonstrations of reaction class definitions and reactive indices. (A) shows the reactant and product connectivity matrices for a reaction class corresponding to atomic dissociation, while the right-hand side shows the application of this reaction class to different atomic indices  $(i, j)$ . (B) shows an example of a three-atom reaction class, namely insertion, as applied to atoms  $(i, j, k) = (3, 2, 4)$ . Note that, in each case, the reactant connectivity matrix constrains which atomic indices might be considered for reaction; for example, in (A), application to atomic indices  $(1, 3)$  would not be considered as a viable reaction for the starting molecule because atoms 1 and 3 are not bonded, as required by the reactant connectivity matrix for this reaction class.

for each reaction. The underlying assumption is, of course, that the reaction is catalysed by the metal atom; however, *all* atoms in our SA search results presented here are active.

These chemical constraints on valences and reactive atoms can always be removed, but the effect in preliminary tests is that the number of possible reaction mechanisms which can be generated by our SA search increases enormously, with most mechanisms involving chemical structures which would simply be thermodynamically or kinetically inaccessible under normal reaction conditions. As a result, chemical constraints are an important feature of graph-based heuristic schemes; however, we note here that imposing common valence constraints and restraining reactions to involve the metal centre represent very weak constraints indeed.

### 1.3 Molecular structure generation

After a reaction mechanism with  $F = 0$  has been identified by our SA search, the final task is to generate molecular structures (and, possibly, initial reaction paths) for each of the  $n_r$  intermediate elementary steps in the mechanism. These molecular structures can then be used in further analysis of thermodynamic and kinetic properties using quantum chemical calculations.

To generate molecular structures, we use the concept of the GRP, as introduced in our previous work. The GRP is a simple analytical PES which depends on both a set of atomic coordinates  $\mathbf{r}$  and a CM  $\mathbf{G}$ , and is constructed such that it is a minimum only when the bonding pattern encoded in the atomic coordinates exactly matches that in  $\mathbf{G}$ . In other words, the GRP imposes the

connectivity pattern in  $\mathbf{G}$  onto the set of atomic coordinates  $\mathbf{r}$ .

The functional form of the GRP is somewhat arbitrary, and is given here as

$$W(\mathbf{r}, \mathbf{G}) = \sum_{j > i} \left[ \delta(G_{ij} - 1)f_1(\mathbf{r}) + \delta(G_{ij})f_2(\mathbf{r}) \right] + V_{mol}(\mathbf{r}, \mathbf{G}), \quad (5)$$

where

$$f_1(\mathbf{r}) = H(r_{ij}^{min} - r_{ij})\sigma_1(r_{ij}^{min} - r_{ij})^2 + H(r_{ij} - r_{ij}^{max})\sigma_1(r_{ij}^{max} - r_{ij})^2 \quad (6)$$

and

$$f_2(\mathbf{r}) = \sigma_2 e^{-r_{ij}^2/(2\sigma_3^2)}. \quad (7)$$

The summation in Eq. 5 runs over all pairs of atoms, the “delta” function is defined as

$$\delta(x) = \begin{cases} 1 & \text{if } x = 0, \\ 0 & \text{otherwise,} \end{cases} \quad (8)$$

and the Heaviside step function is defined as

$$H(x) = \begin{cases} 0 & \text{if } x < 0, \\ 1 & \text{if } x > 0. \end{cases} \quad (9)$$

From these definitions, we see that the first term in square brackets in Eq. 5 only acts between *bonded* pairs of atoms, and acts as a harmonic restraint term which ensures that the bond lengths of these bonded pairs remains between the fixed limits  $r_{ij}^{min}$  and  $r_{ij}^{max}$ . Similarly, the second term, involving the  $f_2(\mathbf{r})$  is a simple

repulsive interaction which only acts between non-bonded pairs of atoms, making sure that these non-bonding pairs simply stay apart from each other. The parameters  $\sigma_1$ ,  $\sigma_2$  and  $\sigma_3$  are user-defined parameters with values of  $0.05 E_h a_0^{-2}$ ,  $0.03 E_h$  and  $2.2 a_0$ , respectively, in the simulations performed in this Article. Similarly, the parameters  $r_{ij}^{min}$  and  $r_{ij}^{max}$  are minimum and maximum appropriate bonding ranges for each atom pair; here, we simply set these to be  $r_{ij}^{cut} \pm 0.1 a_0$  respectively.

Supplementing the pairwise additive terms in Eq. 5 is an intermolecular term which operates between distinct molecular species to simply ensure that they are “kept apart” from each other. This  $V_{mol}(\mathbf{r}, \mathbf{G})$  term has the following form:

$$V_{mol}(\mathbf{r}, \mathbf{G}) = \sum_{J>I} [H(R^{min} - R_{IJ})\sigma_4(R^{min} - R_{IJ})^2], \quad (10)$$

where  $R_{IJ}$  is the distance between the centers-of-mass of two molecules  $I$  and  $J$ ,  $R^{min}$  is a user-defined minimum separation distance between any pair of molecules (typically  $10 \text{ \AA}$ ), and  $\sigma_4 = 0.03 E_h a_0^{-2}$ .

Starting from some arbitrary atomic coordinates  $\mathbf{r}$ , and with a target graph  $\mathbf{G}$ , optimization of the GRP with respect to  $\mathbf{r}$  will force all atoms to move to adopt positions such that the connectivity pattern of  $\mathbf{r}$  matches that of  $\mathbf{G}$ . This procedure can therefore be used with our SA search scheme in order to generate molecular structure representing each reaction intermediate along the string of  $n_r$  reactions for any candidate mechanism with  $F = 0$ . Starting from the initial input coordinates for the reactants,  $\mathbf{r}^R$ , as well as the input reactant CM  $\mathbf{G}^R$ , we then apply the first reaction in the mechanism to  $\mathbf{G}^R$  in order to generate a new CM  $\mathbf{G}^1$  representing the outcome of the first reaction-step. Optimization of the GRP  $W(\mathbf{r}, \mathbf{G}^1)$  starting from the coordinates  $\mathbf{r}^R$  then produces a molecular structure  $\mathbf{r}^1$  which is representative of  $\mathbf{G}^1$ . This procedure can be repeated along the reaction steps to generate coordinates for all steps  $\mathbf{r}^k$  by optimizing under the action of  $W(\mathbf{r}, \mathbf{G}^k)$  starting from the coordinates generated in the previous step  $\mathbf{r}^{k-1}$ . The result is that the sequence of CM updates can be converted into a sequence of atomic coordinates; further quantum chemical calculations can then be used to characterize the thermodynamic and kinetic parameters of each individual reaction step to build up a picture of the full mechanism.

