

# Digital Discovery

Accepted Manuscript

This article can be cited before page numbers have been issued, to do this please use: R. Robidas and C. Y. Legault, *Digital Discovery*, 2025, DOI: 10.1039/D5DD00366K.



This is an Accepted Manuscript, which has been through the Royal Society of Chemistry peer review process and has been accepted for publication.

Accepted Manuscripts are published online shortly after acceptance, before technical editing, formatting and proof reading. Using this free service, authors can make their results available to the community, in citable form, before we publish the edited article. We will replace this Accepted Manuscript with the edited and formatted Advance Article as soon as it is available.

You can find more information about Accepted Manuscripts in the [Information for Authors](#).

Please note that technical editing may introduce minor changes to the text and/or graphics, which may alter content. The journal's standard [Terms & Conditions](#) and the [Ethical guidelines](#) still apply. In no event shall the Royal Society of Chemistry be held responsible for any errors or omissions in this Accepted Manuscript or any consequences arising from the use of any information it contains.

Cite this: DOI: 00.0000/xxxxxxxxxx

Distilling System Complexity to Enable Unbiased and Predictive Computational Reaction Investigations<sup>†</sup>

Raphaël Robidas\* and Claude Y. Legault\*

Received Date  
Accepted Date

DOI: 00.0000/xxxxxxxxxx

Computational modelling is a powerful tool to study chemical reactions. Currently, human guidance is nearly always required to avoid the intractable complexity of all *a priori* possible reaction steps, which consequently greatly limits automated predictive applications. In this work, we present novel theoretical strategies based on the concept of atomic reactivity, as well as a "neophile" kinetic model, demonstrating how they enable unbiased automated reaction modelling with molecules of size typically encountered in experimental methodologies. Our framework allows the identification of unlikely or redundant reaction steps based on first principles and previous analyses, while the neophile kinetic model separates crucial reaction intermediates from inconsequential ones. These advances significantly improved modelling efficiency, allowing us to automatically model 17 unimolecular gold(I)-catalyzed reactions of increasing complexity, starting only from the reactant and catalyst. In 10 reactions, the experimental product distribution is closely reproduced, with 5 additional cases yielding essentially correct reaction networks. Our results demonstrate that it is possible to predictively model catalytic reactions without human guidance through an innovative reformulation of the problem.

Discovering new chemical reactions by computational modelling is a long-standing chemist's dream.<sup>1</sup> The ability to quickly and accurately predict the outcomes of chemical reactions without first empirically testing them in a laboratory would greatly benefit synthetic organic chemistry. Indeed, current limitations in synthetic methodologies can restrain access to organic compounds of interest, such as pharmaceuticals. However, developing new synthetic methodologies is a lengthy, iterative process that demands considerable resources. Computational modelling has the potential to significantly shorten this process by identifying plausible reactions through large-scale computational screenings.

Currently, the state of the art in computational chemistry modelling is quite far from this vision, but studies based on human interaction and intuition are widely successful in the study of known chemical reactions. Computational studies of reaction mechanisms are common in modern chemistry research, as they provide a microscopic view complementary to the experimentally observed outcomes. In addition to providing fundamental insight, the results can be leveraged to reduce side reactions, predict the compatibility of specific reactants, and consider new re-

lated transformations. The use of computational chemistry in this way, however, requires significant work by experts, making it too complex and time-consuming to achieve the ultimate objective of a fully automated *in silico* chemical exploration platform.

As computational chemistry grows more sophisticated and computing power increases, the limits of what can be modelled efficiently and rigorously keep getting pushed back. This means that failures in the theoretical models will keep getting less likely as a reason for discrepancies between theory and experiment. However, this increasing accuracy will not reduce the human time required; if anything, it will increase the complexity of computational studies as more possibilities become tractable and need to be evaluated, rather than being confined to small model systems.

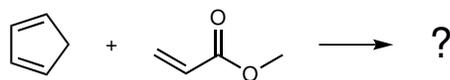
To unlock the full potential of computational modelling, we must embrace a paradigm shift through the development of automated, and fully knowledge- and domain-agnostic (or "unbiased") computational chemistry. Here, "bias" is defined as the modelling scope limitation based on chemical intuition, prior work and, sometimes, human error. This bias is currently *necessary* due to the overwhelming number of reaction processes that are *a priori* possible for any given chemical system. If we view elementary reaction steps as changes in atom bonding, we can enumerate all combinations of bond formation (*F*) and breakage (*B*), as shown in Table 1. This will be named the *combinatorial* approach.

By assuming no prior knowledge about this chemical system, nearly 7 million elementary steps forming and breaking up to two bonds can be enumerated, with only one of them corresponding

Department of Chemistry, Université de Sherbrooke, Centre in Green Chemistry and Catalysis Sherbrooke (Québec) J1K 2R1 (Canada); E-mail: raphael.robidas@usherbrooke.ca, claude.legault@usherbrooke.ca.

<sup>†</sup> Supplementary Information available: all final reaction networks, including molecular structures, Gibbs free energy and original labels. They can be explored interactively using the provided static web pages, which were also used to produce the network illustrations of this paper. See DOI: 00.0000/00000000.



Table 1 *A priori* possible reaction possibilities in a small system (23 atoms)

Bonds formed and broken	Number of possibilities
$B^1$	22
$F^1$	231
$B^2$	231
$F^1B^1$	5082
$F^2$	26,565
$F^1B^2$	53,361
$F^2B^1$	584,430
$F^2B^2$	6,136,515
<b>Total</b>	<b>6,806,437</b>

to the experimentally observed Diels-Alder ( $F^2$ ) reaction.<sup>2,3</sup>

Accordingly, while automatic and unbiased computational mechanism exploration has been an active field of research for at least 25 years,<sup>4</sup> the applicability of such approaches has remained extremely limited. This automated exploration has also been formulated as a high-dimensional coordinate space exploration, although this does not limit the exponential growth of the system complexity with its size. Regardless of the approach, most developments until now involve almost exclusively small ( $\leq 25$  atoms) molecules,<sup>5–15</sup>

To tackle larger systems, some combinatorial approaches use wisely chosen rules to consider only some atoms as reactive<sup>16</sup> or require manual selection of reactive atoms<sup>9,17–19</sup>, which sacrifices complete generality and automaticity. Reaction templates have also been used to reduce the combinatorial explosion, especially in the context of a limited diversity of possible mechanistic steps, like pyrolysis<sup>20,21</sup> and polymerisation.<sup>22</sup> Hammond's postulate has been used to eliminate some reaction possibilities once the product's thermochemistry is known.<sup>23</sup> Additionally, manually curated mechanistic steps have been leveraged recently through an algorithmic approach in order to predict cationic rearrangement outcomes.<sup>24</sup> Although encouraging, this approach cannot be easily generalised without substantial involvement from human experts. Different reactions would necessitate curating the relevant mechanistic steps from the ground up.

We believe that this shift towards fully automated, general, and unbiased computational chemistry is crucial for achieving remarkable progress in the field. In this study, we propose novel strategies to tame the exponential growth of the combinatorial approach. We then provide a proof-of-concept based on the automated investigation of catalytic organometallic reactions, which yields both insightful mechanistic details and reliable predictions of reaction outcomes. These results hold significant value as they can reveal outcomes that deviate from previously established patterns, which are often difficult to predict. This difficulty is intensified by the noticeable bias in the scientific literature towards successful reactions, with failed reactions seldom reported, and the

common absence of identified and quantified side products. An unbiased reaction model offers valuable insights into the reactivity of the reactant in question, including unanticipated outcomes.

