Automated prediction of catalytic mechanism and rate law using graph-based reaction-path sampling

Scott Habershon*

Department of Chemistry and Centre for Scientific Computing, University of Warwick, Coventry, CV4 7AL, United Kingdom

E-mail: S.Habershon@warwick.ac.uk

Abstract

In a recent article [*J. Chem. Phys.*, **143**, 094106 (2015)], we have introduced a novel graph-based sampling scheme which can be used to generate chemical reaction paths in many-atom systems in an efficient and highly-automated manner. The main goal of this work is to demonstrate how this approach, when combined with direct kinetic modelling, can be used to determine the mechanism and phenomenological rate law of a complex catalytic cycle, namely cobalt-catalyzed hydroformylation of ethene. Our graph-based sampling scheme generates 31 unique chemical products and 32 unique chemical reaction pathways; these sampled structures and reaction paths enable automated construction of a kinetic network model of the catalytic system when combined with density functional theory (DFT) calculations of free energies and resultant transition-state theory rate constants. Direct simulations of this kinetic network across a range of initial reactant concentrations enables determination of both the reaction mechanism and the associated rate law in an automated fashion, without the need for either pre-supposing a mechanism or making steady-state approximations in kinetic analysis. Most importantly, we find that the reaction mechanism which emerges from these simulations is exactly that originally proposed by Heck and Breslow; furthermore, the simulated rate law is also consistent with previous experimental and computational studies, exhibiting a complex dependence on carbon monoxide pressure. While the inherent errors of using DFT simulations to model chemical reactivity limit the quantitative accuracy of our calculated rates, this work confirms that our automated simulation strategy enables direct analysis of catalytic mechanisms from first principles.

*To whom correspondence should be addressed

Introduction

Homogeneous catalysis is one of the foundations of the worldwide chemical industry;[1–3] as a result, the development of molecular catalysts with improved turnover frequency, higher selectivity and better resistance to degradation is a central goal of fundamental chemistry research. Computational chemistry provides, at least in principle, a direct route to designing and optimizing novel molecular catalysts with selected functionality;[4] for example, new proline-based derivatives for catalysis of aldol reactions have been reported following computational design,[5,6] and recent work on designer electrocatalysts has also been reported using a scheme based on minimizing free energy changes along catalytic pathways.[7]

However, it is important to note that the complete *de novo* design of molecular catalysts using computational chemistry remains a largely unmet challenge; most catalysts designed with the aid of computer simulations have used an existing catalyst as a starting-point, in a similar manner to the "lead development" approach often taken in pharmaceutical development.[4] Furthermore, simulations aimed at designing new catalytic processes can be limited in scope, for example focussing solely on the thermodynamic aspects of catalytic paths[7,8] or, particularly when investigating the influence of the catalyst properties on reaction kinetics, focussing on a single reaction step corresponding to the rate-determining reaction in the catalytic cycle. The rate-determining step is usually identified by applying chemical "common sense", for example seeking the step in the catalytic cycle which involves breaking the strongest chemical bonds, or by comparison to similar reactions with known intermediate steps. This emphasis on a single intermediate step in understanding catalytic propensity can clearly be successful, but relies on the separation of time-scales arising when a single reaction step with a high free-energy barrier exists.

In many catalytic reactions, the experimentally-observed rate law, the key observable of interest in optimising catalytic activity, often exhibits a complex dependence on reactant concentrations which mirrors a similarly complex kinetic network at play. For example, the cobalt-catalyzed hydroformylation reaction (oxo process[9–11]) which will be the focus of

this Article has been found to exhibit a rate law with non-trivial dependence on reactant concentrations; experimental studies of cobalt-catalyzed hydroformylation of propene have produced rate law expressions with the following forms,[10–12]

$$\frac{d\,[\text{Aldehyde}]}{dt} = k\frac{[\text{CO}]\,[\text{Alkene}]\,[\text{H}_2]^{0.6}\,[\text{Catalyst}]^{0.8}}{(1+k_1\,[\text{CO}])^2}, \qquad (1)$$

$$\frac{d\,[\text{Aldehyde}]}{dt} = k\frac{[\text{Alkene}]\,[\text{H}_2]\,[\text{Catalyst}]}{([\text{CO}])},$$

while a recent theoretical investigation resulted in a rate law of the form,[13]

$$\frac{d\,[\text{Aldehyde}]}{dt} = k\frac{[\text{Alkene}]\,[\text{H}_2]^{0.5}\,[\text{Catalyst}]^{0.5}}{[\text{CO}]}. \qquad (2)$$

These non-trivial rate laws generally arise because of the existence of several competing reactions with comparable activation barriers, as well as a complex interplay of equilibria amongst the various reactive species; in such cases, computational studies focussing on a single reaction steps, or thermodynamic changes only, will be implicitly limited.

Unfortunately, the challenges posed to computational chemistry in modelling complex, multi-step catalytic cycles are well-known; as a result, design of new tailored homogeneous catalysts from first-principles consideration of the entire catalytic cycle is hindered.[14] In particular, the potential energy surface (PES) of the system should preferably be sufficient to achieve "chemical accuracy" (errors of order 1 kcal mol^{-1}), or at least such that the relative errors in calculated energies and reaction barriers are consistent across the relevant set of reaction paths; more often than not, this level of accuracy is extremely difficult to achieve, even when high-level electronic structure methods are applicable.[13] Furthermore, one must have an approach for sampling multiple *conformers* at the reaction end-points, as well as different *isomers*, in order to generate all of the chemically-relevant reactions which can potentially occur for a given system. Finally, an approach to determining transition

state (TS) structures, reaction energy barriers and reaction rates is also required to model the chemical kinetics inherent in the network of possible reaction paths.[15]

The challenges of predictive catalysis are further highlighted by the fact that very few computational approaches have, to date, been shown capable of determining complete catalytic mechanisms in an *automated* and *unbiased* manner. For example, the Scaled Hypersphere Searching (SHS[16–18]) methodology has proven extremely powerful in obtaining PES minima and transition-states for chemically-reactive systems containing up to around 12 atoms,[16] as well as locating isomers of molecular clusters,[19] but has yet to be applied to *ab initio* catalysis. The basin-hopping Monte Carlo (BHMC[20]) strategy has been applied to study catalysis, notably the cobalt-catalyzed hydroformylation reaction studied here; however, it is unclear to what extent these BHMC simulations actually sampled a broad range of chemical reactions for this system, or whether the choice of molecular fragments was guided in order to aid discovery of the main catalytic reaction-steps. Similarly, the artificial force induced reaction (AFIR[21]) approach has also been applied to cobalt-catalysed hydroformylation,[22] but again with significant human input being used to provide guidance on reaction-path searching. Perhaps closest in spirit to our recent work,[23] and this article, is the report of a chemical-connectivity-based approach[24] combined with growing-string TS searching,[25,26] for automated reaction discovery; however our approach is not specifically reliant upon a TS search method such as growing-string, instead defining an implicit dynamic reaction path associated with a Hamiltonian, thereby enabling straightforward conformational *and* chemical sampling, as described below. Finally, it is worth noting the recent report of a computational "nanoreactor" aimed at sampling reactive chemistry by periodically driving bond rearrangements *via* application of high pressures;[27] this methodology has not yet been applied to study catalysis, but has already proven promising in discovering new reaction paths leading to glycine formation in a computational analogue of the Urey-Miller experiment.[28]

In this paper, following on from an initial methodological report,[23] we demonstrate how

a new atomistic reaction-path sampling method can be used to build a complete picture of reaction kinetics from the bottom-up. Our approach combines (i) a novel classical Hamiltonian describing a dynamic reaction path connecting two well-defined chemical isomers, (ii) molecular dynamics (MD) sampling of trajectories in reaction-path-space *and* of the reaction end-points, and (iii) periodic updates of bond connectivity at the reaction-path end-points. Overall, our strategy enables simultaneous sampling of multiple reaction paths connecting multiple chemical isomers and molecular conformations in an automated fashion; the purpose of this paper is to demonstrate that this strategy enables us to systematically proceed from atomistic simulations to complex kinetic networks, ultimately leading to interpretation and rationalization of reaction mechanism and experimental rate law. The significant advantage of our simulation strategy is that the experimental rate law and mechanism directly emerge without any prior mechanistic assumptions being made, and without reducing the mechanism to a few simple steps; this unbiased strategy clearly contrasts against the majority of previous approaches to catalytic simulations, and potentially opens a new door to direct optimization of catalytic reaction kinetics.

The remainder of this manuscript is organized as follows. We first outline our graph-based approach to reaction-path sampling, before describing how the ensemble of reaction paths sampled in this methodology can be subsequently analyzed using geometry optimization and TS searching. We then describe how the combination of automated reaction-path sampling and structural optimization can be used to build a kinetic network model, allowing one to make connection to experimental kinetics. This overall strategy is then applied to study the cobalt-catalysed hydroformylation of ethene;[9–11,13,22,29–32] we find that our bottom-up strategy enables direct identification of the important reactive steps in the catalytic cycle, and also yields a simulated rate law which is comparable to previous experimental and theoretical results *without* pre-guessing the mechanism or adopting steady-state assumptions. Finally, we highlight the fact that, although the predicted rates of chemical reaction of our kinetic model are somewhat low (yet easily corrected), the emergent kinetic mechanism is

robust to these errors. Overall, our simulations allow us to move from atomistic sampling to experimental kinetics in a systematic manner.

Method

The computational scheme adopted in this Article is hierarchical in nature; starting from atomistic simulations for sampling reaction paths, moving to refinement of reaction paths and identification of TSs, and finally using the sampled reaction-path data as input to a kinetic model describing the entire network of reactions, our approach represents a "bottom-up" route to catalytic kinetics. The three stages of our simulation approach are outlined below.

Graph-based reaction-path sampling

Our computational approach (Fig. 1) to automated reaction-path sampling has been outlined in a recent article;[23] here, we briefly review the most important features of this methodology. In contrast to previous methods aimed at automated reaction-path sampling, which most commonly rely on random generation of initial reactant and product configurations, our approach transforms the challenge of chemical reaction sampling into a simple classical MD-like sampling strategy, albeit operating in the space of accessible reaction paths. As well as subsequently being able to exploit the many tools developed to improve the efficiency of MD sampling, such as temperature-accelerated approaches,[33,34] our strategy also ensures that a physically-sensible reaction path is always maintained throughout the simulation, avoiding unrealistic pathways which are unlikely to contribute to the overall kinetics of the system of interest.