#### 1.4 Quantum chemistry calculations

As discussed below, electronic structure calculations are used in two contexts within our simulations. First, once a given GDS simulation has successfully found a reaction mechanism connecting reactants and products, resulting in generation of molecular structures for all intermediates, we routinely perform geometry optimization of the  $n_r$  reaction intermediates to generate representative stationary points (and relative energies) along the full reaction coordinate. Second, as described below, for each elementary reaction comprising the full reaction mechanism, we can also perform MEP searches using NEB if desired. In principle, the geometry optimization and NEB calculations can be performed with any *ab initio* or semi-empirical electronic method which adequately describes the system of interest and is computationally

feasible; in the present work, we use a semi-empirical method as described below.

However, for the homogeneous organometallic catalytic cycles which are a major interest in our work, performing geometry optimizations at a reasonable level of theory (e.g. DFT with polarized basis sets) is extremely time-consuming. Furthermore, performing NEB calculations for all intermediate reaction-paths generated in all GDS simulations is clearly also too computationally-demanding if one is interested in screening multiple candidate reaction mechanisms. As a result, our approach is to instead employ more approximate methods for calculating molecular energies and forces in order to rapidly screen reaction mechanisms as a first pass; the underlying assumption is that the energetics (*i.e.* reaction energy changes and barriers) given by these approximate methods are at least proportional to those which would be given by more accurate methods such as DFT. Depending upon the details of the system, this is of course not guaranteed; however, for the case of the hydroformylation reaction considered herein, our results below suggest that this assumption is appropriate.

All calculations of molecular energies performed here used the self-consistent charge density-functional tight-binding (SCC-DFTB) methodology, as implemented in *DFTB+*.<sup>38,39</sup> Our previous investigations using this approach have shown that SCC-DFTB gives reasonably accurate molecular structures and relative energetics;<sup>3</sup> for example, the SCC-DFTB parameter set used in this work was demonstrated to exhibit qualitative accuracy for a series of organometallic complexes when compared to benchmark DFT B3LYP calculations.<sup>39</sup> Most importantly, however, is the fact that SCC-DFTB calculations are very fast compared to DFT calculations, meaning that they are compatible with our goals of developing a rapid screening approach suitable for predicting catalytic reaction mechanisms.

As a final point, we note that the suitability of SCC-DFTB in the present system is a little fortunate; given the semi-empirical nature of this energy calculation approach, it is of course not guaranteed that the requisite accuracy or energy-ordering in our calculations should emerge. However, we know from previous experience that SCC-DFTB is a suitable description of the reaction dynamics in the hydroformylation system studied below. In addition, we emphasise that there is nothing in our GDS methodology which is tied to any particular energy calculation method; instead, our GDS scheme is a broadly applicable tool which can be used in conjunction with any state-of-the-art energy calculation method which is reasonably applicable to the chemical system under study.

## 2 Application

We now turn to the main aim of this Article, namely addressing the question of whether or not our GDS reaction mechanism generation scheme, when combined with fast approximate quantum chemical calculations, can enable us to unambiguously identify a homogeneously-catalyzed reaction mechanism. For this, we consider the hydroformylation reaction (or ‘oxo’ process) as a prototypical example of a ‘known’ mechanism

## 2.1 Cobalt-catalyzed hydroformylation

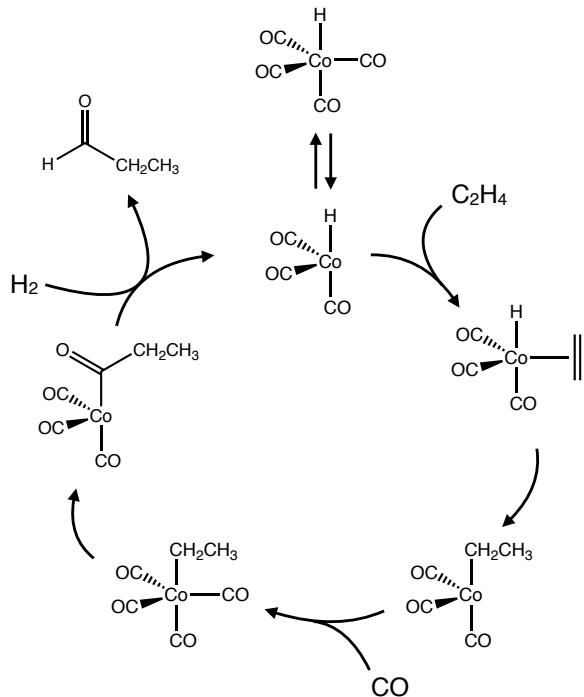
The hydroformylation of ethene by the complex  $\text{HCo}(\text{CO})_4$  is well-studied from both experimental and computational viewpoints.<sup>12,40–47</sup> As shown in Fig. 2, the accepted mechanism comprises six key steps, namely: (i) carbon monoxide dissociation from  $\text{HCo}(\text{CO})_4$  to form the catalytically active  $\text{HCo}(\text{CO})_3$  species, (ii) coordination of alkene, (iii) Alkene insertion into the Co-H bond to form a cobalt alkyl species, (iv) coordination of carbon monoxide, (v) Insertion of carbon monoxide into the Co-C bond, (vi) Oxidative addition of  $\text{H}_2$ , and (vii) Reductive elimination of the aldehyde product and reformation of the active catalyst. The rate law of cobalt-catalyzed hydroformylation has been studied experimentally,<sup>43–45,47</sup> where the necessity of the carbon monoxide dissociation step from  $\text{HCo}(\text{CO})_4$  is supported by the finding that reaction rate decreases as carbon monoxide partial pressure increases. In addition, Harvey *et al* have shown how steady-state approximations applied to a kinetic model of propene hydroformylation also support the experimental rate law and its inverse carbon monoxide pressure dependence.<sup>41</sup>

As a result of the detailed level of experimental and theoretical insight into the Heck-Breslow alkene hydroformylation mechanism, this catalytic cycle has served as an important benchmark problem in the development of a number of recent reaction discovery methods, including both graph-based strategies and methods based on automatic searches for transition-states.<sup>3,12,19,48</sup> For example, by combining an earlier version of our GDS scheme with DFT calculations and microkinetic modelling, we were able to successfully reproduce the key features of the experimental rate law for hydroformylation, notably the well-known inverse dependence on the partial pressure of carbon monoxide.<sup>3</sup> Using MD-based transition-state searches and graph-based characterization of reaction products, Martinez-Nunez and coworkers demonstrated an alternative methodology which was also shown capable of reproducing experimental kinetics,<sup>48</sup> while recent heuristics-based work by Kim and coworkers also showed capable of capturing the key details of hydroformylation.<sup>19</sup>

As such, it is clear that hydroformylation provides a key benchmark test for any automated reaction discovery tool; in this paper, we use our new double-ended GDS scheme to assess whether it is capable of picking out the accepted catalytic mechanism of cobalt-catalyzed hydroformylation, and whether approximate quantum chemistry is suitable in guiding the identification of the mechanism amongst the many possibilities generated by our simulation approach.