We have developed a conceptual framework based on the ansatz of "atomic reactivities" to determine plausible reactions and control the system-wise combinatorial explosion of possibilities without any human interaction. We have also developed the "neophile" kinetic model, which limits the network-wise possibility growth. In this work, we present how these innovative approaches enable the automated exploration of typical unimolecular gold(I)-catalyzed reactions with *predictive* accuracy. The studied reactions involve reactants with 22 to 42 (reactive) atoms as well as catalysts with 35 to 66 atoms, constituting system sizes significantly larger than previously possible. Historically, acquiring detailed mechanistic insights has required significant effort, and accurately predicting outcomes has often been elusive. Herein, we report a method that successfully achieves both.

## 1 Results and Discussion

### 1.1 Atomic reactivity framework

#### 1.1.1 Reaction possibilities

Our general approach to automating reaction explorations is combinatorial, as elementary reactions are defined by combinations of bond formation and breakage, similar to previously reported approaches.<sup>9,12,14,25,26</sup> To eliminate nonsensical elementary reaction possibilities (hereafter "possibilities"), many of these approaches use a valence-based check and also remove duplicates by symmetry. Double and triple bonds are also expected never to be broken in a single elementary step (*i.e.*, pulling both atoms apart). These criteria have been taken as a baseline to evaluate the reduction in plausible reaction possibilities.

Our approach uses bond formations and breakage as driving coordinates to find new elementary reactions. These coordinates are only used to bias the exploration of the potential energy surface through a relaxed scan; they do not restrict the reactions to be discovered, as opposed to methods that generate the product from the expected connectivity.

Thus, reactions that do not involve exactly the expected bond changes are also discovered. This enables the reduction of the number of explicitly considered reaction coordinates. While a common limit is the formation and breakage of up to two bonds each (denoted  $F^{0 \rightarrow 2}B^{0 \rightarrow 2}$ ), we found that all relevant reactions were found with only four combinations:  $F^1$ ,  $B^1$ ,  $F^1B^1$  and  $F^2$ . As shown in Fig. 1, the number of possibilities is reduced by two orders of magnitude by this modification. In addition, possibilities of type  $F^1$  involving the catalyst and every other atom of the reactant were added in order to move the catalyst around and ensure that all complexation isomers are found. These possibilities can also lead to productive reactions due to implicit PES sampling.

#### 1.1.2 Atomic reactivity ansatz

As our initial foray into developing innovative strategies to prevent combinatorial explosion in larger molecular systems, we decided to focus on catalytic systems. In particular, we turned our focus on gold(I)-catalyzed process because of its general interest



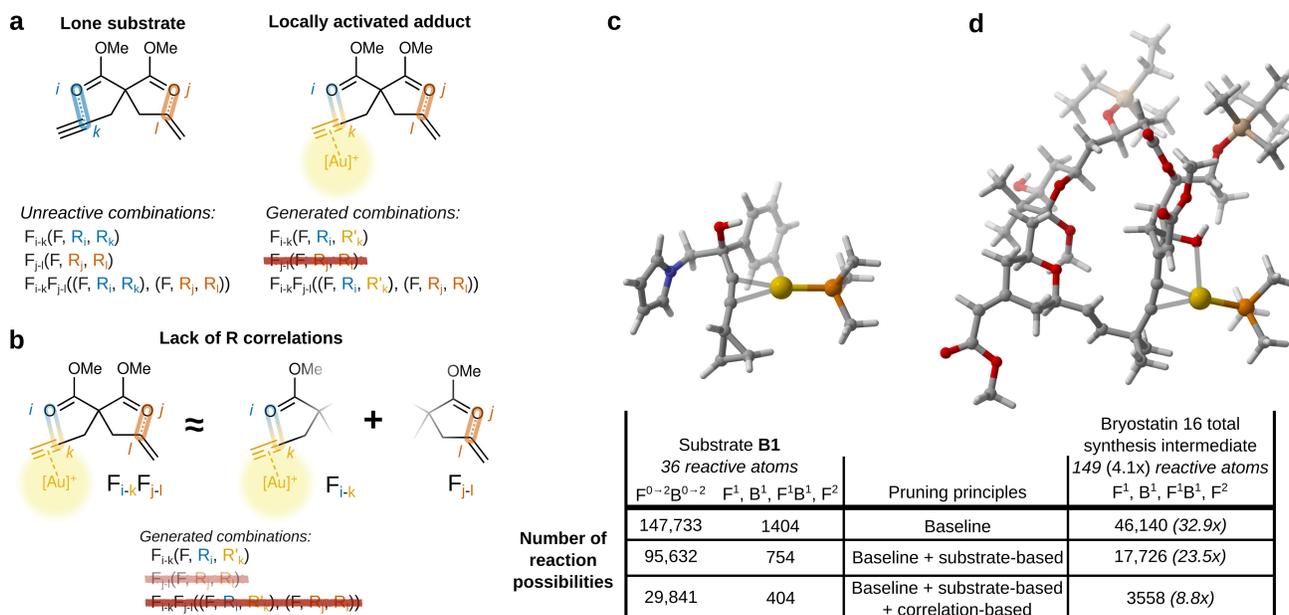


Fig. 1 Example of reaction possibilities pruning through principles derived from atomic reactivity. a) Pruning principle based on the lone reactant reactivity; b) Pruning principle based on atomic reactivity correlation; c) Reactant B1, whose reactions are presented in subsection 1.3.2; d) Intermediate in the total synthesis of bryostatin 16 by Trost.<sup>27</sup> Multiples in parentheses are calculated with respect to reactant B1.

in modern organic synthesis and the availability of high-quality results with quantified side products.

The foundation of our framework comes from what we know to be unreactive. In the case of gold(I)-catalyzed reactions, the reactant itself without a catalyst is stable and does not react by itself. This means that any imaginable reaction involving only the reactant has to be negligible. The role of the catalyst is to activate certain parts of the molecule to enable a reaction. The first step of the reaction then has to involve these activated parts of the molecule, since the rest has the same non reactivity as in the lone substrate.

In order to leverage those logical deductions, we propose the atomic reactivity ansatz, which is defined as such:

1. The molecular reactivity can be broken down into atomic reactivities, which are properties of the atoms.<sup>28</sup>
2. As a result, changing the reactivity of a given molecule inevitably implies alterations in the atomic properties of part of its atoms.
3. The combined atomic reactivities of the atoms involved in a reaction determine the barrier of the reaction.

In terms of atomic reactivities, the deductions above mean that any combination of atomic reactivities in the reactant is associated with a large barrier. In terms of notation, the atomic reactivity of atom  $i$  will be written as  $R_i$ . The formation of a bond between atom 1 and atom 2 will be denoted  $F_{1-2}$ . If this bond formation occurs simultaneously with the breakage of a bond between atom 2 and atom 3, this overall reaction will be denoted  $F_{1-2}B_{2-3}$ .

Since our approach is combinatorial, these reactions can also be described by pairs of atomic reactivities, each pair correspond-

ing to the reactivities of the atoms whose bond is being formed or broken. This formulation allows for the comparison between potential elementary reactions and the negligible reactions of the lone reactant. Thus, elementary reactions that are functionally equivalent to those of the lone reactant will be disregarded without any calculation.