First, we define a classical Hamiltonian which describes a reaction-path connecting two

end-points, $\mathbf{r}_0$ and $\mathbf{r}_P$,

$$H(\mathbf{r}_0, \mathbf{r}_P, \mathbf{p}_0, \mathbf{p}_P, \mathbf{a}, \mathbf{G}^0, \mathbf{G}^P) = \sum_{i=1}^{N_a} \frac{|\mathbf{p}_0^{(i)}|^2}{2m_i} + \sum_{j=1}^{N_a} \frac{|\mathbf{p}_P^{(j)}|^2}{2m_j} + \sum_{k=1}^{P} \frac{|\mathbf{b}^{(k)}|^2}{2\mu} + V_s(\mathbf{r}_0, \mathbf{r}_P, \mathbf{a}, \mathbf{G}^0, \mathbf{G}^P). \tag{3}$$

The Hamiltonian of Eq. 3 describes a reaction path, defined as a set of images transitioning from reactants $\mathbf{r}_0$ to products $\mathbf{r}_P$. In what follows, we assume that the reactive system is described by a set of Cartesian coordinates, although there is no reason that our approach could not employ alternative coordinate systems; as a result, each of the reactant end-points, $\mathbf{r}_0$ and $\mathbf{r}_P$, is a point in the $(3N_a - 6)$-dimensional configurational space for the N_a-atom system. The reaction string has M intermediate configurations; again, each of these intermediates is a configuration in the full $(3N_a - 6)$-dimensional space. The reaction string is parameterized by a set of P Fourier coefficients for each of the $(3N_a - 6)$ degrees-of-freedom, such that the positions of the intermediate images are related to the set of Fourier coefficients, $\mathbf{a}$, according to[35,36]

$$\mathbf{r}_i = \mathbf{r}_0 + \lambda_i(\mathbf{r}_P - \mathbf{r}_0) + \sum_{k=1}^{P} \mathbf{a}_k \sin(k\pi\lambda_i). \tag{4}$$

Here, $\lambda_i \in [0, 1]$ is a linear variable that describes the position of image i along the reaction pathway from reactants ($\lambda = 0$) to products ($\lambda = 1$), such that

$$\lambda_i = \frac{i}{(M+1)}. \tag{5}$$

The reaction path end-points, $\mathbf{r}_0$ and $\mathbf{r}_P$, are associated with conjugate momenta, $\mathbf{p}_0$ and $\mathbf{p}_P$, respectively; similarly, the Fourier coefficients describing the reaction-path are also associated with a set of conjugate momenta, $\mathbf{b}$. The first three terms of the classical Hamiltonian of Eq. 3 are thus the usual kinetic energy contributions associated with the momenta of the reaction start- and end-points, as well as the artificial conjugate momenta associated with the Fourier coefficients describing the reaction path. The final term in Eq. 3, $V_s(\mathbf{r}_0, \mathbf{r}_P, \mathbf{a}, \mathbf{G}^0, \mathbf{G}^P)$, is

the potential energy function for the reaction-path system, and is given by

$$V_s(\mathbf{r}_0, \mathbf{r}_P, \mathbf{a}, \mathbf{G}^0, \mathbf{G}^P) = V(\mathbf{r}_0) + V(\mathbf{r}_P) + \frac{1}{M}\sum_{k=1}^{M}\left[V(\mathbf{r}_k) + \gamma_1|\mathbf{r}_k - \mathbf{r}_{k-1}|^2\right] + W(\mathbf{r}_0, \mathbf{G}^0) + W(\mathbf{r}_P, \mathbf{G}^P). \tag{6}$$

Here, $V(\mathbf{r})$ is the PES describing the interatomic interactions in the system (described below) and γ_1 is a user-defined harmonic constant; thus, the first three terms in Eq. 6 describe the potential energy of the set of images representing the reaction string, with the intermediate coordinates given by Eq. 4. Note that third term in Eq. 6 is comparable to the PES found in the nudged elastic band (NEB[37,38]) method for reaction-path refinement; however, in the present case, we do not employ minimization using projected forces, as in the NEB approach. The harmonic interaction term acting between adjacent images in Eq. 6 helps ensure that a stable and continuous reaction-path is sampled, avoiding spurious "leaps" or "kinks" along the reaction-path which might cause instabilities in time-evolution, as noted in our original report.[23] In practice, we have found that the sampled reaction paths are generally insensitive to the value of γ_1 as long as it sufficiently avoids generating reaction-paths with these spurious features.

The final ingredient in our reaction-path sampling scheme is encapsulated in the final two terms of Eq. 6, $W(\mathbf{r}_0, \mathbf{G}^0)$ and $W(\mathbf{r}_P, \mathbf{G}^P)$. As well as a set of $(3N_a - 6)$ coordinates, each end-point is associated with a *connectivity graph* (or adjacency matrix), $\mathbf{G}^0$ and $\mathbf{G}^P$, which describes the desired atomic connectivity. The elements of these graphs are defined as,

$$G_{ij} = \begin{cases} 1 & \text{if } r_{ij} < r_{ij}^{cut}, \\ 0 & \text{otherwise,} \end{cases} \tag{7}$$

where r_{ij} is the distance between atoms i and j and r_{ij}^{cut} is a cut-off distance which depends only on the atomic types of i and j (here, these cut-off distances were chosen based on preliminary *ab initio* calculations for the species involved, although we note that these values

can be selected based on other criteria such as covalent radii[20]). Thus, the end-point graphs describe the *chemistry* inherent in the system; by generating reaction-paths connecting end-points with different connectivity graphs, we implicitly have a method for sampling chemical reaction pathways. However, there is no guarantee that an atomic configuration at a given end-point will match the desired connectivity graph; for example, simulations using reactive PESs might allow bonding rearrangements to occur. This is where the potential energy contributions $W(\mathbf{r}_0, \mathbf{G}^0)$ and $W(\mathbf{r}_P, \mathbf{G}^P)$ come in to play; in particular, the terms $W(\mathbf{r}_0, \mathbf{G}^0)$ and $W(\mathbf{r}_P, \mathbf{G}^P)$ are *graph-enforcing potentials* which ensure that each end-point $\mathbf{r}_0$ and $\mathbf{r}_P$ can only adopt configurations which are consistent with the target connectivity graphs. There is some flexibility in choosing the exact form of $W(\mathbf{r}, \mathbf{G})$, and the current implementation of our reaction-path sampling approach uses the following constraint potential:

$$\begin{aligned} W(\mathbf{r}, \mathbf{G}) = \sum_{j>i} \Big[& \delta(G_{ij} - 1)[H(r_{ij}^{min} - r_{ij})\sigma_1(r_{ij}^{min} - r_{ij})^2 \\ & + H(r_{ij} - r_{ij}^{max})\sigma_1(r_{ij}^{max} - r_{ij})^2] + \delta(G_{ij})\sigma_2 e^{-r_{ij}^2/(2\sigma_3^2)} \Big] \\ & + V_{mol}(\mathbf{r}, \mathbf{G}). \end{aligned} \tag{8}$$

Here, $\delta(x)$ is the Dirac delta function, $H(x)$ is the Heaviside step function, and σ_1, σ_2 and σ_3 are user-defined constants. The summation in Eq. 8 runs over all pairs of atoms. The potential $W(\mathbf{r}, \mathbf{G})$ constrains *bonded* atoms to lie at bond lengths which are approximately in the range $\left[r^{\min}, r^{\max}\right]$, while *non-bonded* atoms are subject to a repulsive potential term which ensures that they remain non-bonded. Furthermore, the *molecular* constraint term $V_{mol}(\mathbf{r}, \mathbf{G})$ ensures that graphs describing several independent molecular species (easily determined from the connectivity matrix), cannot sample configurations which are incompatible with the expected number of molecules. In this work, we use

$$V_{mol}(\mathbf{r}, \mathbf{G}) = \sum_{\substack{j>i \\ m_i(\mathbf{G}) \neq m_j(\mathbf{G})}} [H(R^{min} - r_{ij})\sigma_4(R^{min} - r_{ij})^2, + H(r_{ij} - R^{max})\sigma_4(R^{max} - r_{ij})^2] \tag{9}$$

where R^{min} and R^{max} are, respectively, minimum and maximum allowed distances between a pair of atoms, each in a different molecule, and the label m_i identifies the molecule to which each atom i is assigned to, according to the connectivity matrix $\mathbf{G}$. Finally, we note that the sampled reaction paths are again insensitive to the choice of parameters σ_{1-4}, provided that the chosen parameters are sufficient to impose the appropriate graph constraints; here, these parameters were chosen based on our preliminary studies for systems such as formaldehyde.[23]

Together, Eqs. 3 to 9 define a Hamiltonian system describing a reaction path. The reaction path is defined by the end-point positions, $\mathbf{r}_0$ and $\mathbf{r}_P$, and a set of Fourier coefficients $\mathbf{a}$, as well as conjugate momenta for each of these variables; furthermore, the connectivity graphs $\mathbf{G}^0$ and $\mathbf{G}^P$ define the chemical bonding at the reaction end-points. Applying Hamilton's equations-of-motion, it is straightforward to determine time-evolution equations for $\mathbf{r}_0$, $\mathbf{r}_P$ and $\mathbf{a}$ (as well as their conjugate momenta); these can be integrated using standard methodologies, such as the velocity Verlet algorithm, in order to sample the conformational space associated with the reaction path. Furthermore, we note that standard MD approaches,[39–41] including efficient thermostatting and temperature-accelerated sampling,[33,34] can also be readily introduced.

However, while the Hamiltonian dynamics of Eqs. 3 to 9 enables conformational sampling of the reaction paths between *fixed* end-point chemical connectivities, to drive the search for reaction paths between different *chemical isomers* we introduce *graph moves* into our sampling scheme. Here, during the time-evolution of the reaction path, we introduce changes to the end-point graphs, $\mathbf{G}^0$ and $\mathbf{G}^P$, with a small probability P_u; specifically, at each time-step during evolution, we modify the end-point graphs if a random number $\eta \in [0, 1] \leq P_u$. The graph-moves allowed in the current work are described in more detail below, but can generally be viewed as simple "bit flips" in a small number of graph elements. Following graph update, the end-point configurations will no longer be consistent with the new connectivity graph; relaxation of the reaction-path under the action of the constraint potential $W(\mathbf{r}, \mathbf{G})$ then yields a configuration which is consistent. Following graph updates,

Hamiltonian sampling of the reaction-path resumes, although the reaction path will now be sampling reaction paths between updated chemical species at the end-points. In this way, combining Hamiltonian sampling and stochastic graph updates, our simulation system can automatically sample multiple reaction paths connecting multiple chemical species, thereby building up a picture of the allowed chemistry in a given system.

The enormous advantage of employing graph-based constraints is that it is straightforward to focus the search for reaction paths on the "chemically-relevant" region of graph space. For an N atom system, the total number of possible graphs which can be generated is $2^{N(N-1)/2}$; however, the large majority of these connectivity graphs will be chemically-irrelevant in the sense that they disobey basic atomic valence rules. For example, six-coordinate carbon atoms and four-coordinate hydrogen atoms simply do not contribute to observed reactive chemistry, so the graphs corresponding to such structures do not need to be sampled in assessing allowed reaction paths in a given system. To implement this valence screening in the current work, we simply reject graph moves if they violate valence sum rules. In particular, we implement the following valence constraints:

$$\begin{aligned} 0 &\leq \mathrm{val(H)} \leq 1, \\ 1 &\leq \mathrm{val(C)} \leq 4, \\ 4 &\leq \mathrm{val(Co)} \leq 6, \\ 1 &\leq \mathrm{val(O)} \leq 2, \end{aligned} \tag{10}$$

where val(X) indicates the atomic valence of X as determined from the connectivity graph. We note that these valence constraints play a similar role as energetic constraints implicit in previous reaction-path searching studies;[20,22,42] furthermore, these constraints can, in principle, be relaxed to allow more exotic chemical species to be generated. However, in the current case, where we are interested in assessing the extent to which a catalytic mechanism can be seen to emerge by combining atomistic and kinetic simulations, the constraints of Eq. 10 are employed.

As a final implementation issue, we note that all of the path-sampling simulations performed here employed self-consistent charge density-functional tight-binding (SCC-DFTB) to describe the PES of the reaction path ($V(\mathbf{r})$ in Eq. 6).[43–45] Our previous exploratory investigations,[23] as well as previous BHMC simulations,[20] have shown that this level of theory is sufficiently accurate to generate approximate molecular and TS structures for the catalytic system studied herein. However, we emphasize that SCC-DFTB is only used to generate *initial* reaction paths; these are subsequently refined at a higher-level of theory before further kinetic analysis, as described below.