## 2.2 Mechanism identification

In order to investigate whether our double-ended GDS scheme, in combination with fast approximate quantum chemistry calculations, can correctly identify the “correct” ethene hydroformylation mechanism (*i.e.* hopefully the Heck-Breslow scheme of Fig. 2), we generated molecular models for typical reactant and product configurations. Here, the reactant configuration comprised  $\text{C}_2\text{H}_4 + \text{CO} + \text{H}_2 + \text{HCo}(\text{CO})_4$  and the product configuration comprised  $\text{HCOCH}_2\text{CH}_3 + \text{HCo}(\text{CO})_4$ . In the initial molecular configurations for reactants and products, the individual molecules



**Fig. 2** The Heck-Breslow catalytic cycle for  $\text{HCo}(\text{CO})_4$ -catalyzed hydroformylation. Dissociation of CO from  $\text{HCo}(\text{CO})_4$  gives the active catalyst  $\text{HCo}(\text{CO})_3$ , which then undergoes addition of alkene and insertion into the Co–H bond. Further addition and insertion of CO, followed by oxidative addition of  $\text{H}_2$ , ultimately leads to generation of aldehyde product and reformation of the catalyst.

were well-separated; an initial geometry optimization using DFTB was then performed and the resulting configurations were used as the target mechanism end-points for GDS.

The set of reaction classes used in our SA searches for mechanisms is shown in Table 1. The set of eight reaction classes includes generic association/dissociation reactions, insertions and three-atom rearrangements, as well as four-atom rearrangement reactions; as shown below, this set of reactions is broad enough that a large number of possible reaction mechanisms connecting reactants and products for the hydroformylation reaction can be readily generated. We note that the only reactivity constraint imposed during our SA search is that one of the atoms must be the Cobalt atom, a relatively weak assumption based on well-known organometallic chemistry. Table 1 also shows the atomic valence constraints used in our SA searches; if any CM updates lead to atoms which disobey one of these atomic valence constraints, the move will be rejected. Again, we note that these allowed valence ranges are sufficiently generic and broad that one can find a large number of possible reaction mechanisms. We note that the valence ranges chosen here are a broad as they can be without allowing the generation of chemically improbable species, such as ‘bare’ hydrogen (with a valence of zero). As such, the success rates of our simulations should be viewed as the ‘worst case scenario?’, and we expect the success rates for finding reaction mechanisms would improve as one places more constraints on

the allowed valence ranges, thereby narrowing the mechanism search-space.

**Table 1** (A) Reaction classes used in graph-drive simulated annealing or reaction mechanisms; note that the inverse of each reaction is also included as a possible reaction class. (B) Allowed atomic valence ranges during simulated annealing searches.

(A) Reaction classes	
Reactants	Products
$A - B$	$A + B$
$A - B + C$	$A - C - B$
$A - B - C$	$B - A - C$
$A - B - C - D$	$D - A - B - C$

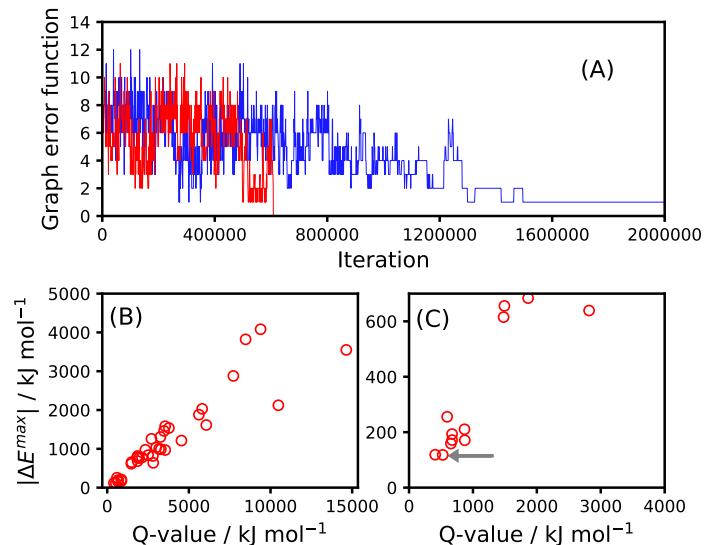
(B) Atomic valence ranges	
Atom type	Valence ( $v$ ) range
C	$1 \leq v \leq 4$
H	$1 \leq v \leq 1$
O	$1 \leq v \leq 2$
Co	$4 \leq v \leq 6$

We performed 50 independent GDS simulations, running SA of the graph error function for a maximum of  $2 \times 10^6$  iterations. At each iteration, our SA algorithm attempts to modify either (i) the atoms involved in a randomly-selected reaction step, or (ii) both the atoms and the reaction class of a randomly-selected reaction step. In all calculations, the number of reactions in each generated mechanism was  $n_r = 10$ ; however, we note that “null” reactions are allowed in our GDS scheme, such that  $n_r$  represents the maximum number of active reactions which could be used in each reaction mechanism. This value of  $n_r$  is sufficiently large that it should enable a wide range of reaction mechanisms to be generated (including, potentially, the mechanism of Fig. 2).

Of the 50 GDS simulations, 47 located a reaction mechanism (*i.e.* sequence of  $n_r$  reaction steps) which led from reactant to product CMs. Figure 3(A) shows the progression of the graph error function  $F$  in two different representative GDS simulations, one successful and one unsuccessful. Both start with  $F \approx 10$  for the initial sequence of  $n_r = 10$  reactions and, as the SA simulation proceeds, the graph error function fluctuates but generally decreases as expected. After around  $600 \times 10^3$  iterations, one of the calculations falls into a minimum on the  $F$  hypersurface with  $F = 0$ , demonstrating that a mechanism has been located which successfully leads to generation of the target product CM after  $n_r = 10$  reaction steps. In the case of the other calculation, the graph error function converges to  $F = 1$  after around  $1.5 \times 10^6$  SA iterations; this calculation is clearly trapped in a local minimum with a single incorrect CM element after  $n_r = 10$  reaction steps. However, the fact that more than 95 % of our GDS calculations managed to locate a mechanism highlights the overall simplicity and efficiency of our reaction-mechanism-finding scheme.