In this work, we used the pair  $(q, \alpha)$  as the formulation of  $R$ , where  $q$  is the atomic partial charge and  $\alpha$  is the atomic polarisability. This formulation is derived in the subsection 2.3.1 of the ESI.<sup>†</sup>

### 1.1.3 R correlations

Possibilities that involve multiple reaction coordinates can be *de facto* redundant. For example, if two bonds to be formed or broken are on opposite edges of the molecule and the electronic conjugation is negligible between the two reaction sites, we can eliminate this possibility: each reaction site is isolated from the other and will produce the same result as if the other reaction did not occur as well.

More formally, we define *reactivity correlation* as  $\partial R_i / \partial R_j$ . If  $\partial R_i / \partial R_j \approx 0$ , then atoms  $i$  and  $j$  are said to be *uncorrelated*. This means that changes in the properties of atom  $i$  (because of a bonding change or other fluctuation in electronic density) will not influence how reactive atom  $j$  is. For the possibility  $F_{i-k}F_{j-l}$ , if  $i$  and  $k$  are both uncorrelated to  $j$  and  $l$ , it is reasonable to assume that:

$$\Delta G_{F_{i-k}F_{j-l}}^\ddagger \approx \Delta G_{F_{i-k}}^\ddagger + \Delta G_{F_{j-l}}^\ddagger \quad (1)$$

While  $F_{i-k}$  and  $F_{j-l}$  are relevant to evaluate individually,  $F_{i-k}F_{j-l}$  is not. As a first approximation, we considered that only directly bonded atoms have correlated reactivities (see section



2.3.1 of the ESI† for details about the implementation).

### 1.1.4 Data-driven Pruning

As the reaction network is being explored, elementary reactions will be evaluated through quantum chemical calculations (see section 2.1 of the ESI† for the full methodology). This process will yield relative Gibbs energies (hereafter shortened to "energies") for the discovered intermediates and the transition states leading to them. It becomes possible to associate combinations of atomic reactivities with the computed reaction barriers throughout the different refinement stages (for example:  $((F, R_i, R_j)) \Rightarrow \Delta G_{F_{R_i-R_j}}^\ddagger$ ). These results can be used to filter out negligible reactions without any additional calculation based on the equivalence of atomic reactivities. As the reaction network exploration progresses, these reference barriers accumulate and allow increasing pruning of reaction possibilities.

### 1.1.5 Effect on reaction possibilities

When these novel principles, derived from atomic reactivity, correlation, and data pruning are also applied, the number of reaction possibilities further decreases significantly. For reactant **B1** (Fig. 1a), the number of possibilities is further reduced by a factor of 3.5 by the two principles. Moreover, the reduction in possibilities becomes more significant for larger reactants. For example, the number of possibilities for the intermediate in the total synthesis of bryostatin 16 (Fig. 1a) is 46,140 with standard pruning rules, yet decreases to only 3558 with the addition of atomic reactivity principles. Furthermore, this intermediate presents only 8.8 times more feasible possibilities despite containing 4.1 times more reactive atoms than reactant **B1**, indicating less than quadratic growth. Although larger reactants may have more relevant complexation isomers, the number of possibilities is still manageable.

Fig. 2a plots the reduction in the number of evaluated reaction possibilities for all the 17 considered reactions during the reaction network explorations. The number of possibilities that would have been considered without atomic reactivity pruning principles has been recalculated to quantify the reduction. Using the reactant's atomic reactivities reduces the possibilities by a factor of 1.5, on average. Considering atomic reactivity correlations and past data greatly increases the reduction factor, especially for large numbers of total possibilities. Explorations that would require more than 60,000 possibility evaluations are reduced by a factor of 4 to 5 with both additional principles. This means that networks normally involving 140,000 reasonable reaction possibilities (with baseline principles) will require only 28,000 to 35,000 possibilities to be evaluated. Fig. 2b shows that the percentage of pruned possibilities due to past data increases steadily with each possibility evaluation cycle, often reaching 60% beyond 14 cycles in the one case that reached that stage. The combinatorial explosion of possibilities is thus further contained by exploiting its limited novelty in terms of chemical reactivity.

## 1.2 Neophile Kinetic Simulations

Another key factor leading to the combinatorial explosion is the formation of the reaction network. Indeed, the elementary reactions starting from each new product also need to be considered;

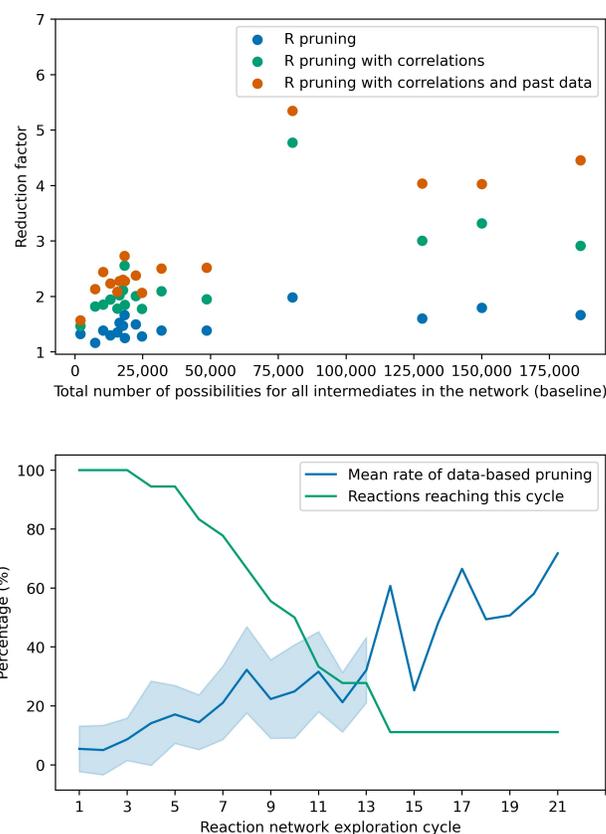


Fig. 2 Global reduction in the number of evaluated reaction possibilities. a) Reduction factor for entire network explorations depending on pruning principles and total possibilities; b) Pruning rate statistics across network exploration cycles, with the shaded area spans one standard deviation from the mean.

in nearly all systems, this number of elementary reactions will increase exponentially with the number of steps considered, since each product leads to multiple other products. Not all reactions will be significant in the final reaction network, but it is challenging to identify significant reactions in the incomplete reaction network. Reported approaches include considering all reactions with a barrier below a defined threshold<sup>29</sup> and considering only the fastest reaction per intermediate.<sup>9</sup> A more detailed review of the problem and suggested solutions has recently been published.<sup>30</sup>

Alternatively, the changes in concentration of all intermediates can be modelled using a system of differential equations. The reaction rate for intermediate  $i$  is given by:

$$dC_i = \sum_k C_k K_{k \rightarrow i} - C_i \sum_k K_{i \rightarrow k} \quad (2)$$

where  $C_k$  represents the concentration of intermediate  $k$  at a given moment and  $K_{i \rightarrow k}$  is the rate constant of the reaction from intermediate  $i$  to intermediate  $k$ . Rate constants are estimated using calculated energy barriers and the Eyring model. Numerical integration of this system of differential equations yields the concentration as a function of time.

With a full and exact reaction network, a kinetic simulation will determine which products accumulate and in which proportions.



However, when the network is not quite complete, the kinetic simulation can be misleading. With a "short" simulation time, the most significant intermediates will be those who are kinetically favoured, while the true reaction pathway might involve a higher barrier that leads to a much more stable intermediate. In this case, the latter would be considered negligible, as it did not form sufficiently during the simulation time. With a "long" simulation time, the system tends towards the thermodynamic equilibrium, and the final distribution of products will depend only on their thermodynamic stabilities.