In summary, Eqs. 3 to 9 define a system which enables conformational sampling of the reaction-path end-points as well as the reaction-path itself. Unlike other approaches, where connectivity graphs have been used to either analyze reaction products[20,27] or impose restrictions on sampled reaction-path end-points,[24] our methodology provides a route to *driving* the exploration of chemical (bonding) space in which an implicit reaction-path is always present between end-points. By combining graph-driven exploration of *chemical* space with Hamiltonian-based sampling of *conformational* space, our overall strategy provides a route to generating multiple reaction pathways connecting multiple chemical products; this is illustrated in Fig. 1. Once a set of initial reaction paths has been sampled, a range of further simulation approaches can then be used to refine these paths and model the resulting kinetic network; the approaches taken in this work towards these tasks are now described.

Structure optimization and transition-state searching

The graph-based sampling scheme described above results in a large number of reaction paths, connecting a range of different conformers and isomers; the second stage in our approach to computational kinetics is to determine optimised structures for all of the reaction-path end-points, as well as the relevant TSs.

With this aim in mind, the initial set of reaction-paths is first filtered to select a single representative reaction-path for each *unique* sampled pathway. For each remaining reaction

path, the end-point structures then undergo geometry optimization, again using SCC-DFTB. The reaction path itself, defined by the M intermediate structures, is also refined using the climbing-image nudged elastic band (CI-NEB[46]) approach. Overall, the CI-NEB refinement and energy minimization allow identification of locally-optimised structures for the reaction-path end-points, as well as determination of an (approximate) TS on the DFTB PES.

To generate a higher-quality description of the reaction-path stationary points, these SCC-DFTB-optimized geometries are further refined using a higher level of electronic structure theory. In the present case, we employ density functional theory (DFT) with a B3LYP exchange-correlation functional and a 6-31G(d,p) basis set. This choice was primarily motivated by the fact that DFT/B3LYP has already been shown to provide sensible relative energies and reaction barriers for the cobalt-catalyzed hydroformylation system to be studied here.[20,22] All calculations were performed using standard geometry optimization and TS location methods in *Gaussian03*,[47] starting from the molecular geometries identified by SCC-DFTB CI-NEB calculations. For each TS, normal-mode frequency calculations confirmed that a saddle-point had been located. These TS determination calculations generally proceeded smoothly with no user intervention, demonstrating that the SCC-DFTB CI-NEB calculations were sufficiently close to the DFT TS to be useful as starting points for optimization. Once stationary points for each reaction-path had been located, the relative free energies were calculated *via* the standard (harmonic) vibrational and (rigid) rotational methodology.

Kinetic network simulations

The graph-sampling strategy described above, combined with CI-NEB reaction-path refinement, geometry optimization and TS location, allows one to automatically generate a kinetic network model describing the reactive system. Here, the rates of all sampled reactions were determined using standard transition-state theory (TST[15,40,41]). Barrierless reactions were assumed to proceed at a diffusion-limited rate of $k_{diff} = 4.49 \times 10^{10}$ s^{-1}; this value is appro-

priate for the solvent (toluene) and temperature (423 K) assumed in the simulations below.[13] Furthermore, for barrierless reactions, the rate of the reverse reaction was chosen such that the correct equilibrium constant was obtained.

Once reaction rates had been determined, the kinetics of the network were simulated using the direct stochastic simulation algorithm (or Gillespie method[48–50]). Such kinetic Monte Carlo (KMC) methods[51,52] are well studied and have found extensive use in modelling, for example, large biochemical reaction networks[53,54] or interfacial processes such as adsorption, desorption and surface-mediated reactivity.[55–57] Here, following selection of the initial reactant concentrations, the time-evolution of chemical species concentrations is simulated such that reactive events occur with frequency correctly determined by the reaction rate and reactant concentration. As a result, these direct kinetic simulations can be used to interrogate both the mechanism and the rate law which emerges from our graph-sampled kinetic model. To the best of our knowledge this Article marks the first time that this methodology has been combined with an automated scheme for reaction path sampling with the specific aim of investigating a catalytic mechanism.

Results and discussion

The methodology described above represents a hierarchical approach to catalytic kinetics; starting from an atomic description of reaction paths, we identify and refine relevant stationary points (minima and TSs) and use these to build a kinetic model. In other words, automated graph sampling enables us to move from a microscopic sampling of reaction paths to macroscopic rate laws in a clear, well-defined manner.

The main result of this paper is the first direct application of our hierarchical approach to catalytic kinetics. Specifically, we consider the hydroformylation of ethene catalysed by $HCo(CO)_4$ as a challenging test case. Alkene hydroformylation[9–11,13,22,29–31] is one of the world's most important industrial chemical processes, with the annual global production

of resulting aldehydes and alcohols measured in millions of tons. In the case of cobalt-catalyzed hydroformylation, the active catalyst is $HCo(CO)_4$, formed *in situ* by dissociation of $Co_2(CO)_8$; the generally-accepted mechanism of Heck and Breslow[9] then proceeds as shown in Fig. 2. Following dissociation of CO, generating $HCo(CO)_3$, the alkene coordinates and subsequently inserts into the Co-H bond, forming a cobalt-alkane species. Coordination of CO and insertion into the Co-C bond is then followed by oxidative addition of hydrogen, and reductive elimination finally leads to aldehyde product formation and regeneration of the catalyst. The reaction is typically run at temperatures of 150°C and pressures of *ca.* 50 bar, usually in a solvent of toluene.[10,13] Furthermore, as in all previous mechanistic studies of this system, we assume that the reaction proceeds *via* singlet ground-states, an assumption which is further reinforced by work suggesting that the lowest-lying triplet state for the cobalt catalyst lies around 100 kJ mol^{-1} higher than the singlet ground-state.[58]

As noted in the Introduction, the experimental rate-law for cobalt-catalyzed hydroformylation is generally found to be of the forms shown in Eq. 1. Recently, Harvey and coworkers[13] were successful in deriving a rate law (Eq. 2) which went some way to reproducing experimental observations, but this rate law was derived by (i) making assumptions about the mechanism and the relative importance of side-reactions, and (ii) adopting steady-state assumptions for selected reactant species. However, this paper marks an attempt to derive a catalytic rate law in an *automated* fashion; in other words, we are seeking to determine whether the experimental rate law can be derived in a unguided, hands-off simulation approach which starts from first principles, with only weak chemical constraints. If successful, this approach could mark a shift away from *assuming* a mechanism (and then testing whether it is correct) towards enabling the reaction rate law and associated mechanism to naturally *emerge* from simulations of reaction kinetics.

Our reaction-path sampling simulations began with four molecular fragments, H_2, CO, C_2H_4 and $HCo(CO)_3$, arbitrarily arranged in space (a total of $N_a = 18$ atoms). The initial start- and end-points of our reaction path were simply chosen to be the same configurations;

the natural dynamics implicit in the Hamiltonian of Eq. 3, as well as graph-sampling moves, mean that these end-points rapidly diverge as the simulation proceeds. The reaction-path was discretized using $P = 18$ images. The Hamiltonian equations-of-motion were integrated using the standard velocity Verlet algorithm with a time-step of 0.25 fs; we note that this time-step is comparable to that typically used in *ab initio* simulations of reactive chemistry. A simple Anderson thermostat was used to maintain the temperatures of the reaction-path end-points and Fourier coefficients at $T = 300$ K. Each reaction-path sampling trajectory was run for up to 10^4 time-steps. The reaction-path was stored every 250 time-steps for subsequent CI-NEB analysis; CI-NEB calculations were performed with the standard approach using steepest descent optimization. The remaining parameters for the reaction-path sampling simulations were the same as employed in our previous methodological work.[23]

The graph moves employed in these simulations included the following transformations:

1. Single bit-flip. Select a random off-diagonal element G_{ij}; if $G_{ij} = 1$, replace it with 0, and if $G_{ij} = 0$, replace it with 1;

2. 1,2-insertion: M-X + Y-Z $\rightarrow$ M-Y-Z-X. Note that the Y-Z bond remains intact;

3. Metathesis: M-X + Y-Z $\rightarrow$ M-Y + X-Z;

4. 1,1-insertion: Z-X + Y $\rightarrow$ Z-Y-X. Here, Y represents an arbitrary species, not just a single atom; the important point to note is that both Z and X are bonded to the *same* atom in Z-Y-X.

In the above, M represents the cobalt atom, and Y and Z are arbitrary ligands. The inverse of each of the above transformations was also incorporated as possible graph moves, and each of the total set of graph moves was performed with equal probability during each simulation. As in our previous work,[23] we note that the set of graph moves considered here are comprised of common organometallic transformations, and are in no way tailored to the $HCo(CO)_4$-catalyzed system; these graph moves, as well as the existence of an arbitrary

"bit-flip" move, do not exclude the discovery of "exotic" chemical species, as shown in the set of sampled structures below. However, we note that the choice of graph-moves has an important influence over the overall efficiency of our graph-sampling scheme. For example, employing only random bit-flips would most commonly produce structures in which the valence constraints noted above are violated; as a result, such graph moves would be rejected, resulting in poor overall sampling of the chemical space associated with the connectivity graphs. Instead, the graph-moves noted above generally have a much greater chance of avoiding violations of valences and thus being accepted.

To further ensure that a full-range of reaction-paths were sampled for the system, we adopt a "cascade" strategy. Here, when a new molecular structure has been generated as a result of a graph-move, this new structure was then used to initialize further reaction-path sampling calculations. This process was repeated until no further reaction-paths were sampled, at least within the typical time-frame of the simulations. We note that this strategy cannot *guarantee* that we have sampled *all* possible reaction-paths consistent with the chemical constraints described above; however, visualization of the sampled configurations suggests that no obvious chemically-important reactions are missing from our collection of reaction-paths. The challenge of knowing when to stop sampling is akin to that encountered in global optimization; technically, one does not know whether the global minimum of an optimization function has been located without sampling *every* local minimum. In higher-dimensional problems, this is usually impossible, so the best one can achieve is to perform repeated optimizations and be satisfied with the best result available. Furthermore, we note that this cascade process for sampling reaction paths starting at different molecular structure could be trivially parallelized, a strategy which we are currently exploring. In total, we performed 45 independent graph-sampling simulations, with each sampling between 2×10^3 and 10^4 time-steps. In practice, many of these simulations were redundant in that they resulted in reaction-paths which had been previously discovered; in fact, only around 20 sampling simulations resulted in truly unique reaction-paths. The efficiency of SCC-DFTB ensured

that these graph-sampling simulations could be performed in a short time. For example, each individual sampling simulation took around 12 hours; employing up to ten independent computational nodes, the sampling simulations could be performed in a few days. Of course, while our reaction-path sampling approach is not particularly tied to SCC-DFTB, we note that using a more expensive *ab initio* method during sampling, such as DFT, would clearly increase simulation times (hence our choice of SCC-DFTB in this case).

Minima and Transition-States

Ultimately, our graph-based sampling scheme produced a total of 31 unique reaction products and 32 reaction paths; of these reaction paths, five were found to be barrierless. TS structures were located for the 27 remaining activated reactions. All reaction products (minima) and TS structures were optimised using DFT (B3LYP, 6-31G(d,p)) calculations, and the associated free energies were determined using the standard translational, rotational and (harmonic) vibrational thermal corrections. The thermal corrections were calculated assuming that the temperature was $T = 423$ K (150°C), which is typical of the operating temperature of cobalt-based hydroformylation setups, and is the same temperature as considered in our kinetic simulations below. The 31 reaction products are shown in Fig. 3, and the relative free energies of these molecules are given in Table 1. Furthermore, Fig. 4 shows the relative free energy profiles for each of the sampled reactions; as would be expected in any complex reactive system such as that studied here, we find a range of exothermic and endothermic reactions, as well as TS free energies which span a wide range from *ca.* 1 to *ca.* 420 kJ mol^{-1}. The information in Table 1 and Fig. 4 is the raw input data used to directly generate our kinetic network model for this catalytic system.