To further investigate the reaction mechanisms proposed by our GDS scheme, we have performed a comparative analysis of the 47 successful reaction-mechanism searches; the results are shown in Fig. 4. In Fig. 4(A), we show a color-coded diagram illustrating the sequence of reactions taken in each of the 47 mechanisms, with each color corresponding to one of the allowed elementary reaction-types shown in Table 1. A simple visual comparison of

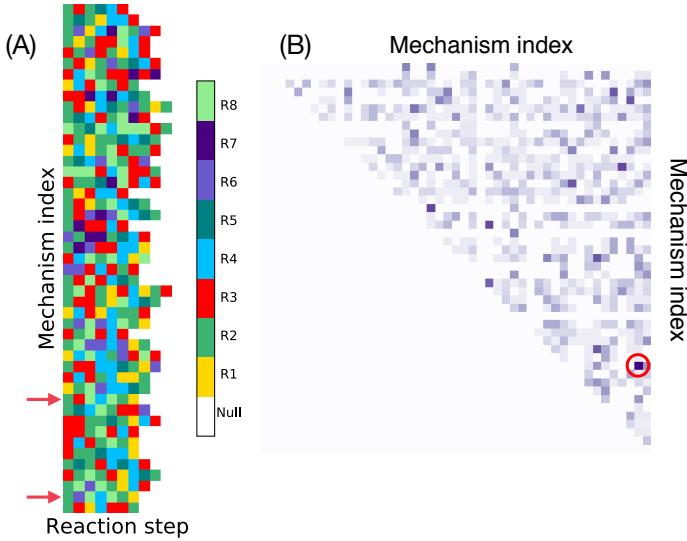
each row (corresponding to a single proposed mechanism) shows a wide variety of mechanisms and sequence-lengths (once the “null?? reactions have been removed). Figure 4(B) further quantifies this comparison of mechanisms by illustrating the calculated similarity between all pairs of reaction mechanisms; this similarity is evaluated by counting the number of common reaction-steps shared by each pair of reaction mechanisms. Here, we find that all 47 of the successful reaction mechanisms are unique; the highest degree of similarity found between any pair of reaction mechanisms is around 85% (as highlighted in Figs. 4(A) and 4(B)), whereas the average similarity is around 11%. These results suggest that most mechanisms only share one or two elementary reaction steps in common; this is also borne out in the energetic considerations outlined below.



**Fig. 3** (A) Progress of graph-error function  $F$  during two different SA optimization calculations, one successful (red) and one unsuccessful (blue). (B)  $Q$ -values and  $|\Delta E^{max}|$  values for all successful GDS simulations of the 50 calculations performed. (C) shows the same data as (B), but zoomed into the region of lower  $Q$ -values.

In order to seek to identify the most plausible reaction mechanism, we adopted the common idea that an idealized reaction profile for a catalytic cycle should be “flat”, in that both energetic changes for each reaction step and kinetic barriers should be minimized.<sup>49</sup> However, as already noted above, the evaluation of energy barriers for many-step reactions can be computationally-demanding; this is exacerbated if one must evaluate energy barriers for the large number of mechanisms generated here. As a result, we instead calculated two readily-accessible descriptors which approximately characterise the “flatness” of the energy landscape associated with a given reaction mechanism based on the structures of the reaction-mechanism intermediates (*i.e.* the end-points of each reaction step) alone. First, we calculate the  $Q$ -value, defined as the sum of energy changes for each elementary reaction step,

$$Q = \sum_{i=1}^{n_r} |\Delta E_i|, \quad (11)$$



**Fig. 4** (A) shows a graphical representation of each of the 47 successful reaction mechanisms located in our GDS approach. Each row represents a different mechanism, and each color-coded box represents one of the nine possible elementary reaction-types shown in Table 1. Where applicable, mechanisms have been concatenated such that contain ?null? reactions (which result in no changes to bonding) appear at the end of the  $n_r = 10$  reaction set. (B) shows a similarity matrix (with more intense color showing greater similarity) calculated by counting the number of reaction-steps each pair of mechanisms have in common. In the case of mechanisms with different lengths (ignoring null reactions), the similarity is assumed to be zero. The red circle indicates the mechanism pair with maximum similarity; these mechanisms are highlighted by red arrows in (A).

where  $\Delta E_i$  is the energy change for reaction  $i$  in the proposed mechanism (e.g. calculated using DFTB for GRP-optimized elementary reaction end-points). In all of the calculations which follow, we use the electronic energies of reaction intermediates to evaluate  $Q$ , to avoid calculation of the Hessian and hence further accelerate this screening process.

For a mechanism, with a “flat” energy landscape, we would expect  $Q$  to be small; we note that absolute values of the energy changes appear in Eq. 11, so that mechanisms with intermediates with either very high energy or very low energy will be disfavoured when searching for mechanisms with low  $Q$ -values. As a second descriptor to categorize each reaction mechanism, we calculated the maximum absolute energy change,  $|\Delta E^{\max}|$ , along the mechanism. We note that both of these descriptors are straightforward to calculate after geometry optimization of molecular configurations for each intermediate in a proposed reaction mechanism; no NEB or TS-finding calculations are required. However, as we show below, while these descriptors are a very useful initial screening tool, the ultimate test of a proposed mechanism must incorporate reaction barrier information too.

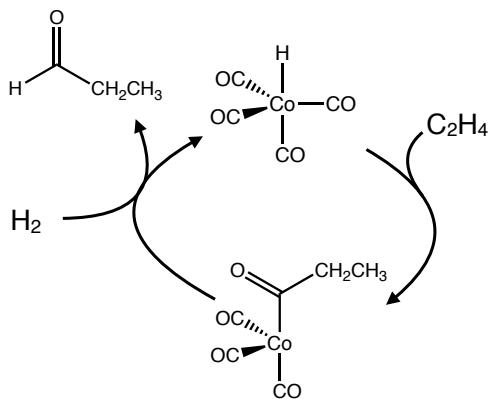
Figure 3(B) shows the  $Q$ -values and  $|\Delta E^{\max}|$  values plotted for all 47 of the successful GDS simulations; as a reminder, each of the points shown corresponds to an entire candidate mechanism which connects the reactants and products, and all energies were

calculated using DFTB following geometry optimization of reaction intermediates initially-generated by GRP optimization. The  $|\Delta E^{\max}|$  values range up to about  $4000 \text{ kJ mol}^{-1}$ , showing that some GDS-generated mechanisms involve extremely high-energy intermediates; such mechanisms can be readily discounted as viable candidates for the “correct” mechanism. In the case of the high-energy intermediates, it is essentially impossible for a thermal reaction mechanism to generate such paths with any significant probability, bearing in mind that the kinetic barrier to reaction, if it exists, must be at least  $|\Delta E^{\max}|$  for the reaction leading to the highest-energy intermediate. In the case of very low-energy intermediates, the reverse argument applies; catalytic cycles with very low-energy intermediates should be disfavoured in order to avoid kinetic trapping and reduced catalytic turnover.<sup>49</sup>

Figure 3(C) shows the same plot as Fig. 3(B), but focussing on the region with small  $Q$ -value and small  $|\Delta E^{\max}|$ . If  $Q$  and  $|\Delta E^{\max}|$  are good descriptors to categorize different mechanisms, we expect that the “correct” Heck-Breslow mechanism would fall in the lower left-hand corner of Fig. 3(B) and, preferably, should indeed be the point with the lowest values of both descriptors. However, as highlighted in Fig. 3(C), we actually find two mechanisms which both have much lower  $Q$  and  $|\Delta E^{\max}|$  values than the other generated mechanisms. One of these mechanisms (referred to hereafter as mechanism I) has  $Q = 527 \text{ kJ mol}^{-1}$  and  $|\Delta E^{\max}| = 118 \text{ kJ mol}^{-1}$ , whereas the other mechanism (referred to hereafter as mechanism II) has  $Q = 409 \text{ kJ mol}^{-1}$  and  $|\Delta E^{\max}| = 119 \text{ kJ mol}^{-1}$ .