One approach would be to aim to be in the intermediate regime where both kinetic and thermodynamic effects are balanced. However, this can be problematic if the next intermediate in the reaction pathway is relatively high in energy and thus will not accumulate meaningfully. Additionally, determining the appropriate simulation time is challenging and depends on the overall reaction barrier. Any systematic errors in the reaction barrier will also affect the optimal simulation time. Consequently, we opted to abandon this approach and developed an alternative kinetic model instead, coined neophile kinetic simulation.

The key challenge to address is the incompleteness of the network. The edges of the nascent network induce errors, since further reactions are missing. The matter that reaches the edge intermediates is kept within the limited network, although they might undergo a very rapid and favourable reaction in the fully explored network. Neophile kinetic simulations thus use one simple rule: *intermediates whose reactions have not been fully evaluated are formed irreversibly*. Thus, these intermediates will accumulate during the simulation proportionally to their ease of formation, and their final concentrations can be used to identify significant intermediates. This approach is extremely robust, and we were unable to find any scenario where it induced errors. By design, it is suitable for reaction networks of any energy span.

A related kinetic strategy has been independently developed by the Savoie group this year.<sup>31</sup> In their case, a graphical distance to the nearest unexplored intermediate is defined and used to further modify the reaction kinetics in combination with an exploration hyperparameter.

Quantifying the computational savings from neophile kinetic simulations is challenging, as conducting complete network explorations without them would be intractable. Nevertheless, the substantial number of energetically favourable intermediates identified, yet not prioritised in the neophile simulations, suggests that the potential reaction pathways are diminished by at least an order of magnitude compared to a simplistic threshold-based approach.

### 1.3 Reaction Results

To assess the viability and scope of automated reaction studies using these new concepts, we selected 17 known gold(I)-catalyzed reactions that had a single reactant as starting material, with the goal of reproducing product distributions and automatically elucidating their formation mechanism. We categorised them into three classes based on their anticipated mechanistic complexity. We chose these reactions to assess the accuracy of the predic-

tions across a wide range of transformations. Reactions producing more than one identified and quantified product were particularly significant to verify how accurately the product distribution can be reproduced. All reactions resulted in good to quantitative yields of identified products and were reported in reputable journals.

The A-class category is composed of 1,6-enyne and 1,5-enyne isomerisations, well-known gold(I)-catalyzed reactions which form multiple new carbon-carbon bonds. The expected reaction mechanisms are likely to involve relatively few steps (2-4) and are not anticipated to require intermolecular proton transfers. The reactants bear only one or two functional groups other than the enyne motif.

The B-class category displays more complexity and chemical transformation diversity. The reactants generally contain multiple reactive functional groups. One or multiple proton transfers, and dehydration, are expected to be important in the mechanisms.

Finally, the C-class category may involve considerable skeletal changes and larger reactants. Sizable alkyl groups were modelled as-is without simplification and many reactive functional groups may be present. Proton transfers and unfavoured tautomers are also expected to be important in multiple mechanisms. Multiple reactants bear relative or absolute stereochemistry, although in this first proof-of-concept we did not address this explicitly.

In every case, starting from only the reactant and the experimental ligand, the entire relevant reaction network was explored until no further reaction was predicted to occur significantly. Products with predicted yields below 1% are not shown. All quantum chemistry calculations with more than one electron are imperfect, and shortcomings in the theory levels used may affect the results. As such, a perfect match between experimental and calculated results is exceedingly rare. In some cases, errors linked to the theory level could explain incorrect predictions, and such cases will be discussed.

#### 1.3.1 A-class reactions

The cycloisomerisation reaction of enyne **A1-1** to **A1-2** discovered as the main reaction pathway, leading to a predicted yield of 99%. This result is in excellent agreement with the very high experimental yield of 97%.<sup>32</sup>

This unbiased exploration provides an unprecedented level of insight into the mechanistic study of complex systems (Fig. 3). Indeed, our approach now allows for the investigation of both productive and unproductive pathways, enabling access to the true energetic span of complex catalytic processes. This will be crucial for addressing, for example, the structure-activity relationship in ligand design.

In this graph, the lone reactant **A1-1** is fixed to have an energy of 0 kJ/mol. Multiple complexation sites are available for the catalyst (alkene, alkyne, esters), leading to multiple complexation pathways to intermediates **A1-i1** to **A1-i3**. The most stable complexation isomer **A1-i2** is also defined to have an energy of 0 kJ/mol since the complexation energy of the catalyst is not relevant in the case of pseudo-unimolecular reactions. All other relative energies are calculated with respect to this intermediate. The complexation barriers are considered to be negligible. All



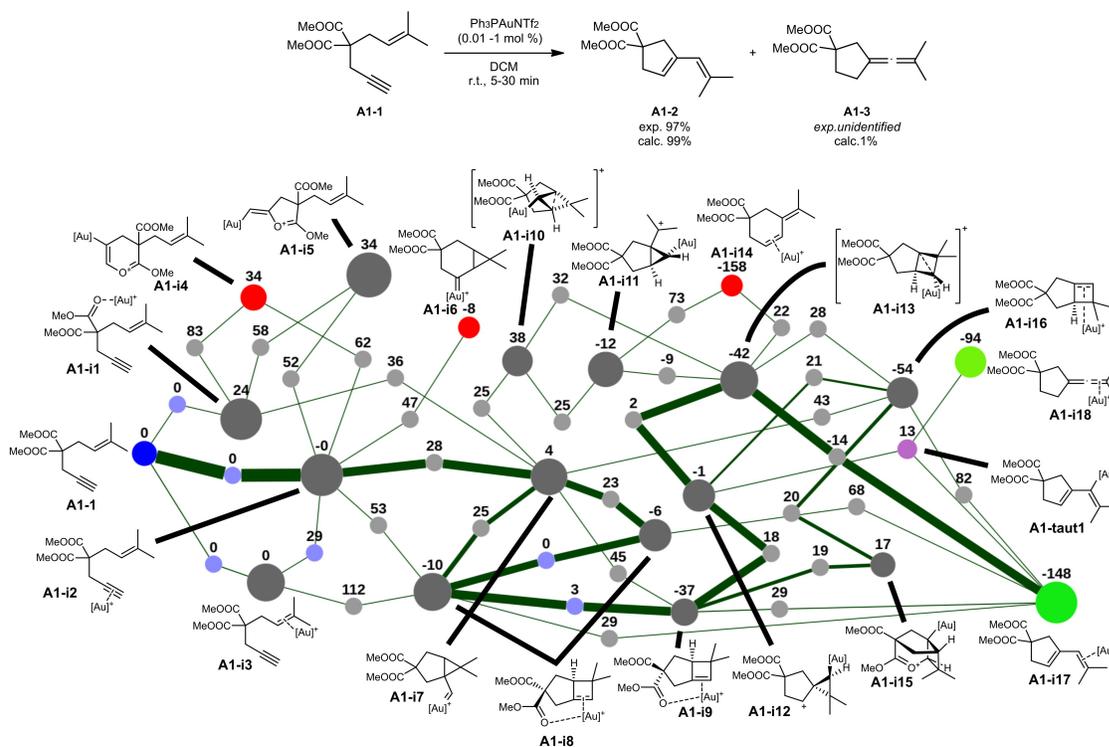


Fig. 3 Transformation of reactant A1-1<sup>32</sup> and selected portion of its reaction network (relative Gibbs free energies in kJ/mol). Some complexation isomers and negligible reaction pathways have been removed to reduce clutter (view the full networks in the ESI†). The blue node corresponds to the lone reactant, light blue nodes indicate complexation isomerisation dummy transition states, light purple nodes indicate tautomerization pseudo-transition states, red nodes indicate intermediates that were formed in negligible amounts and not explored further, and light green nodes indicate the most significant intermediates at the end of the kinetic simulation. The thickness of the edges represents the final net flux that has passed through those edges.

nodes between the identified intermediates correspond to transition states, their structure not being illustrated in the figure but available in the ESI† through the full networks.