The structures shown in Fig. 3 span the range of chemically-relevant products which would be expected for this system. For example, we locate structure S10, the starting point for the catalytic cycle, and find that the reaction S1 + S3 → S10 is barrierless as generally expected. We also determine structures corresponding to coordination of the ethene (S4) to

the active cobalt catalyst (S1), giving rise to two different products, S13 and S18, with the coordinated alkene lying either parallel to or perpendicular to the cobalt-hydrogen bond. Further structures corresponding to the products of alkene insertion (*e.g.* S6) and CO insertion (*e.g.* S12) are also located. Finally, we note that the final aldehyde product is located (S26), as is an alternative alkane product formed by hydrogenation of ethene (S27).

The set of reaction products (Fig. 1) and sampled reaction paths can be combined to create a direct visualisation of the entire kinetic network associated with the $HCo(CO)_3$ + CO + H_2 + C_2H_4 system; this is shown in Fig. 5. Such kinetic networks can be useful in developing a broad overview of possible chemical reactions available. For example, Fig. 5 directly illustrates that the aldehyde product S26 can be reached *via* four pathways, corresponding to reductive elimination from S23, S23, S30 or S31, respectively. We note that these four different reactions were *automatically* generated, in contrast to the more common approach of using human chemical intuition to guide the search for the relevant reactive pathways. Most importantly, this picture of the chemical kinetic network is key to further analysis, notably determination of the dominant catalytic mechanism and the associated phenomenological rate law, as will now be discussed.

Extracting the Catalytic Mechanism

Given that we have now constructed a kinetic model describing the $HCo(CO)_3$ catalytic hydroformylation system, we can ask the following question: what is the predicted mechanism of ethene hydroformylation? We emphasise that our approach is somewhat different in philosophy from the usual route to interrogating reaction mechanisms, whereby one suggests a mechanism and then uses simulations of reaction energy barriers and energy changes to justify whether the suggested mechanism is plausible. In contrast, our reaction network model (Fig. 5) describes a range of possible reaction pathways; for example, the following pathways, in which the cobalt-containing species are indicated, all lead to formation of the

aldehyde product,

$$\begin{aligned} &\text{S10} \rightarrow \text{S1} \rightarrow \text{S18} \rightarrow \text{S6} \rightarrow \text{S11} \rightarrow \text{S12} \rightarrow \text{S30} \rightarrow \text{Product S26} \\ &\text{S10} \rightarrow \text{S1} \rightarrow \text{S13} \rightarrow \text{S6} \rightarrow \text{S11} \rightarrow \text{S12} \rightarrow \text{S31} \rightarrow \text{Product S26} \\ &\text{S10} \rightarrow \text{S9} \rightarrow \text{S20} \rightarrow \text{S8} \rightarrow \text{S28} \rightarrow \text{Product S26}. \end{aligned} \tag{11}$$

Thus, kinetic modelling of this system will enable us to directly analyze the relevant reaction pathways to deduce the mechanism; in other words, the predicted mechanism will automatically emerge from the set of graph-sampled reaction paths without presupposing a catalytic cycle.

Using the direct SSA methodology described previously,[48–50] we simulated the kinetics of the network illustrated in Fig. 5. Here, the simulation volume was chosen to be $v = 5 \times 10^{-19}$ m^3, and the total simulation time was $T = 10^{-2}$ s; in passing, we note that our reaction-path sampling scheme, which employs picosecond time-scale simulations, has been used to generate a kinetic model which enables simulations on experimentally-relevant time-scales beyond the millisecond regime. The initial concentrations (or pressures, for gaseous reactants) of the reactant species S2 (H_2), S3 (CO), S4 (C_2H_4) and S10 ($HCo(CO)_4$) were 46.7 bar, 46.7 bar, 1.33 mol dm^{-3} and 0.013 mol dm^{-3}, respectively. Again, these initial concentrations and pressures were chosen to be representative of typical conditions employed in $HCo(CO)_4$-catalyzed hydroformylation.[10,13]

Figure 6 shows the net reactive flux through the various sampled reactions (Fig. 5). The net flux for reaction k was determined as the difference between the number of forwards and backwards reactive events,

$$F_k = \frac{n_k^f - n_k^b}{N_r}, \tag{12}$$

where N_r is the total number of observed reactive events and $n_k^{f/b}$ are the number of forward/backwards reactions which occur during the time-scale of the SSA modelling.

The most important conclusion from Fig. 6 is that the generally-accepted Heck-Breslow

mechanism[9] (Fig. 2) *automatically* emerges from direct kinetic simulations of our sampled reaction network. First, significant flux is observed from S10 to S1+S3; this reaction corresponds to dissociation of CO from the starting catalyst $HCo(CO)_4$, leading to the active species $HCo(CO)_3$. This species then undergoes alkene coordination and insertion, CO insertion and H_2 insertion, ultimately leading to elimination of the product aldehyde S26. We do not observe formation of product S26 by any other series of reactions; the Heck-Breslow mechanism is automatically implied in our simulations. Thus, our kinetic network model of Fig. 5 can be reduced to a simpler version, as shown in Fig. 7; this minimal kinetic model can be seen to map exactly onto the Heck-Breslow mechanism of Fig. 2. It should be emphasized, however, that this reduced mechanistic model was determined automatically from the set of all possible reactions illustrated in Fig. 2.

The results of Fig. 6 also imply several further interesting observations. We find an indication that S9 and S10 exist in equilibrium (zero net flux), although analysis indicates that the forward flux through this reaction is around six orders of magnitude smaller than the reaction leading to S1+S3; in other words, direct insertion of a CO moiety into the Co-H bond is irrelevant to further reaction. Furthermore, we find that alkene coordination can lead to two different products differing in the orientation of the alkene; S13 has the alkene C-C bond lying parallel to the Co-H bond, whereas S18 has this bond lying perpendicular to the Co-H bond. The calculated flux indicates that the route leading to S18 is very slightly favoured, although both routes are ultimately followed in product formation. Finally, we note that the coordination of the alkene S4 to the Co-COH species S9 is also observed, although the reversibility of this reaction is again implied by the zero net flux, and this pathway does not lead to further reaction.

As a final point, we highlight the fact that the aldehyde product S26 is formed exclusively by reductive elimination from S30 in our kinetic simulations; no elimination is observed from S31. Closer analysis of the free energies of S12 and S29 demonstrates that S29 is around 25.2 kJ mol^{-1} lower in energy; however, the free energy barrier to hydrogen addition to S29

(leading to S31) is 92.2 kJ mol^{-1}, whereas the barrier to addition to S12 (leading to S30) is 67.0 kJ mol^{-1}, indicating a kinetic preference for the reaction pathway which ultimately leads to S30 and then elimination of S26. Again, we note that our automated graph-sampling methodology enables one to address these mechanistic questions in a relatively straightforward manner; the only provisos are (i) whether all relevant reaction paths have been sampled, and (ii) whether the calculated free energy values for minima and TSs are sufficiently accurate to distinguish between plausible mechanisms. The first of these challenges has already been noted above; the second challenge plays a role in the calculated rate law, as we now discuss.

Phenomenological Rate Law

As well as elucidating the reaction mechanism, direct SSA kinetics simulations can also be used to discern the rate law for the reaction. To do this, we calculate the reaction rate as a function of concentration of each reactant species (S2, S3, S4 and S10); here, the reaction rate is defined as the rate of change of the concentration of the product aldehyde, S26. These simulations were performed in the same manner as those used to analyze the catalytic mechanism; unless otherwise stated, the concentrations of those reactant species which are not varying were the same as given in the previous section. Furthermore, note that the *full* kinetic network of Fig. 5 was used in all kinetic simulations performed here.

Figure 8(a) illustrates the time-dependence of the concentration of the aldehyde product during three independent SSA simulations; the average result is also shown. By taking the slope of the average line between $t = 8 \times 10^{-3}$ s and $t = 10^{-2}$ s, we find that the calculated reaction rate is 3.9×10^{-5} mol dm^{-3} s^{-1}. While modern experimental kinetic data for ethene is somewhat difficult to establish, data is available for hydroformylation of propene under similar experimental conditions to those considered here.[10] In the case of propene, reaction rates are typically of the order of 10^{-3} mol dm^{-3} s^{-1}; in other words, the rate of ethene hydroformylation predicted from our reaction network is around a factor of 25 slower than

the typical experimental rate of propene hydroformylation. Although the rate of ethene hydroformylation would clearly be expected to differ from that of propene, not least because the propene system offers the possibility of linear and branched product formation, one might expect that the calculated ethene hydroformylation rate would at least be of a similar order of magnitude to these experimental results.

As a result of this comparison, it is interesting to investigate why our predicted rates are lower than the corresponding experimental propene hydroformylation system. Given that the correct (*i.e.* generally-accepted Heck-Breslow) *reaction mechanism* has been identified in our graph-sampling/kinetics simulation scheme, the only conclusion can be that the calculated relative free energies and barriers are in error. To investigate this point, we consider the data shown in Fig. 8(b), which shows calculated free energies for several molecular structures and reaction barriers as determined in this work and in the work of Morokuma and coworkers.[22] In both cases, DFT (with standard thermal corrections) was used to calculate the minima and TS free energy values; however, different exchange-correlation functionals were employed in the two sets of calculations (M06 and B3LYP). As a result, one would not expect perfect agreement between the two sets of calculated results, and the diagonal line in Fig. 8(b) is instead intended as a guide to the eye. As expected, we find broad correlation between the two sets of DFT results, with the differences in calculated free energies typically being 5-20 kJ mol^{-1} (not uncommon for DFT comparisons). However, we also find two significant outliers compared to the previous AFIR simulation study (indicated by black arrows in Fig. 8(b)): (i) S30 has $G = +25.1$ kJ mol^{-1} in this work, but $G = -21.8$ kJ mol^{-1} in the AFIR study, and (ii) S26 (product aldehyde) has $G = -36.7$ kJ mol^{-1} here, but $G = -88.9$ kJ mol^{-1} in the AFIR study.

As noted above, S30 acts as the main gateway structure to the formation of the product aldehyde S26 in our predicted catalytic mechanism; as a result, errors in the energies and barriers associated with S30 may be expected to play a significant role in the observed rate for product formation. However, the absolute energy value of S30 is not a particularly

significant error with regards to the reaction mechanism and reaction rate (for example, the barrier height for reductive elimination from S30 is 6.35 kJ mol^{-1}, comparable to the value of 7.3 kJ mol^{-1} from the AFIR study). The most notable possible error lies in the fact that the free energy of S30 lies above the TS for S2 + S12 $\rightarrow$ S30; this in turn suggests that the reverse reaction (S30 $\rightarrow$ S2 + S12) can effectively compete with product formation. Furthermore, this competition is also suggested by the low flux related to product formation, as shown in Fig. 6.

To correct this error, we chose to reduce the free energy of S30 by 37 kJ mol^{-1} while keeping all other reaction barriers and relative energies the same as in the previous calculations. We emphasize here that this *ad hoc* correction by hand is in no way condoned as an acceptable strategy. Instead, our goal is to demonstrate that the kinetic network sampled by our graph-based strategy is representative of experimental rate results if one accounts for inaccuracy in the *ab initio* reaction rate data; however, we note that this correction does not change the emergent reaction mechanism, as highlighted below. A much more appealing long-term strategy will be to combine our hierarchical kinetics modelling strategy with uncertainty quantification approaches for assessing predictive reliability; this is an approach we will explore in the future.