Based on the  $Q$  and  $|\Delta E^{\max}|$  values alone, one might therefore hope that mechanism II is the well-known Heck-Breslow mechanism; this is not the case, although it is found that the Heck-Breslow mechanism does indeed correspond to mechanism I, with slightly larger  $Q$ -value than mechanism II. Closer visual inspection of the intermediate structures generated along mechanism II reveals a slightly different catalytic cycle than that shown in Fig. 2. As shown in Fig. 5, mechanism II actually proceeds by direct insertion of  $\text{C}_2\text{H}_4$  into the cobalt-carbonyl bond  $\text{HCo}(\text{CO})_4$ , in a concerted reaction which also leads to hydrogen transfer onto one end of the ethene. This concerted step is quite different to the Heck-Breslow scheme, which suggests that CO must initially dissociate from  $\text{HCo}(\text{CO})_4$  to initiate reaction. However, the energetics of the reaction intermediates for mechanism II is clearly slightly more favourable than that of mechanism I; this is a consequence of the concerted nature of the reaction.

So, based on intermediate energetics and the concept of “flat energy profile” alone, our results suggest not the usual Heck-Breslow scheme, but an alternative concerted reaction scheme (Fig. 5) as the “most likely” mechanism. To truly distinguish between these two mechanisms obviously requires further calculations relating to the kinetic barriers for each intermediate reaction in the two proposed catalytic cycles. To this end, we performed NEB calculations for all  $n_r = 10$  reaction paths in each of mechanism I and II, giving the full energy profile for the entire mechanism in each case. Here, to generate a fine resolution of the energy profile of each reaction step in each mechanism, we used the AutoNEB scheme<sup>50</sup> with a total of 20 images describing each reaction step.



**Fig. 5** A summarized version of an alternative hydroformylation reaction mechanism located by our GDS simulations. This mechanism, involving a concerted insertion of the alkene into  $\text{HCo}(\text{CO})_4$ , has similar reaction energetics to the expected Heck-Breslow mechanism (Fig. 2).

The results of these NEB optimizations are shown in Figs. 6 and 7. Somewhat satisfyingly, we find that the highest energy barrier to reaction is around  $213 \text{ kJ mol}^{-1}$  in mechanism I (R7) and  $264 \text{ kJ mol}^{-1}$  in mechanism II (R5). In other words, although mechanism II seems slightly more favourable based on energetics alone, there is a clear kinetic preference (by  $\approx 50 \text{ kJ mol}^{-1}$ ) for mechanism I; the Heck-Breslow mechanism comes out on top as the kinetically favoured catalytic cycle. We conclude that a combination of (i) pre-screening based on reaction intermediates, and (ii) direct energetic barrier estimation enables unambiguous identification of the preferred catalytic mechanism. In the case considered here, we find that the mechanism with both the most favourable energetics of reaction intermediates and the lowest maximum energy barrier corresponds to the Heck-Breslow scheme.

### 3 A closer look at GDS reaction paths

Although our GDS scheme, when combined with screening based on reaction energetics and reaction barriers, enables identification of the hydroformylation mechanism, these challenging simulations reveal a great deal of additional information about graph-driven reaction discovery. Here, we highlight two key features which emerge from our combination of GDS with NEB simulations; importantly, these features of our simulation approach do not impact on the ability of our overall scheme to discern the most likely overall mechanism, but do suggest some ways forward in terms of improving our simulation scheme.

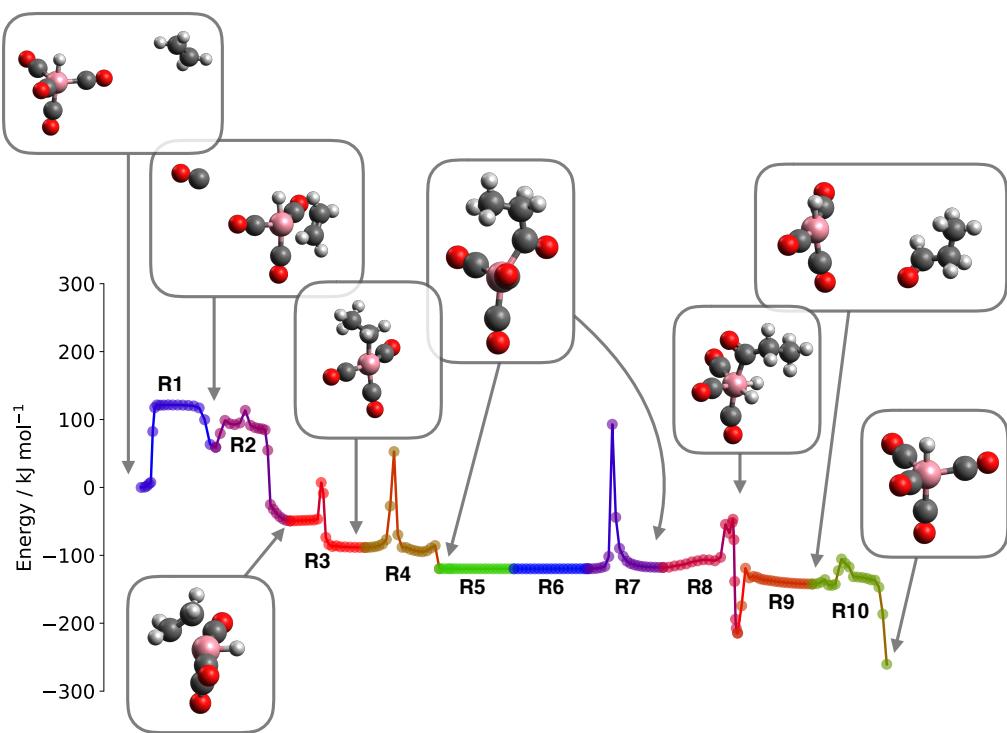
**Irrelevant isomerization reactions.** As noted above, the largest energy barrier to reaction in mechanism I is associated with reaction R7. However, closer inspection of the reaction intermediates in R7 in mechanism I reveals that this step is actually an isomerization step which rearranges two adjacent carbonyl lig-

ands on an alkyl-cobalt intermediate but does not ultimately lead to a change in chemical structure. This is highlighted in Fig. 8, which shows more detailed structures along R7; as shown by the labelling of carbonyl groups, R7 involves a ligand exchange reaction which swaps the carbonyl group within the cobalt-bound  $\text{COCH}_2\text{CH}_3$  group for one of the other CO ligands. The net effect of this reaction is no overall change in chemical structure or energy; in other words, if this high-barrier step in mechanism I is simply removed, the mechanism as a whole does not change (other than skipping an unnecessary isomerization step) and the remaining maximum energetic barrier is then reduced to  $141 \text{ kJ mol}^{-1}$ , around  $120 \text{ kJ mol}^{-1}$  lower than the lowest barrier in mechanism II.