Starting from the complexation isomers, cyclisations on the esters are found to be possible (A1-i4 and A1-i5), but endergonic and not ultimately significant. The expected 6-endo-dig and 5-exo-dig cyclopropanations are also found, leading to A1-i6 and A1-i7 respectively, with the latter possessing a much lower formation barrier. This favoured product rearranges mainly to the bicycle A1-i8, which has other complexation isomers (A1-i9, A1-i16). Those intermediates can notably rearrange to carbocation A1-i12, which can further rearrange to the non-classical carbocation A1-i13. This intermediate directly leads to the final product bound to the catalyst (A1-i17). No further reaction is predicted to be significant and this product accumulates. In addition to this product, trace amounts of allene A1-i18 are formed through the elimination and reprotonation of intermediate A1-i12.

This unbiased exploration of the mechanistic manifolds of this transformation provides an unprecedented level of insight on other reasonable reactions, albeit negligible in this case. The starting adduct A1-i2 can directly form bicycle A1-i8 with a barrier of 53 kJ/mol, although the 5-exo cyclisation followed by a rearrangement (A1-i2→A1-i7→A1-i8) has a much lower rate-determining step of 28 kJ/mol. Similarly, bicycles A1-i8 and A1-i9

can directly form the final product A1-i17, although those reactions are not favoured. The formation of intermediate A1-i15 was also investigated, although it ended up being unproductive and inconsequential. Intermediate A1-i14, which has been identified but not significantly formed, is analogous to the main product A2-3 of the second reaction (see Fig. 4). The reaction network thus shows that the selectivity between those two possible products is entirely determined by the reaction barriers of intermediate A1-i13.

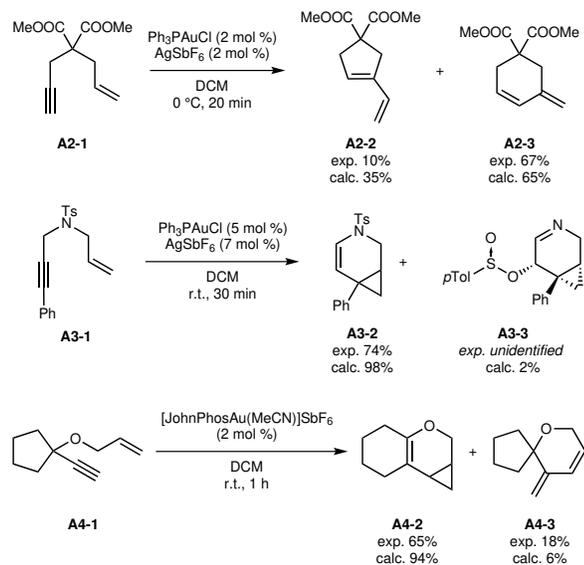
This subgraph thus shows a glimpse of the profound mechanistic understanding that is inherently accessible from the generated data results.

Fig. 4 shows other 1,6-enyne isomerisation reactions that were explored in the same way. Gratifyingly, the predicted product distributions closely follow the experimental ones in every case. These results indicate that the influence of functional group changes has been accurately modelled.

The reactions of 1,5-enynes A5-1<sup>36</sup> and A6-1<sup>37</sup> were explored with similar success (see Fig. 5). The reaction from A5-1 to A5-2 was also found to be the only significant reaction, in agreement with the high experimental yield of 99%.

The reaction of A6-1 is slightly less straightforward. Experimentally, the aldehyde direct product A6-2 was not isolable without significant degradation, so the authors carried out a reduction



Fig. 4 Other modelled 1,6-enyne isomerisation reactions.<sup>33–35</sup>

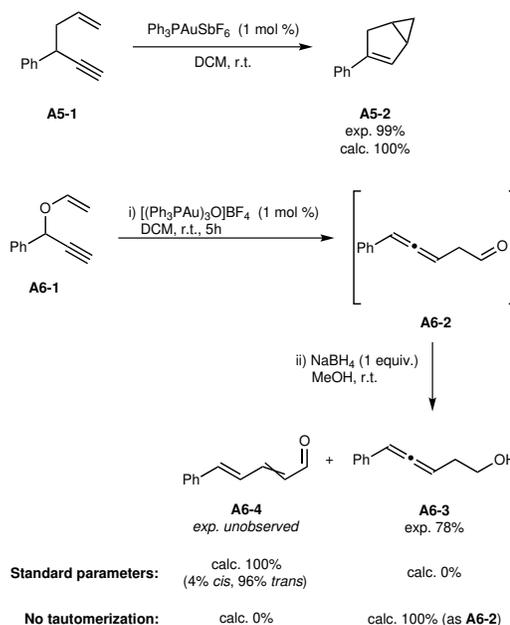
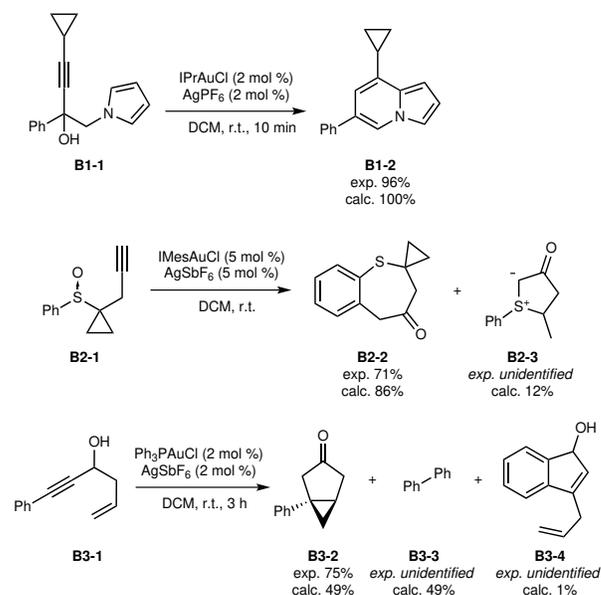
step to isolate the alcohol **A6-3**. Of course, the results do not take into account the reduction step, so the expected product is the aldehyde **A6-2**. This product is indeed formed, but quickly undergoes proton transfers to form the thermodynamically favored products **A6-4** as a mixture of *cis/trans* isomers. The proton transfers are estimated to have a low barrier of 32 kJ/mol, making them much faster than the formation of the aldehyde itself (barrier of 68 kJ/mol). The intermediate immediately preceding the aldehyde can also undergo a proton transfer to ultimately yield **A6-4**. As explained in subsection 3.4, proton transfers are not modeled completely explicitly, in part due to the uncertain and variable species involved. In particular, the approximate proton transfer barriers are calculated with the reactant as proton transfer agent. In dilute solutions, this intermolecular process will be significantly slower, which could explain the apparent discrepancy. Indeed, the effect of substrate concentration during the reaction was not taken into account at this time.

If tautomerizations are considered to be too slow and thus effectively impossible, the automated exploration does predict the sole formation and accumulation of aldehyde **A6-2**. It is thus unclear whether the calculated Gibbs free energies are affected by an error or if the overall kinetic modelling requires additional sophistication and experimental parameters.