The effect of reducing the energy of S30 is that the reverse (dehydrogenation) reaction (S30 $\rightarrow$ S2 + S12) is slowed significantly, and thus the net flux towards product is increased; this is illustrated in Fig. 9, which compares reactive flux in the original and modified networks. Importantly, we also find that reducing the energy of S30 has no effect on the mechanism of reaction, and the same set of reaction channels are sampled during our kinetics simulations. Instead, the effect of reducing the energy of S30 is to increase the flux towards product, as expected, leading to more rapid formation of product (Fig. 9).

With the most important error in the calculated relative energies identified and corrected, we are now in a position to investigate the form of the rate law predicted by our kinetic network. Figure 10 illustrates the rate of product formation as a function of the

concentration of the initial reactants. Most importantly, we find that our kinetic simulations yield a rate dependence for each initial reactant which is in broad agreement with previous experimental and computational studies.[10,11,13] As shown in Fig. 10, we find that the rate of aldehyde formation is first-order with respect to the partial pressure of H_2 and the concentration $[HCo(CO)_4]$. In the case of the alkene, we observe a dependence of the form $[C_2H_4]^{0.55}$; however, if we limit the fitting to the range of concentrations previously studied in experiments[10] on propene ([alkene] > 0.6 mol dm^{-3}), we observe a dependence which is much closer to first-order, and thus similar to that observed previously.

The most complex rate-dependent behaviour is observed for CO. A fit of the form $a\,[\mathrm{CO}]\,/(1 + b\,[\mathrm{CO}])^2$, the rate expression suggested in some previous studies, gives a relatively poor reproduction of the simulation data, while a function of the form $a\,[\mathrm{CO}]\,/(1 + b\,[\mathrm{CO}] + c\,[\mathrm{CO}]^2)$ gives a much better agreement with the calculated rates. There are two points to make here: (i) the experimental rate law of Bourne and coworkers[10] was determined based on measurements at just three independent alkene concentrations, and none of these concentrations was sufficiently low to observe the rate at very low concentrations, and (ii) we cannot rule out the fact that errors in the DFT barrier heights and relative energies are playing a role in the calculated rate dependence, as in all simulation studies. However, and perhaps most encouragingly with regards to the overall aim of this study, we find that the calculated rate law is physically consistent with all previous experimental and computational studies, exhibiting a complex dependence on CO concentration with reaction rate increasing at low CO concentrations (up to a point) and decreasing at higher CO concentrations. Our automatically-generated rate calculations have therefore proven capable of reproducing one of the key experimental feature of this system, without pre-supposing anything about the reaction mechanism itself or making any steady-state assumptions.[13,22]

Finally, it is interesting to investigate what the calculated flux tells us about the reaction mechanism in the high and low CO pressure regimes. As shown in Fig. 11, we find that the set of reactions with non-zero forwards or backwards flux does not change in the high and low

CO pressure regimes. The most important difference between these two regimes can be found in the sub-set of reactions which are associated with alkene insertion into the Co-H bond of S18 and S13. In particular, we find that, in the low CO pressure regime, the reactive flux associated with alkene insertion (S13 $\rightarrow$ S6) is *negative*, implying that the reverse reaction dominates; in contrast, in the high pressure regime, the reactive flux S13 $\rightarrow$ S6 is positive. These results serve to illustrate that, when CO pressure is low, the coordination of CO to form S11 (S6+S3 $\rightarrow$ S11), and subsequently proceed to aldehyde product, is hindered, thereby directly indicating a mechanism for the linear dependence on p_{CO} at low pressures.

Conclusions

This Article has presented a "bottom-up" approach to determining the mechanism and associated rate law of a complex catalytic cycle from first-principles; to the best of our knowledge, this is the first time such an approach, integrating automated reaction-path sampling and kinetic simulations, has been successfully demonstrated.

The advantages of our computational strategy are that it is highly automated, unbiased and unguided. This is in strong contrast to more standard routes to investigating catalytic mechanisms, which generally focus on (i) proposing a mechanism, and then (ii) using computer simulations to assess whether the proposal is consistent with experiments. In contrast, in our approach, we automatically generate a kinetic network describing possible reactions in the system, and then use direct kinetic simulations which enable the mechanism and rate law to naturally emerge from all the available possibilities. We have shown in this Article that our approach is capable of elucidating the Heck-Breslow mechanism of catalytic hydroformylation, as well a phenomenological rate law which is comparable to those derived from previous experimental and theoretical investigations.

Importantly, our simulation approach is computationally-efficient; relatively inexpensive DFTB calculations are used to drive the search for reaction-paths, while a higher level of

theory (DFT B3LYP in this case) can be used to optimize the PES stationary points in order to generate reactions rates for subsequent kinetic modelling. Current work is focussed on further streamlining our overall methodology by (i) code parallelization, and (ii) developing strategies to focus the reaction-path search on unique reactions, avoiding redundancy in sampled paths. These additions promise to offer significant reductions in total calculation times.

However, while our sampling strategy has clearly proven successful in elucidating the reaction mechanism in this case, it is clear that the underlying problem of PES accuracy can be influential. In particular, we have found that the relative inaccuracy of DFT led to predictions for the overall reaction rate which were too low; we found that an *ad hoc* change of a single barrier and associated reaction rate brought the predicted reaction rate into better agreement with experimental results, although this approach is, of course, not acceptable in general. Instead, we are currently investigating the application of uncertainty quantification methods aimed at assessing the overall reliability of the kinetic network models generated by our graph-based sampling approach and *ab initio* calculations; combined with related methods such as sensitivity analysis,[52,59] these strategies have potential to at least highlight when problems relating to accuracy of calculated rates might be expected to be important.

Furthermore, there are clearly remaining opportunities to improve our reaction-path sampling strategy. For example, as well as computational refinements, such as the adoption of parallel computing methods, methodological improvements such as the use of temperature-accelerated sampling strategies[33,34,41] would also be expected to improve the range of application of our strategy. Such refinements, particularly accelerated sampling, will help in accessing a wider range of possible reaction paths, including the "roaming" pathways which have been identified as being key to the dynamics of a number of molecular species.[60–62] As a further challenge, we are also investigating methods for addressing reaction paths associated with changes in electronic state; however, we also note that such simulations will also require more accurate (multi-reference) electronic structure methods, which will have clear

consequences for computational expense which must be similarly addressed. As above, these directions are work in progress.

Finally, we note the exciting possibility that our simulation strategy can be used to *optimize* catalytic function; for example, once a kinetic network model has been generated for a given catalytic system by graph-based reaction-path sampling, investigating the influence of ligand substitution on the catalytic kinetics should be reasonably straightforward. This is again an area of research which we are now exploring.

Acknowledgement

The author is grateful to the Centre for Scientific Computing at the University of Warwick for providing computational resources.

Table 1: Calculated free energies (at DFT/B3LYP/6-31G(d,p) level) for structures sampled in graph-based reaction-path sampling for H_2 + CO + C_2H_4 + $HCo(CO)_3$. In each case, the free energy relative to the sum of free energies of the constituents is given; for example, $G_{rel}(S10) = G_{calc}(S10) - G_{calc}(S1) - G_{calc}(S3)$. For reference, these relative free energies were determined using the following calculated values: $G_{calc}(S1) = -1723.260201\ E_h$, $G_{calc}(S2) = -1.186302\ E_h$, $G_{calc}(S3) = -113.333215\ E_h$ and $G_{calc}(S4) = -78.576220\ E_h$.

Structure	**G / kJ mol^{-1}**
S5	24.91
S6	-39.48
S7	40.55
S8	-9.94
S9	3.71
S10	-89.46
S11	-73.32
S12	-46.74
S13	-16.44
S14	44.18
S15	44.78
S16	93.54
S17	42.96
S18	-28.77
S19	97.30
S20	19.77
S21	78.55
S22	27.27
S23	98.72
S24	68.99
S25	18.83
S26	-36.71
S27	-95.59
S28	94.50
S29	-71.70
S30	25.13
S31	17.71

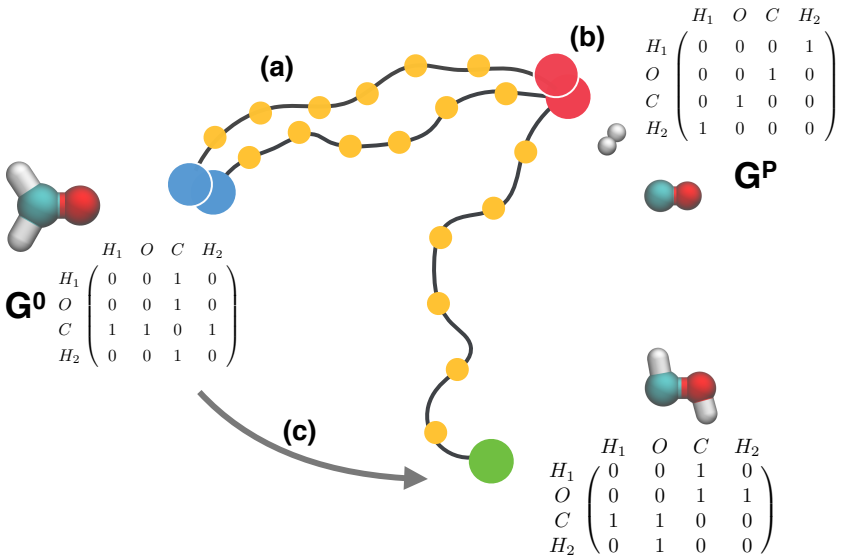

Figure 1: Schematic overview of graph-based reaction-path sampling,[23] demonstrated here for the four-atom H_2CO system. Our approach comprises (a) a dynamic reaction-path defined as a set of images (yellow dots) lying along a path described by a set of Fourier coefficients, and (b) connectivity graphs ($\mathbf{G}^0$ and $\mathbf{G}^P$) describing the bond arrangement at the end-points of the reaction-path. During the Hamiltonian-based dynamic trajectories of the reaction-path, periodic graph moves, (c), drive the search for new paths connecting different chemical species; in this way, one can straightforwardly sample reaction paths in a given chemical system.

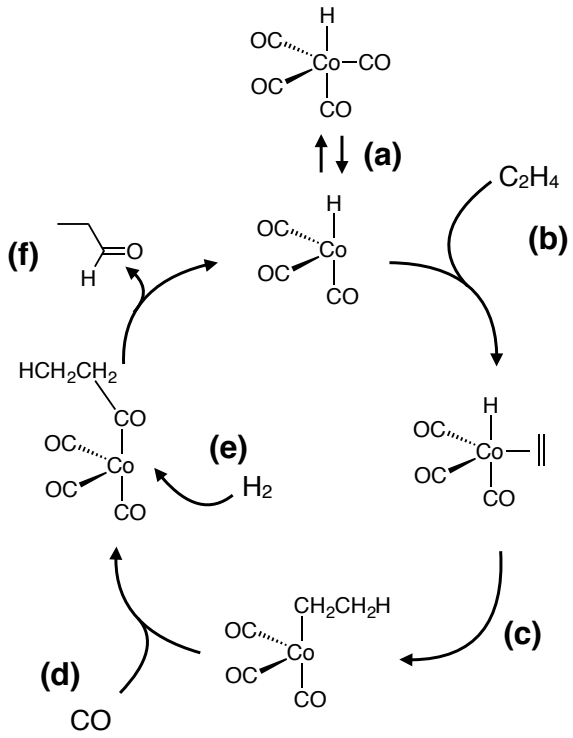

Figure 2: The Heck-Breslow mechanism for alkene hydroformylation.[9] The initial species $HCo(CO)_4$ undergoes CO dissociation (a) to generate the active catalyst. Subsequently, (b) the alkene coordinates and (c) inserts into the Co-H bond. Coordination and insertion (d) of CO is followed by addition of H_2 (e). Finally, reductive elimination leads to formation of the product aldehyde and regeneration of the catalyst (f).