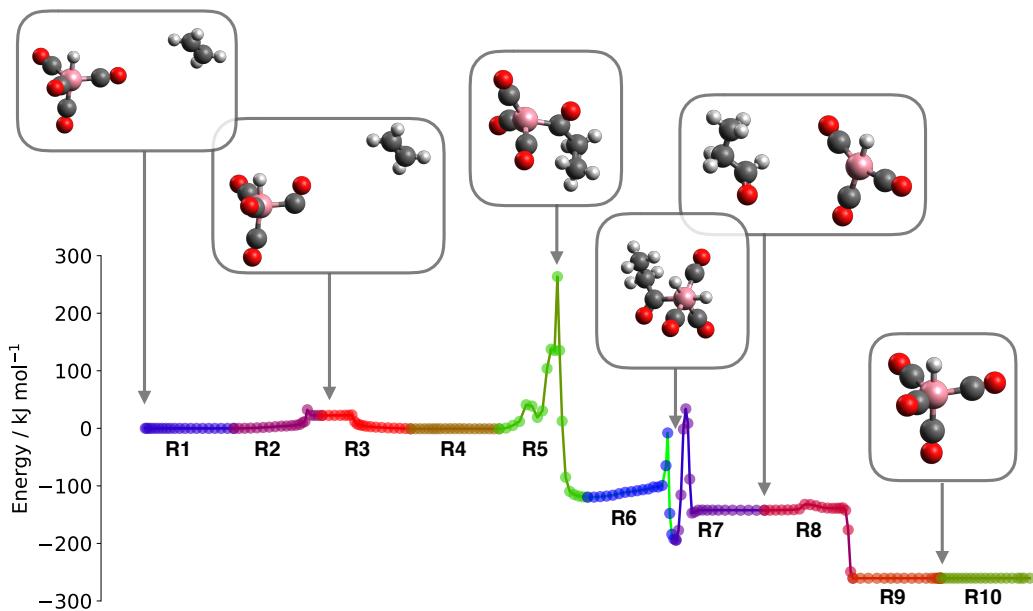
**Conformational changes.** As well as the unnecessary isomerization reactions which have been highlighted above, we also find that several reaction steps can exhibit barriers as a result of conformational changes, where no change in bonding is observed but changes in molecular geometry induce either an energetic barrier or reaction energy change. In such cases, as with the isomerization reactions noted above, these conformational changes can serve to complicate the assessment of a given reaction mechanism.

Two clear examples of conformational change occur in mechanisms I (Fig. 6). In the first example, in R2, it is found that the initial  $\text{HCo}(\text{CO})_3(\text{C}_2\text{H}_4)$  species differs from the product of the reaction by a conformational change in which the  $\text{C}_2\text{H}_4$  adduct differs in its rotational orientation relative to the catalyst. This conformational change decreases the energy by about  $70 \text{ kJ mol}^{-1}$ ; more importantly, as in the case of the isomerization reactions considered above, it does not lead to significant progress along the reaction coordinate towards products. That said, there is a clear energetic preference for the conformation which results from R2 for further reaction, indicating that conformational searching could be implemented to search for not only the set of chemical reactions which lead from reactants to products, but also the most energetically-favourable molecular conformations of the associated intermediate structures. As a second example, in R4 in Fig. 6, it is found that the large barrier of about  $141 \text{ kJ mol}^{-1}$  actually arises from a conformational change which corresponds to an umbrella inversion of the  $\text{CH}_2$  group in the cobalt-bound alkane species; the shallow minimum and smaller barrier at the end of R4 then corresponds to binding of CO and insertion into the cobalt-alkyl bond, reactions which do actually progress the system towards the target products.

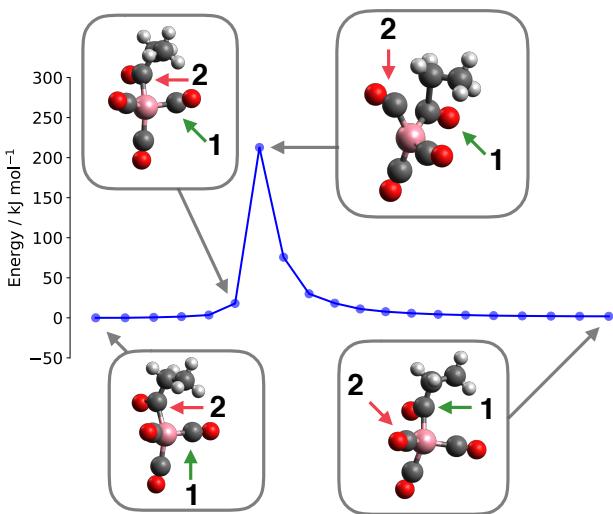
As a final point, it is interesting to note what mechanism I would look like if these non-progressing isomerizations and conformational changes were screened out. This would remove the large barriers in R7 and R4, leaving the next largest energetic barriers as  $121 \text{ kJ mol}^{-1}$  (alkene replacing CO in R1) and  $95 \text{ kJ mol}^{-1}$  (dissociation of aldehyde product in R9). Furthermore, we note that similar consideration of isomerizations and conformations do not significantly change the larger barriers in mechanisms II (Fig. 7); in other words, it is clear that accounting for these features leads to an even stronger steer towards mechanism I as the ‘most likely’ mechanism, as expected based on previous



**Fig. 6** One of two mechanisms with the lowest  $Q$  and  $|\Delta E^{\max}|$  values. NEB calculations were performed for all  $n_r = 10$  reactions in this mechanism; key intermediate structures are shown, with non-involved reactant molecules removed for clarity. This mechanism maps onto the well-known Heck-Breslow reaction mechanism of Fig. 2.