### 1.3.2 B-class reactions

The results of the first three reactions are shown in Fig. 6. The formation of indolizine **B1-2** was correctly found, including the elimination of a water molecule, and without predicting significant reaction with the cyclopropyl group.<sup>38</sup> The experimental yield of 96% is coherent with a clean transformation without side reaction, as found *in silico*.

The conversion of sulfoxide **B2-1** was predicted to yield mostly the expected product **B2-2**<sup>39</sup> with a small amount of ylide **B2-3**. Although not experimentally identified and quantified, it is possible that ylide **B2-3** is indeed formed and could explain part

Fig. 5 Results of modelled 1,5-enyne isomerisation reactions.<sup>36,37</sup>Fig. 6 First set of B-class reactions.<sup>38–40</sup>

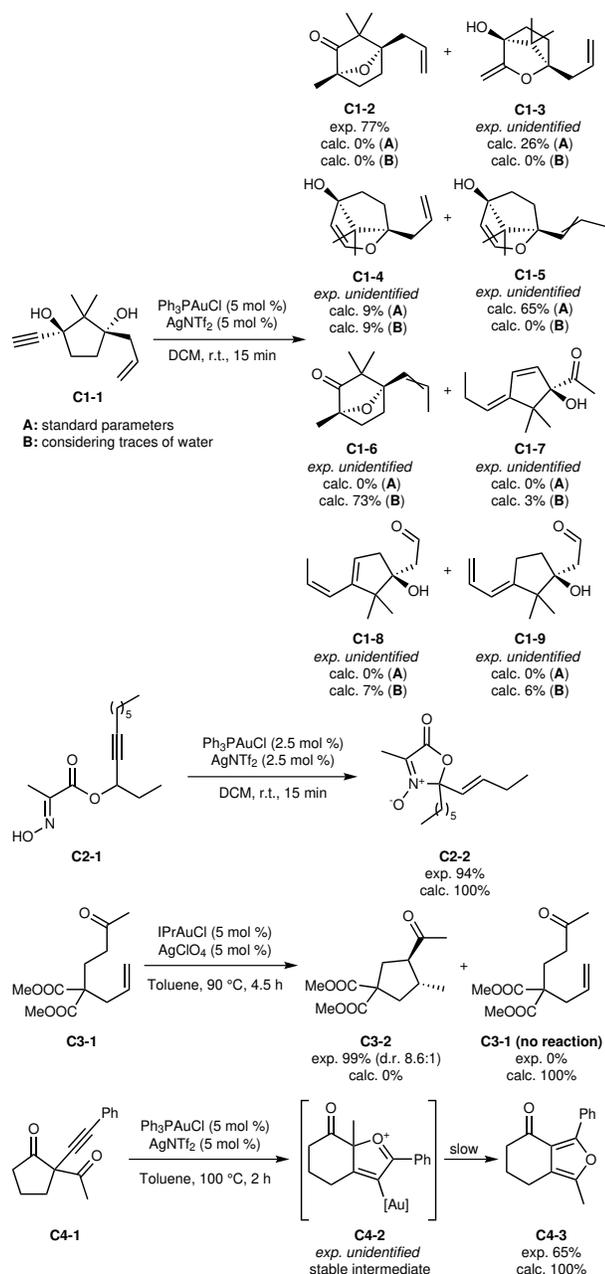
of the lower experimental yield, as **B2-3** could even react and degrade with **B2-2**.

The transformation from **B3-1** to **B3-2** was automatically identified,<sup>40</sup> in addition to the formation of biphenyl **B3-3**. Although the predicted yield is somewhat too low, the lost mass balance might be explained in parts by the formation of the biphenyl **B3-3**, which is very apolar and could go unnoticed during purification. Additional insight is obtained from the reaction network (Fig. 7).

The major product **B4-2** of the transformation of **B4-1**<sup>41</sup> has been correctly predicted to be formed, albeit as minor product. The experimentally observed secondary product **B4-3** was not





Fig. 9 First set of C-class reactions.<sup>43–46</sup>

poorly reproduced when using only continuum solvation models, while small water clusters yield results much closer to experiment. Based on this hypothesis, we calculated the energy associated with the formation of hydronium and hydroxide ions microsolvated by four water molecules.

Using those values to calibrate the tautomerization barriers, the exploration of the reaction was carried out again. With these parameters, other proton transfers become possible. In particular, alkene isomerisations become more accessible, leading to a mixture of terminal and internal alkenes on the side chain. In addition, intermediates with the same core as **C1-3** can lose a proton to form a gold alkoxide, which can then be reprotonated at the exocyclic enol position. This process forms an activated alkoxide-

carboxonium species which readily undergoes the semi-pinacol rearrangement and leads to **C1-6** as major product. This product is correct except for the alkene position, which suggests that alkene isomerisation barriers are underestimated. The formation of the identified minor side products (**C1-4**, **C1-7**, **C1-8** and **C1-9**) is a possible reason for the lost mass balance, although it is not possible to confirm this hypothesis from the available data.

The reaction of **C2-1**<sup>44</sup> yields a zwitterion bearing a quaternary centre, which could plausibly be thought to react further to a more thermodynamic product. However, the automated exploration correctly predicts the sole and complete formation of the main product, in agreement with the experimental yield of 94%. It is worth specifying that the hexyl chain of the reactant has not been simplified in the modelled reactant, as is common practice. The modelling approach keeps computational costs sufficiently low that such a chain can be modelled as-is, thus preventing discrepancies caused by (unknowingly) altering the reactivity with simplifications. Moreover, this result suggests that our conformational sampling process does not become an issue with larger, more flexible reactants.

Reactant **C3-1** can cyclize nearly quantitatively under gold(I) catalysis into product **C3-2**, presumably through an enol intermediate.<sup>45</sup> The automated exploration did not find any reasonable reaction, predicting that the reactants remain unreacted. We once again considered the influence of traces of water on the reaction. With this modification, tautomerizations become accessible, leading to the expected enol intermediates. Moreover, the automated exploration led to the discovery of 5 possible cyclisation transition states, which vary in terms of internal hydrogen bonding and of alkene position. All transition states were associated with significant barriers (112 kJ/mol to 147 kJ/mol), and thus did not lead to any significant reaction.

It is possible that a second molecule of reactant could be involved in the reaction. The catalyst's counterion,  $\text{ClO}_4^-$ , may also play a role in the transformation. In the original work, the authors screened several counter-ions and found that  $\text{BF}_4^-$  lead to no conversion,  $\text{SbF}_6^-$  lead to a yield of 19%, while  $\text{PF}_6^-$ ,  $\text{TfO}^-$  and  $\text{ClO}_4^-$  lead to nearly quantitative yields (96–99%), with  $\text{ClO}_4^-$  reducing the required reaction time greatly. Although not entirely

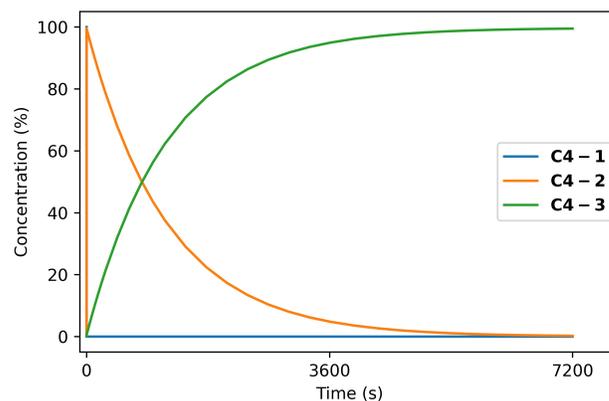


Fig. 10 Result of the kinetic simulation of the C4 reaction network.