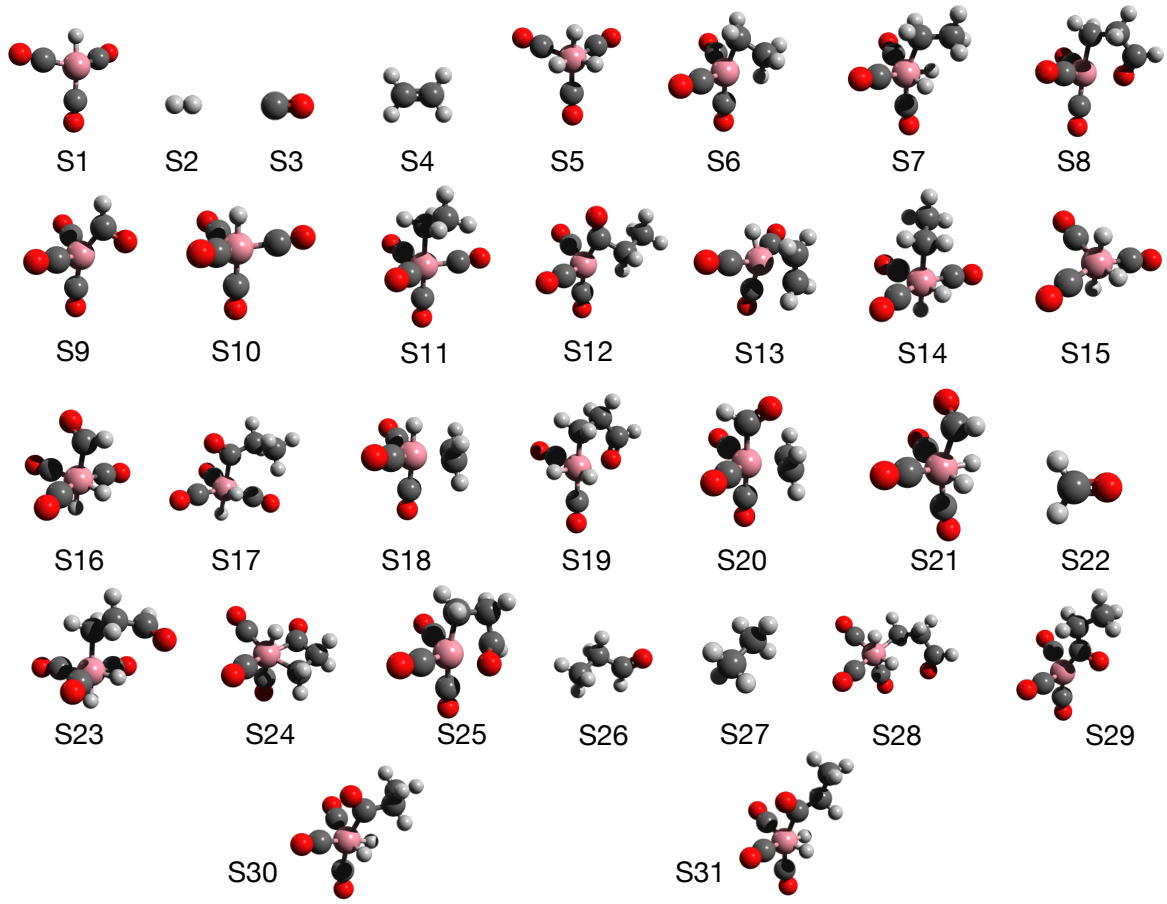

Figure 3: Molecular structures sampled during graph-based reaction path search; free energy values are given in Table 1. All structures were generated during DFTB-based path-sampling, and subsequently optimised with DFT (B3LYP 6-31G(d,p)).

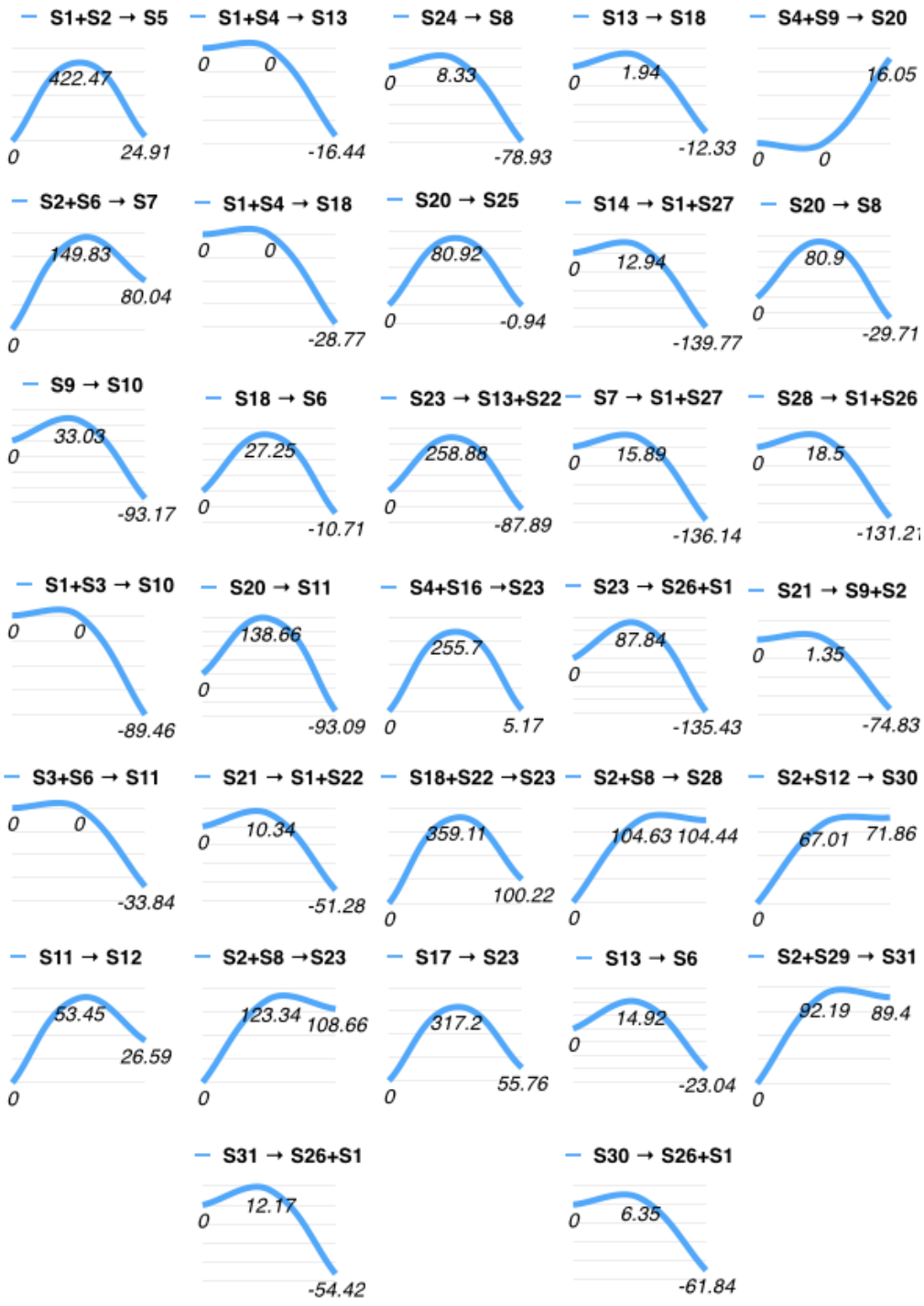

Figure 4: Free energy reaction profiles for the set of sampled reaction paths in the H_2 + CO + C_2H_4 + $HCo(CO)_3$ system. In each case, the free energy of the products is given relative to the reactants; all energies are given in kJ mol^{-1}. In cases where the TS energy is given as zero, the reaction was found to be barrierless.

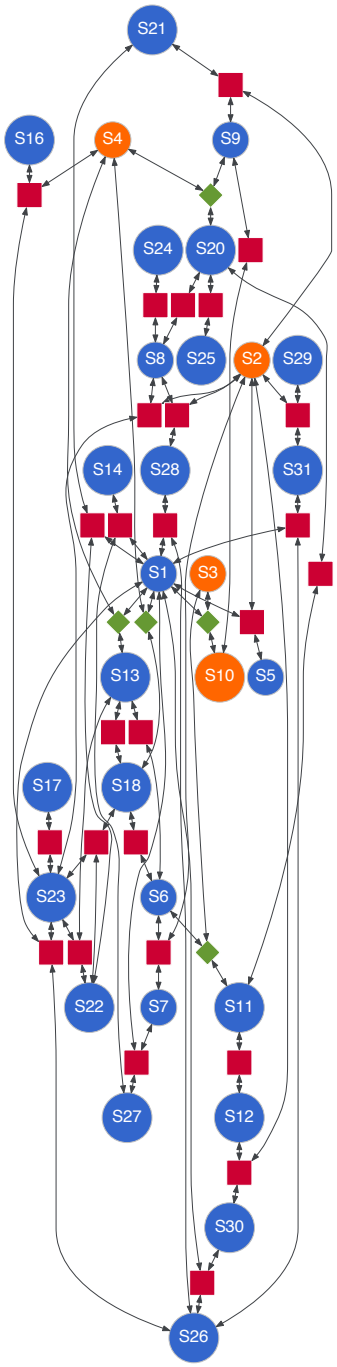

Figure 5: Kinetic network model generated for the H_2 + CO + C_2H_4 + $HCo(CO)_3$ system.[63] Each blue node represents one of the sampled structures shown in Fig. 3. The arrows connect structures which are accessible *via* either a TS (red squares) or a barrierless reaction (green diamonds). The initially-populated chemical species used in the kinetics simulations reported here are shown in orange (S2, S3, S4 and S10), while the aldehyde product S26 is shown at the bottom of the network graph.

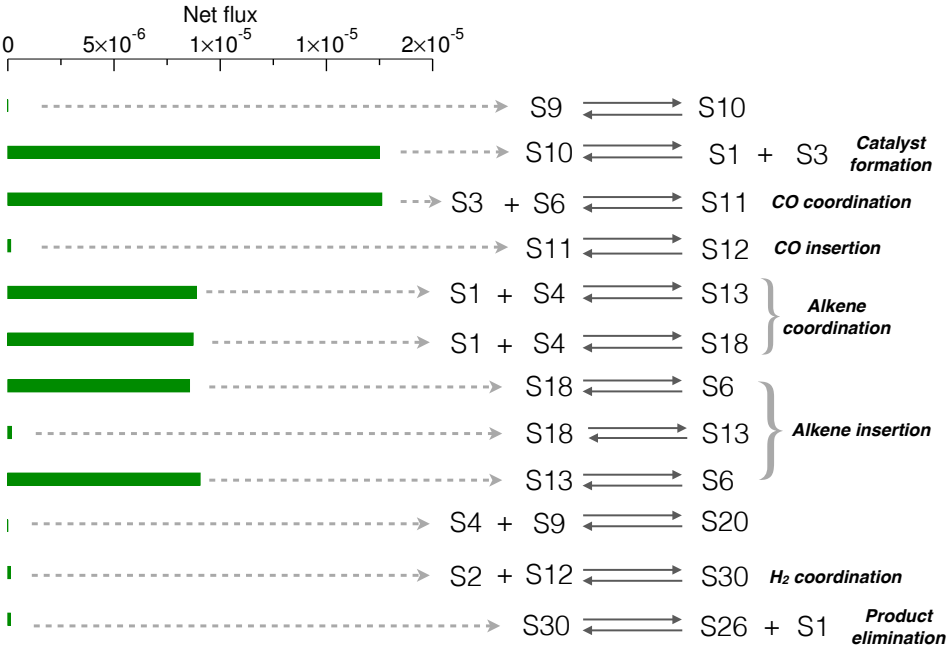

Figure 6: Reactive flux calculated for the graph-sampled kinetic network shown in Fig. 5; the flux is calculated using Eq. 12, and the initial concentrations of reactants is as described in the main text. Note that all other reaction paths exhibited zero net flux; in other words, the reactions shown here were the only reactions which took place during the kinetic modelling.