**Fig. 7** A further mechanism postulated on the basis of low  $Q$  and  $|\Delta E^{\max}|$  values. In contrast to Fig. 6, the key step here (R5) is the concerted addition of  $C_2H_4$ , simultaneously forming a  $Co-(CO)-CH_2$  group and also adding hydrogen to the other alkene  $CH_2$  to form  $CH_3$ . The remainder of the mechanism after this step follows the usual Heck-Breslow scheme of Figs. 2 and 6.



**Fig. 8** An example of an isomerization reaction which does not progress the reaction. Here, the CO ligand labelled “1” is swapped for the CO group labelled “2”; as a result, the initial and final structures in this isomerization have the same energy and no net reaction has been performed.

simulations and experiments.

**GDS improvements.** Both the irrelevant isomerization reactions and conformational changes observed in our GDS-generated reaction paths arise as a consequence of two details of our GDS implementation. First, we have not explicitly assumed that we know the correct number of steps in the target mechanism; instead, we have chosen some maximum value of  $n_r$  (here,  $n_r = 10$ ) which we expect to be “large enough”. In other words, there is a greater degree of flexibility in our generated mechanisms than is absolutely necessary and, as a result, it is possible to find reaction mechanisms which successfully connect reactant and product CMs but also make some unnecessary excursions (e.g. isomerizations). Second, we note that adapting our GDS scheme to account for permutational invariance of atomic labelling might have an impact on minimizing these unnecessary isomerization reactions; for example, if permutational invariance under changes of atomic labelling was accounted for, this could be used to identify the fact that both the reactants and products of R7 (Figs. 6 and 8) are chemically identical. As such, this permutationally-invariant monitoring of chemical structures could be used to remove any isomerization reactions which do not progress the reaction mechanism towards the target products. Accounting for permutational invariance, as well as investigating strategies for trimming reactions from maximum-length mechanisms, are currently under active investigation and will be reported shortly.

At this point, it is also important to highlight the stochastic nature of our current GDS scheme; because our approach is based on stochastic generation of trial mechanisms, there is no guarantee that any of the mechanisms located by a given number of such searches will actually correspond to the “correct” mechanism. Instead, in common with many stochastic approaches, one must simply run enough calculations to be sufficiently confident about the results, but it is not clear *a priori* exactly what this number should be. However, we note that the next stage in development

of this algorithm will be to couple our approach to an outer global optimization algorithm, enabling us to not only search for any reaction mechanism connecting reactants and products, but to seek out that reaction mechanism which is most likely (as judged, for example, using  $Q$  and  $|\Delta E|$  values).

Although our GDS scheme is compatible with any energy calculation scheme (given that the energy evaluations are used as a post-processing analysis tool), some further adaptations will be necessary to deal with catalytic systems in which the electronic state (e.g. spin or redox state) change during a reaction. In such cases, we anticipate that one could calculate the energies of all relevant electronic states for each reaction intermediate; connecting these electronic manifolds together when assessing the suitability of a given proposed reaction mechanism would then enable greater insight into the role of electronic state changes, and this will be explored in the near future.

## 4 Conclusions

To summarize, we have shown that a combination of novel reaction discovery algorithm, combined with approximate evaluation of reaction energetics, can act as a strong filter in seeking to determine the most likely reaction mechanism for a given catalytic cycle. However, we have also found that, in order to truly distinguish between two conflicting mechanisms with similar energetic descriptors, evaluation of approximate energy barriers (here performed using NEB calculations) is invaluable and enables unique identification of the thermodynamically and kinetically favourable mechanism. In the case illustrated here, we have shown that this strategy enables identification of a mechanism of ethene hydroformylation which maps directly onto the well-known Heck-Breslow scheme.

Although this Article has successfully demonstrated our graph-driven scheme for reaction discovery in the context of homogeneous catalysis, there are obviously a large number of avenues for improvement and expansion. For example, screening such a large number of candidate reaction mechanisms is obviously a computational burden, and can only be achieved if efficient yet accurate approximations to the PES of general molecular systems are available. In the case considered here, DFTB has proven sufficient in enabling us to pick out the most likely reaction mechanism, but this may not be the case in all systems, particularly those for which the performance of DFTB is not well characterized. As we have emphasised above, our GDS scheme can be used in combination with any energy calculation method, so there is scope to explore how one might most efficiently achieve this to improve overall predictive accuracy. As a further example of remaining challenges, we note that our graph-driven scheme (in common with other graph-based schemes) does not currently contain information about stereochemistry; in order to evaluate stereoselectivity, one could envisage performing separate NEB calculations for each different possible stereochemistry, although this would become very time-consuming when multiple stereocentres are present. As noted above, an important next step in development of our approach will also be to couple our GDS scheme to an outer global optimization procedure which guides the search over mechanisms to not only find any mechanism, but the most likely

mechanism, as judged, for example, by energetic considerations such as calculation of  $Q$  and  $|\Delta E|$ . In addition, the SA scheme which underlies our GDS calculations also has room for improvement; we have not yet sought to increase its efficiency or reliability, and there is scope to do both by, for example, changing how SA moves are proposed or optimizing the annealing schedule. Finally, we note that searching over reactive conformers at each reaction intermediate is not performed here; again, this approach is quite common in the field of reaction discovery, but forgoes the possibility of conformer-dependence in the reaction kinetics. So, there are challenges (and possible solutions) ahead, but this Article represents an important step towards fully-automated reaction discovery and mechanism identification for homogeneous catalytic systems.

## Conflicts of interest

There are no conflicts to declare.

## Acknowledgements

The authors gratefully acknowledge the award of funding by the Engineering and Physical Sciences Research Council (EPSRC; EP/R020477/1). We also gratefully acknowledge high-performance computing facilities provided by the Scientific Computing Research Technology Platform at the University of Warwick.

Data from Figures 3, 4, 6, 7 and 8 can be found at [wrap.warwick.ac.uk/122793](http://wrap.warwick.ac.uk/122793).