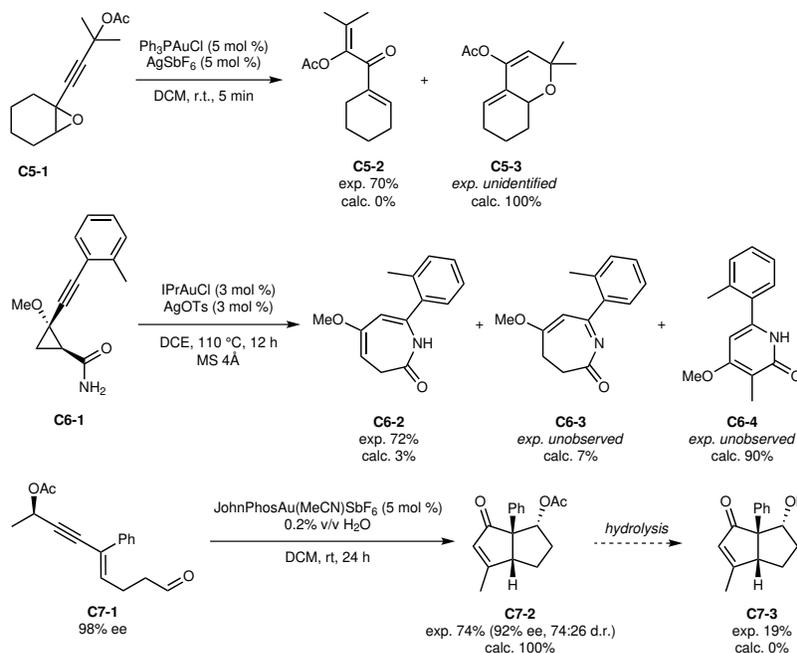


Fig. 11 Second set of C-class reactions. 48–50

conclusive, these results indicate that the counterion affects the reaction, which was not in the scope of our current implementation. A dimeric catalyst could also be involved in the reaction, which is not unheard of in gold(I) catalysis.<sup>51</sup> In any case, these factors may complexify the exploration process beyond what is handled by our current implementation. Nonetheless, the robustness observed in A- and B-class methodologies suggests that our platform could be used to uncover uncommon mechanistic behaviours in developing methodologies, in which case additional experimental studies would be required to address the true nature of the active catalytic species or trace species involved in the overall process.

Under gold(I) catalysis, reactant **C4-1** selectively cyclizes on the exocyclic ketone, resulting in the formation of product **C4-3**.<sup>46</sup> Our automated exploration of the reaction network indicates that intermediate **C4-2** is formed faster than **C4-3** and initially accumulates. However, because of the greater thermodynamic stability of **C4-3**, interconversion occurs at a slower rate, and may account for the increased temperature requirement (see Fig. 10). With this additional mechanistic insight regarding the lifetime of intermediate **C4-2**, one could envision exploiting its formation and reactivity to develop novel methodologies to access new compounds bearing multiple quaternary centres. While outside the scope of the current implementation of the platform, bimolecular automated reaction explorations could be used to identify valuable transformations from **C4-2**. Furthermore, the lost mass balance may stem from side reactions between **C4-2** and either **C4-1** or **C4-3**, which were not examined computationally. The mechanism automatically found also differs from the authors' proposed mechanism,<sup>46</sup> which omits **C4-2**, preventing further exploration of this intermediate. This exemplifies the insights provided by such unbiased automated studies.

The reaction of **C5-1** experimentally yields product **C5-2**,<sup>48</sup> while only product **C5-3** was found significant *in silico*. Considering the length of the pathways leading to this product (at least 7 elementary steps), it is difficult to pinpoint the error source. Despite the accuracy of current computational methods, they remain subject to small intrinsic errors, which can shift the course of reactions. While such errors can be identified in mechanistic studies with known products, predictive applications require highly accurate results at each step. The outcome of reaction **C5** exemplifies the challenge of outcome prediction with errors compounding over 7+ reaction steps.

The reasonable reaction pathways of **C6-1** into **C6-2** and its isomer **C6-3** were found,<sup>49</sup> but the formation of **C6-4** was found to be more favourable. In this reaction, the selectivity between the two kinds of products comes from a competition between a proton transfer and the cyclopropyl ring opening. As proton transfers are modelled semi-explicitly (see Methods), their barriers are only approximate. The error linked to this approximation influences a key step and can change the reaction outcome. In these conditions, however, it is not obvious what the true proton-transfer agent would be, and so fully explicit modelling is quite challenging. Even with this approximation, the actual product is expected to be significant, representing approximately 10% of the predicted products.

Finally, reactant **C7-1** is highly enantioenriched and can cyclize under gold(I) catalysis to form enantioenriched products **C7-2** and **C7-3**.<sup>50</sup> The experimental ratio between the two products relies on specific conditions. The authors of this study found that the reaction benefitted from adding water, likely because it facilitated the hydrolysis of a reaction intermediate. The modelled conditions did not involve explicit water molecules, and so this hydrolysis was naturally not accounted for. However, product



**C7-3** is still found as the sole product of the reaction. As mentioned before, we elected not to strictly enforce stereochemistry considerations in our current implementation. Nonetheless, the predicted product bears the correct absolute stereochemistry at all three chiral centres. This suggests that the rigorous consideration of stereochemistry will require only modest changes to our approach.

The shortest reaction pathway from **C7-1** to **C7-2** involves 6 elementary steps, confirming that the accuracy of computed energies is generally excellent and can also be applied to complex transformations with many elementary steps.

## 2 Conclusion

In this work, we have demonstrated that it is possible to address the combinatorial explosion of possibilities in unbiased mechanistic explorations through the use of innovative pruning strategies. The use of an atomic reactivity framework and neophile kinetic simulations reduces the number of reaction possibilities sufficiently to render completely tractable automated investigations of organometallic reactions involving organic reactants of typical size. These improvements become increasingly valuable as the molecular size grows, as they bring the reaction possibilities growth under control. They have been proven very effective in systems with up to 42 reactive atoms, much more than was previously tractable. We even already used our approach to automatically discover a complex reaction mechanism, which served as a starting point for a DFT study we independently published.<sup>52</sup> Moreover, the improvements are very general and *a priori* applicable to any molecular system. Limitations in the accuracy of the predictions have been linked to approximations in treating tautomerizations and unspecified experimental details, such as traces of water, which are challenging to account for. A more sophisticated approach to handling tautomerization barriers would reduce those limitations and further narrow down the possible sources of error.

The atomic reactivity framework introduced here has been applied to organometallic catalysis, but we expect to develop it further to create pruning principles for both catalysed and uncatalysed polymolecular reactions. Developing more accurate atomic reactivity formulations will enable greater reaction clustering and reuse of existing data. A sufficiently precise formulation, would enable the reuse of reaction data *across* reaction networks, thus greatly minimizing the occurrence of unexplored unproductive reactions to evaluate. We also envision exploiting machine learning to leverage this data to provide a very fast way to estimate reaction barriers, thus further accelerating reaction network explorations. The resulting capabilities would have a remarkable potential in many fields of science which involve studying how organic molecules react and transform.