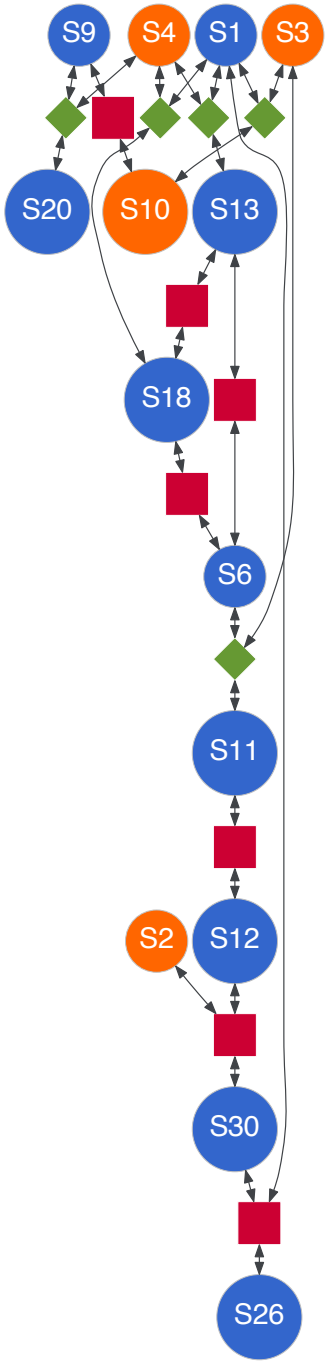

Figure 7: Minimal kinetic network model describing catalytic hydroformylation of ethene. This network was derived by removing all nodes and edges which represented molecular structures and reaction paths which were not sampled in direct stochastic simulations of the full network of Fig. 5. The remaining structures and reactions shown here can be directly mapped onto the Heck-Breslow mechanism of Fig. 2. The color scheme is the same as in Fig. 5.

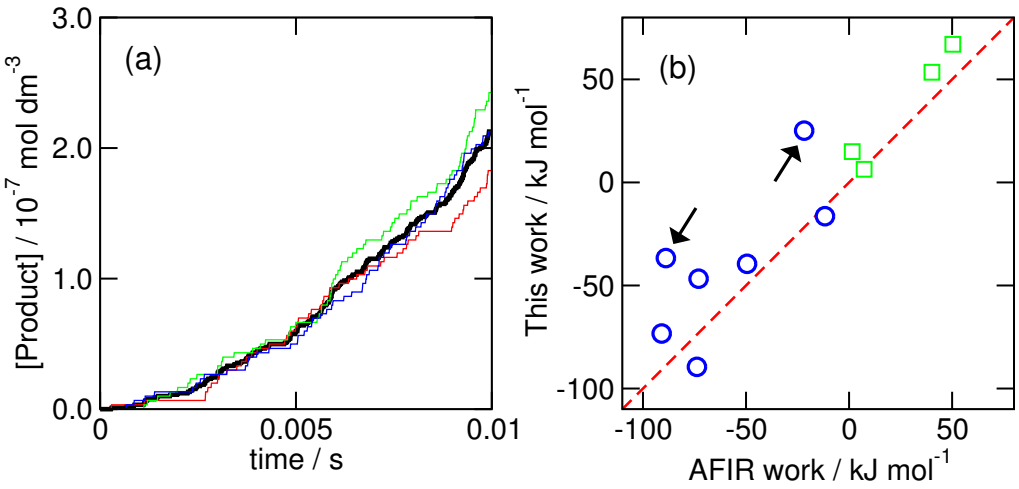

Figure 8: (a) Time-dependence of the concentration of S26, the product aldehyde. Three independent stochastic kinetics simulations are shown (thin lines), as is the average behaviour (thick black line). The slope between $t = 8 \times 10^{-3}$ s and $t = 10^{-2}$ s was used to determine the reaction rate. (b) Correlation between free energies of minima (blue circles) and TSs (green squares) determined in this work and in a previous AFIR study using different DFT functional and basis set.[22] The red dashed line is simply a guide for the eye; one would not expect perfect correlation between these two sets of results, although there is clearly broad agreement. The black arrows indicate minima S26 and S30, which exhibit the largest difference relative to the previous work.

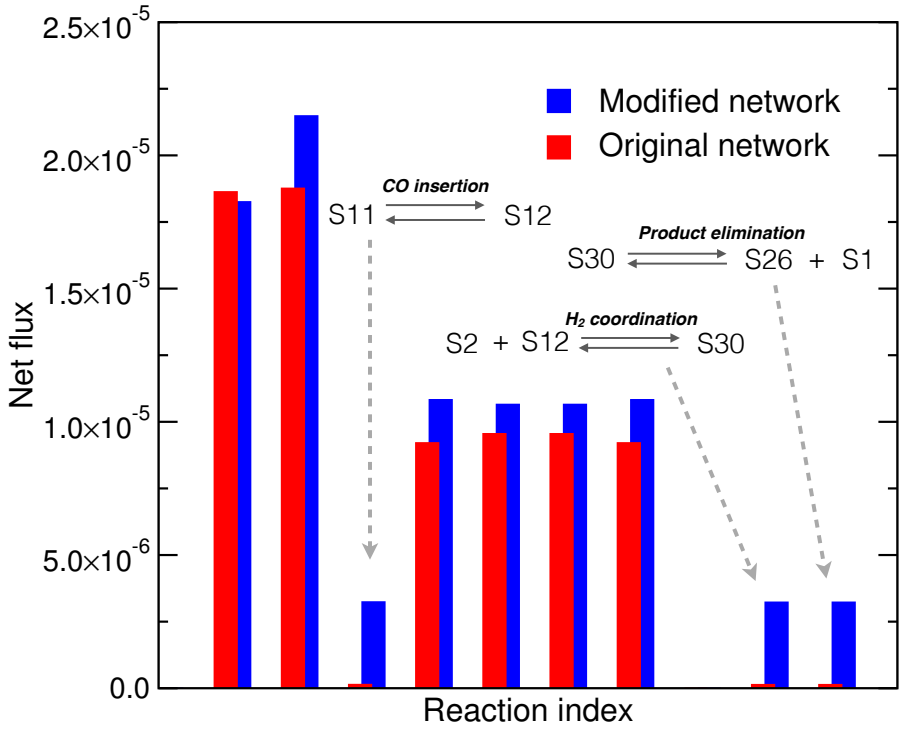

Figure 9: Calculated net flux in original (red) and modified (blue) kinetic network models; as noted in the text, the only difference between these models is that the free energy of S30 has been lowered by 37 kJ mol^{-1}. The reactions are shown in the same order as in Fig. 6, and no additional reactions are sampled in the modified network. The most important differences responsible for the increase in reaction rate in the new model are highlighted explicitly.

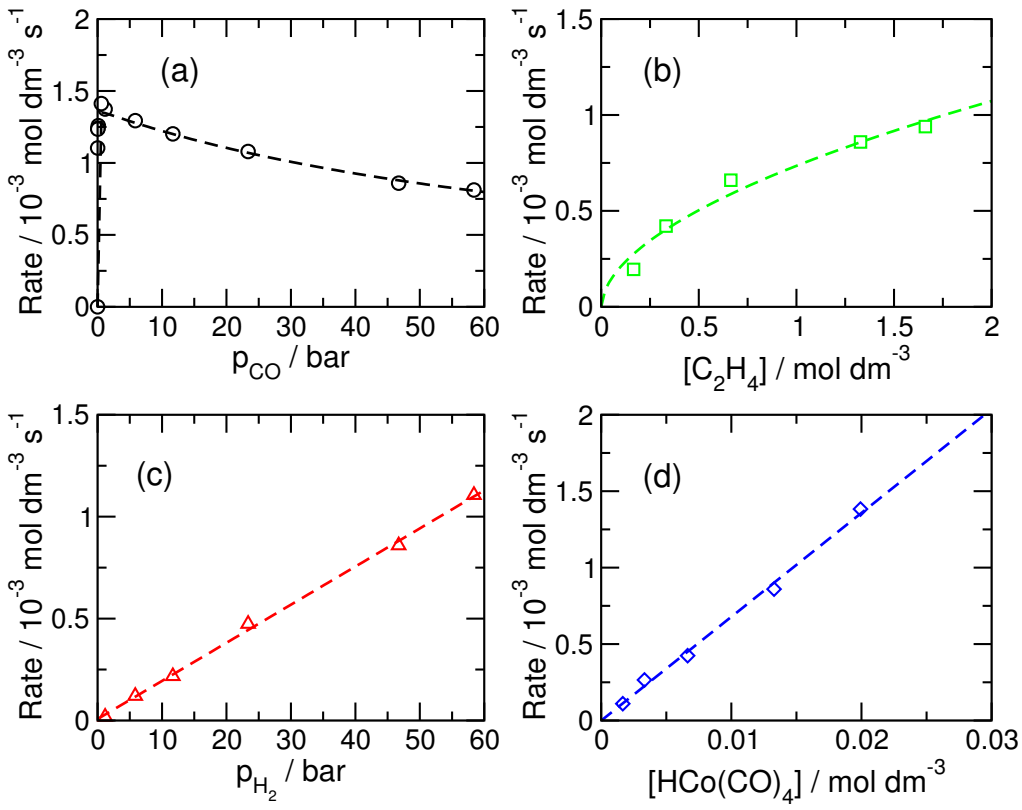

Figure 10: Reaction rates as a function of partial pressures or concentrations of initial reactants. Unless explicitly varied, the initial concentrations (or pressures, for gaseous reactants) of the reactant species S2 (H_2), S3 (CO), S4 (C_2H_4) and S10 ($HCo(CO)_4$) were 46.7 bar, 46.7 bar, 1.33 mol dm^{-3} and 0.013 mol dm^{-3}. In each panel, the dashed lines represent best-fit lines of the functional forms discussed in the main text.