## Notes and references

- 1 E. Martínez-Núñez, *J. Comput. Chem.*, 2015, **36**, 222 – 234.
- 2 K. Ohno and S. Maeda, *Phys. Scripta*, 2008, **78**, 058122.
- 3 S. Habershon, *J. Chem. Theory Comput.*, 2016, **12**, 1786–1798.
- 4 C. F. Goldsmith and R. H. West, *J. Phys. Chem. C*, 2017, **121**, 9970 – 9981.
- 5 C. A. Class, M. Liu, A. G. Vandeputte and W. H. Green, *Phys. Chem. Chem. Phys.*, 2016, **18**, 21651–21658.
- 6 I. Ismail, H. B. V. A. Stuttaford-Fowler, C. Ochan Ashok, C. Robertson and S. Habershon, *J. Phys. Chem. A*, 2019, **123**, 3407–3417.
- 7 L.-P. Wang, A. Titov, R. McGibbon, F. Liu, V. S. Pande and T. J. Martínez, *Nature Chem.*, 2014, **6**, 1044–8.
- 8 G. N. Simm, A. C. Vaucher and M. Reiher, *J. Phys. Chem. A*, 2019, **123**, 385–399.
- 9 S. Maeda and K. Ohno, *J. Phys. Chem. A*, 2005, **109**, 5742–5753.
- 10 A. L. Dewyer, A. J. Argüelles and P. M. Zimmerman, *Wiley Interdisciplinary Reviews: Computational Molecular Science*, 2018, **8**, e1354.
- 11 S. Habershon, *J. Chem. Phys.*, 2015, **143**, 094106.
- 12 S. Maeda and K. Morokuma, *J. Chem. Theory Comput.*, 2012, **8**, 380–385.
- 13 D. T. Gillespie, *J. Phys. Chem.*, 1977, **81**, 2340–2361.
- 14 D. T. Gillespie, A. Hellander and L. R. Petzold, *J. Chem. Phys.*, 2013, **138**, 170901.
- 15 D. T. Gillespie, *Annu. Rev. Phys. Chem.*, 2007, **58**, 35–55.
- 16 M. Besora and F. Maseras, *Wiley Interdisciplinary Reviews: Computational Molecular Science*, 2018, **8**, e1372.
- 17 M. Keçeli, S. N. Elliott, Y.-P. Li, M. S. Johnson, C. Cavallotti, Y. Georgievskii, W. H. Green, M. Pelucchi, J. M. Wozniak, A. W. Jasper and S. J. Klippenstein, *Proc. Combust. Inst.*, 2019, **37**, 363 – 371.
- 18 Y. Kim, S. Choi and W. Y. Kim, *J. Chem. Theory Comput.*, 2014, **10**, 2419–2426.
- 19 Y. Kim, J. W. Kim, Z. Kim and W. Y. Kim, *Chem. Sci.*, 2018, **9**, 825–835.
- 20 P. M. Zimmerman, *J. Comput. Chem.*, 2013, **34**, 1385–1392.
- 21 A. Rodríguez, R. Rodríguez-Fernández, S. A. Vázquez, G. L. Barnes, J. J. P. Stewart and E. Martínez-Núñez, *J. Comput. Chem.*, 2018, **39**, 1922–1930.
- 22 P. M. Zimmerman, *J. Chem. Theory Comput.*, 2013, **9**, 3043–3050.
- 23 C. W. Coley, W. Jin, L. Rogers, T. F. Jamison, T. S. Jaakkola, W. H. Green, R. Barzilay and K. F. Jensen, *Chem. Sci.*, 2019, **10**, 370–377.
- 24 J. S. Schreck, C. W. Coley and K. J. M. Bishop, *ACS Central Science*, 2019, **5**, 970–981.
- 25 P. Schwaller, T. Laino, T. Gaudin, P. Bolgar, C. Bekas and A. A. Lee, *ChemRxiv*, 2019.
- 26 K. J. Laidler, *Chemical Kinetics*, Harper Collins, New York, 3rd edn, 1987.
- 27 M. E. Tuckerman, *Statistical Mechanics: Theory and molecular simulation*, Oxford University Press, 2012.
- 28 D. G. Truhlar, B. C. Garrett and S. J. Klippenstein, *J. Phys. Chem.*, 1996, **100**, 12771–12800.
- 29 K. J. Laidler and M. C. King, *J. Phys. Chem.*, 1983, **87**, 2657–2664.
- 30 G. Henkelman, B. P. Uberuaga and H. Jónsson, *J. Chem. Phys.*, 2000, **113**, 9901.
- 31 L. Xie, H. Liu and W. Yang, *J. Chem. Phys.*, 2004, **120**, 8039–8052.
- 32 G. Henkelman and H. Jónsson, *J. Chem. Phys.*, 2000, **113**, 9978.
- 33 N. González-García, J. Pu, A. González-Lafont, J. M. Lluch and D. G. Truhlar, *J. Chem. Theory Comput.*, 2006, **2**, 895–904.
- 34 N. Govind, M. Petersen, G. Fitzgerald, D. King-Smith and J. Andzelm, *Comp. Mat. Sci.*, 2003, **28**, 250 – 258.
- 35 B. Peters, A. Heyden, A. T. Bell and A. Chakraborty, *The Journal of Chemical Physics*, 2004, **120**, 7877–7886.
- 36 S. Kopec, E. Martínez-Núñez, J. Soto and D. Peláez, *International Journal of Quantum Chemistry*, 2019, **119**, e26008.
- 37 C. W. Gao, J. W. Allen, W. H. Green and R. H. West, *Computer Physics Communications*, 2016, **203**, 212 – 225.
- 38 B. Aradi, B. Hourahine and T. Frauenheim, *J. Phys. Chem. A*, 2007, **111**, 5678–5684.
- 39 G. Zheng, H. A. Witek, P. Bobadova-Parvanova, S. Irle, D. G. Musaev, R. Prabhakar, K. Morokuma, M. Lundberg, M. Elstner,

- C. Köhler and T. Frauenheim, *J. Chem. Theory Comput.*, 2007, **3**, 1349–1367.
- 40 T. Kegl, *RSC Adv.*, 2015, **5**, 4304–4327.
- 41 L. E. Rush, P. G. Pringle and J. N. Harvey, *Angew. Chemie Int. Ed.*, 2014, **53**, 8672–8676.
- 42 C.-F. Huo, Y.-W. Li, M. Beller and H. Jiao, *Organometallics*, 2003, **22**, 4665–4677.
- 43 O. M. K. Raghuraj V. Gholap and J. R. Bourne, *Ind. Eng. Chem. Res.*, 1992, **31**, 1597–1601.
- 44 M. F. Mirbach, *J. Org. Chem.*, 1984, **265**, 205–213.
- 45 M. Orchin and W. Rupilius, *Cat. Rev. - Sci. Eng.*, 2006, **6**, 85–131.
- 46 G. Natta, R. Ercoli, S. Castellano and F. H. Barbieri, *J. Am. Chem. Soc.*, 1954, **76**, 4049–4050.
- 47 R. F. Heck and D. S. Breslow, *J. Am. Chem. Soc.*, 1961, **83**, 4023–4027.
- 48 J. A. Varela, S. A. Vázquez and E. Martínez-Núñez, *Chem. Sci.*, 2017, **8**, 3843–3851.
- 49 S. Raugei, D. L. DuBois, R. Rousseau, S. Chen, M.-H. Ho, R. M. Bullock and M. Dupuis, *Acc. Chem. Res.*, 2015, **48**, 248–255.
- 50 E. L. Kolsbjerg, M. N. Groves and B. Hammer, *J. Chem. Phys.*, 2016, **145**, 094107.