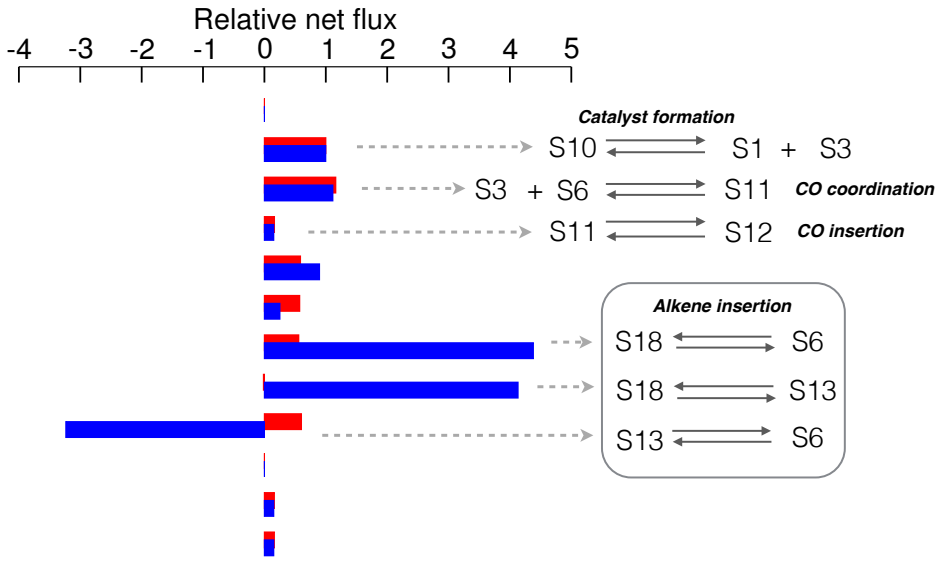

Figure 11: Calculated flux in the modified kinetic network model in the high-pressure (red; $p_{CO} = 46.7$ bar) and low-pressure (blue; $p_{CO} = 0.01$ bar) kinetic networks; in each case, we illustrate the flux relative to that for the reaction S10 $\rightarrow$ S1+S3. In the low p_{CO} regime, it is found that flux through the set of reactions associated with alkene insertion is increased due to low probability of CO coordination *via* S3 + S6 $\rightarrow$ S11.

References

(1) Duca, G. *Homogeneous catalysis with metal complexes*; Springer series in Chemical Physics; Springer-Verlag: New York, 2012.

(2) Bhaduri, S.; Mukesh, D. *Homogeneous catalysis: Mechanisms and industrial applications*; Wiley, 2014.

(3) Parshall, G. W.; Ittel, S. D. *Homogeneous catalysis : the applications and chemistry of catalysis by soluble transition metal complexes*, 2nd ed.; Wiley: New York, 1992.

(4) Houk, K. N.; Cheong, P. H.-Y. *Nature* **2008**, *455*, 309–313.

(5) Tang, Z.; Jiang, F.; Yu, L.-T.; Cui, X.; Gong, L.-Z.; Mi, A.-Q.; Jiang, Y.-Z.; Wu, Y.-D. *J. Am. Chem. Soc.* **2003**, *125*, 5262–5263.

(6) Shinisha, C. B.; Sunoj, R. B. *Org. Biomol. Chem.* **2007**, *5*, 1287–1294.

(7) Raugei, S.; DuBois, D. L.; Rousseau, R.; Chen, S.; Ho, M.-H.; Bullock, R. M.; Dupuis, M. *Acc. Chem. Res.* **2015**, *48*, 248–255.

(8) Sperger, T.; Sanhueza, I. A.; Kalvet, I.; Schoenebeck, F. *Chem. Rev.* **2015**, *115*, 9532–9586.

(9) Heck, R. F.; Breslow, D. S. *J. Am. Chem. Soc.* **1961**, *83*, 4023–4027.

(10) Raghuraj V. Gholap, O. M. K.; Bourne, J. R. *Ind. Eng. Chem. Res.* **1992**, *31*, 1597–1601.

(11) Natta, G.; Ercoli, R.; Castellano, S.; Barbieri, F. H. *J. Am. Chem. Soc.* **1954**, *76*, 4049–4050.

(12) Chaudhari, R. V.; Seayad, A.; Jayasree, S. *Catal. Today* **2001**, *66*, 317–380.

(13) Rush, L. E.; Pringle, P. G.; Harvey, J. N. *Angew. Chemie Int. Ed.* **2014**, *53*, 8672–8676.

(14) Ananikov, V. P., Ed. *Understanding organometallic reaction mechanisms and catalysis: Computational and experimental tools*; WILEY-VCH Verlag: Weinheim, Germany, 2015.

(15) Laidler, K. J. *Chemical Kinetics*, 3rd ed.; Harper Collins: New York, 1987.

(16) Ohno, K.; Maeda, S. *Phys. Scripta* **2008**, *78*, 058122.

(17) Maeda, S.; Ohno, K. *J. Phys. Chem. A* **2005**, *109*, 5742–5753.

(18) Ohno, K.; Maeda, S. *Chem. Phys. Lett.* **2004**, *384*, 277–282.

(19) Maeda, S.; Ohno, K. *J. Phys. Chem. A* **2007**, *111*, 4527–4534.

(20) Kim, Y.; Choi, S.; Kim, W. Y. *J. Chem. Theory Comput.* **2014**, *10*, 2419–2426.

(21) Maeda, S.; Morokuma, K. *J. Chem. Theory Comput.* **2011**, *7*, 2335–2345.

(22) Maeda, S.; Morokuma, K. *J. Chem. Theory Comput.* **2012**, *8*, 380–385.

(23) Habershon, S. *J. Chem. Phys.* **2015**, *143*, 094106.

(24) Zimmerman, P. M. *J. Comput. Chem.* **2013**, *34*, 1385–1392.

(25) Zimmerman, P. M. *J. Chem. Theory Comput.* **2013**, *9*, 3043–3050.

(26) Nett, A. J.; Zhao, W.; Zimmerman, P. M.; Montgomery, J. *J. Am. Chem. Soc.* **2015**, *137*, 7636–9.

(27) Wang, L.-P.; Titov, A.; McGibbon, R.; Liu, F.; Pande, V. S.; Martínez, T. J. *Nature Chem.* **2014**, *6*, 1044–8.

(28) Miller, S. L.; Urey, H. C. *Science* **1959**, *130*, 245–251.

(29) Kegl, T. *RSC Adv.* **2015**, *5*, 4304–4327.

(30) Huo, C.-F.; Li, Y.-W.; Beller, M.; Jiao, H. *Organometallics* **2003**, *22*, 4665–4677.

(31) Mirbach, M. F. *J. Org. Chem.* **1984**, *265*, 205–213.

(32) Orchin, M.; Rupilius, W. *Cat. Rev. - Sci. Eng* **2006**, *6*, 85–131.

(33) Abrams, J. B.; Tuckerman, M. E. *J. Phys. Chem. B* **2008**, *112*, 15742 – 15757.

(34) Abrams, C. F.; Vanden-Eijnden, E. *Proc. Nat. Acad. Sci. USA* **2010**, *107*, 4961–6.

(35) Passerone, D.; Ceccarelli, M.; Parrinello, M. *J. Chem. Phys.* **2003**, *118*, 2025–2032.

(36) Fujisaki, H.; Shiga, M.; Kidera, A. *J. Chem. Phys.* **2010**, *132*, 134101.

(37) Mills, G.; Jónsson, H. *Phys. Rev. Lett.* **1994**, *72*, 1124–1127.

(38) Mills, G.; Jónsson, H.; Schenter, G. K. *Surf. Sci.* **1995**, *324*, 305 – 337.

(39) Allen, M. P.; Tildesley, D. J. *Computer Simulation of Liquids*; Oxford University Press: Oxford, 2005.

(40) Frenkel, D.; Smit, B. *Understanding molecular simulation: From algorithms to applications*; Academic Press: San Diego, USA, 2002.

(41) Tuckerman, M. E. *Statistical Mechanics: Theory and molecular simulation*; Oxford University Press, 2012.

(42) Maeda, S.; Taketsugu, T.; Ohno, K.; Morokuma, K. *J. Am. Chem. Soc.* **2015**, *137*, 3433–3445.

(43) Elstner, M.; Porezag, D.; Jungnickel, G.; Elsner, J.; Haugk, M.; Frauenheim, T.; Suhai, S.; Seifert, G. *Phys. Rev. B* **1998**, *58*, 7260–7268.

(44) Aradi, B.; Hourahine, B.; Frauenheim, T. *J. Phys. Chem. A* **2007**, *111*, 5678–5684.

(45) Zheng, G.; Witek, H. A.; Bobadova-Parvanova, P.; Irle, S.; Musaev, D. G.; Prabhakar, R.; Morokuma, K.; Lundberg, M.; Elstner, M.; Köhler, C.; Frauenheim, T. *J. Chem. Theory Comput.* **2007**, *3*, 1349–1367.

(46) Henkelman, G.; Uberuaga, B. P.; Jónsson, H. *J. Chem. Phys.* **2000**, *113*, 9901.

(47) Frisch, M. J. et al. Gaussian 03, Revision C.02. Gaussian, Inc., Wallingford, CT, 2004.

(48) Gillespie, D. T. *J. Comp. Phys.* **1976**, *22*, 403–434.

(49) Gillespie, D. T. *J. Phys. Chem.* **1977**, *81*, 2340–2361.

(50) Gillespie, D. T. *Annu. Rev. Phys. Chem.* **2007**, *58*, 35–55.

(51) Fichthorn, K. A.; Weinberg, W. H. *J. Chem. Phys.* **1991**, *95*, 1090.

(52) Gillespie, D. T.; Hellander, A.; Petzold, L. R. *J. Chem. Phys.* **2013**, *138*, 170901.

(53) Wilkinson, D. J. *Nature Rev. Genet.* **2009**, *10*, 122–133.

(54) Slepoy, A.; Thompson, A. P.; Plimpton, S. J. *J. Chem. Phys.* **2008**, *128*, 205101.

(55) Stamatakis, M. *J. Phys.: Condens. Matter* **2015**, *27*, 013001.

(56) Cuppen, H. M.; Karssemeijer, L. J.; Lamberts, T. *Chemical Reviews* **2013**, *113*, 8840–8871, PMID: 24187949.

(57) Meng, B.; Weinberg, W. H. *J. Chem. Phys.* **1994**, *100*, 5280.

(58) Ziegler, T.; Cavallo, L.; Berces, A. *Organometallics* **1993**, *12*, 3586–3593.

(59) Tur'anyi, T.; Tomlin, A. S. *Analysis of kinetic reaction mechanisms*; Springer, 2014.

(60) Townsend, D.; Lahankar, S. A.; Lee, S. K.; Chambreau, S. D.; Suits, A. G.; Zhang, X.; Rheinecker, J.; Harding, L. B.; Bowman, J. M. *Science* **2004**, *306*, 1158–1161.

(61) Heazlewood, B. R.; Jordan, M. J. T.; Kable, S. H.; Selby, T. M.; Osborn, D. L.; Shepler, B. C.; Braams, B. J.; Bowman, J. M. *Proc. Nat. Acad. Sci. USA* **2008**, *105*, 12719–12724.

(62) Dey, A.; Fernando, R.; Abeysekera, C.; Homayoon, Z.; Bowman, J. M.; Suits, A. G. *The Journal of Chemical Physics* **2014**, *140*.

(63) Gasner, E. R.; North, S. C. *Software Pract. Exper.* **2000**, *30*, 1203–1233.

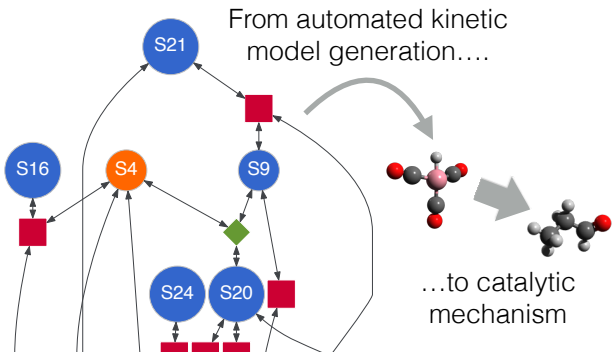

Table of Contents graphic