## 3 Computational Methodology

### 3.1 Reaction Modeling

Gold(I)-catalyzed reactions involve an organic reactant, a gold(I) catalyst (e.g.,  $\text{Ph}_3\text{PAuNTf}_2$ ) or a gold(I) precatalyst (e.g.,  $\text{Ph}_3\text{PAuCl}$ ) with an activator (e.g.,  $\text{AgOTf}$ ), and usually

a somewhat polar, aprotic and non-coordinating solvent, often dichloromethane or dichloroethane. Due to the noncoordinating nature of the catalyst counterion, we proposed in this study that it is safe to neglect it. Although it has been proposed to be involved as proton transfer agent in some reactions,<sup>53</sup> we aimed to address this effect using an alternative approach to explicit modeling (see section 2.2 of the ESI†), to limit model size and keep it unimolecular.

Gold catalysts involve a ligand, which has a great influence on the reaction outcome<sup>54</sup> and can be used to induce stereoselectivity.<sup>55</sup> However, the size of the ligand (typically 35 to 90 atoms) signifies a substantial increase in computational costs when using DFT and post-Hartree-Fock methods, which is difficult to manage, particularly in this context. Indeed, many computational studies involving such ligands rely on smaller model ligands, such as  $\text{PH}_3$  or  $\text{PMe}_3$ , to simplify calculations.<sup>56</sup>

To address this issue, we aimed to create a composite scheme for calculating free energies with reasonable computational costs and sufficient accuracy. In particular, recent semi-empirical tight-binding methods such as GFN2-xTB<sup>57,58</sup> have been reported as much faster alternatives to DFT methods for many applications.<sup>59</sup> Although those methods are typically not as accurate and reliable as a carefully selected DFT functional paired with a large basis set, they prove extremely useful when high accuracy is unnecessary or when DFT methods are simply impractical. Solvation is taken into account implicitly for every GFN2-xTB calculation using the ALPB solvation model with parameters for dichloromethane.<sup>60</sup>

Through a preliminary benchmark (see section 2.4 of the ESI†), we found that GFN2-xTB geometries were adequate for achieving good results, provided that the electronic energies are calculated using DFT. This made large complexes computationally manageable while ensuring the accuracy of the results. Single-point energy calculations, even with large ligands, were found to be perfectly reasonable. Furthermore, we devised pruning strategies to minimize the number of such calculations, as outlined in subsequent sections. After evaluating several computational methods, we decided on B2K-PLYP-D3/def2-mTZVP/SMD(DCM) with a reduced integration grid to perform single-point calculations energies. This method is expected to provide sufficiently accurate results compared to the more accurate and greatly more computationally expensive coupled-cluster methods.<sup>61–63</sup>

We have not explicitly evaluated the impact of different free energy corrections on the results. By default, the xtb package uses Grimme's entropy correction for low-lying frequency modes,<sup>64</sup> and Gibbs free energy corrections were only calculated at the GFN2-xTB theory level.

### 3.2 Automated Reaction Evaluation

The reaction evaluation process starts with the enumeration of all *a priori* reasonable reaction possibilities. We consider only  $\sigma$  bond formation and breakage, which also corresponds to atoms coming together or moving apart. This means that a Diels-Alder reaction would be a  $F_2$  reaction, and not  $F_3B_3$ . Relaxed coordinate scans are carried out to bring together or push apart the reactive atoms. This can be seen as an exploration of the potential en-



ergy surface (PES) biased by the expected reaction coordinates. The direct consequence of this is that the net transformation that occurs is *not limited to the expected one*. For example, by considering an  $F_1$  reaction approaching  $Cl^-$  from the carbon of  $CH_3I$ , we will discover an SN2 reaction, which is formally an  $F_1B_1$  reaction. Similarly, any formation or breakage of  $\pi$  bonds will occur naturally as the electron density moves during the reaction. This means that the considered reaction coordinates *do not need to be complete* and can have reduced dimensionality. Indeed, with this approach, we found that it was sufficient to consider  $B_1$ ,  $F_1$ ,  $F_1B_1$  and  $F_2$  reactions only.

In order to reduce the error rate related to spurious GFN2-xTB geometries, DFT gradients are calculated for each new product and compared to a threshold value determined from preliminary spurious and valid GFN2-xTB structures. If above this threshold, the structure is perturbed once following the steepest energy descent according to the DFT gradient, then reoptimized using the same theory level and loose convergence parameters. The optimized structure is reoptimized by GFN2-xTB with the expectation of converging to the same local minimum. This final structure is not modified further, even if it converges back to its initial state. For this process, the M06-2X/def2-mSVP/CPCM(DCM) theory level combined with geometrical counter-poise<sup>65</sup> (GCP(DFT/SVP)) was found to be adequate. Since this theory level is only used to break out of a potentially spurious local minimum, having a high accuracy is not necessary.

Relaxed scans provide crude reaction barriers to eliminate certain reaction possibilities at this stage. If a product emerges, its geometry will be optimized, and its energy will be computed. It can then be used with the reactant to perform a transition state search using the double-ended growing string method (GSM) developed by Chakraborty<sup>66</sup> and Zimmerman.<sup>67,68</sup> We used its implementation in the Pysisyphus Python package<sup>69</sup> combined with the GFN2-xTB level of theory. The approximate transition state obtained with the GSM is fully optimized using the RS-P-RFO algorithm as implemented in Pysisyphus. The vibrational modes of transition state are expected to contain one negative frequency below  $-75\text{ cm}^{-1}$  and that this frequency mode involves the movement of the expected reactive atoms. Spurious transition states are thus identified and filtered.

The immediate reactant and product are verified by carrying out an IRC calculation on the transition state, also with GFN2-xTB as level of theory. It is possible that the true reactant or product are not equivalent to the initial species that lead to this transition state. This is not an issue as long as the correct connectivities are considered in the reaction network.

### 3.3 Data-driven Pruning

The reaction  $F_{i-j}$  for intermediate  $A$  will be associated with a reaction barrier once evaluated. To reuse this data, we just need to consider the atomic reactivities as the relevant dimensions instead of the atom indices and nature of the intermediate. In other words, instead of storing  $(F_{i-j}, A) \Rightarrow \Delta G_{\alpha}^{\ddagger}$ , we will store  $(F, R_i, R_j) \Rightarrow \Delta G_{\alpha}^{\ddagger}$ . The major difference is that the former is necessarily unique (there is just one intermediate  $A$  and one possibility

$F_{i-j}$ ), while the latter is not ( $R_i$  and  $R_j$  stem from the reactivity of the system, not its label). However, in both cases, the same key will be linked with the same value (*i.e.*, reaction barrier).

In practice, the implementation adds the following modifications (see extended discussion in section 2.3 of the ESI<sup>†</sup>):

- To maximize the error cancellation, two reactions are considered equivalent if they involve equivalent atomic reactivities *and the same atom indices* (or atoms equivalent by symmetry). This partially accounts for geometrical constraints, which are difficult to account for in  $R$ . For example, hydrogens at different positions of a substituted cyclohexane might have equivalent atomic reactivities, but will differ in their feasible reactions due to their different positions on the ring (for example due to ring strain). However, if the steric effects around an atom or its ring strain change significantly, it implies that a reaction has taken place nearby, and so its electronic properties are expected to change. As such, changes in steric and ring strain effects are expected to also involve changes in  $R$ . This restriction related to atom indices then reduces the number of reaction possibilities that can be pruned, but is expected to avoid a very large percentage of invalid prunings due to steric and ring strain effects.
- The atomic reactivities are considered equivalent if they are within a certain range of each other (see threshold determination in subsection 2.3.2 of the ESI<sup>†</sup>).
- When using the barrier of an equivalent reaction to determine whether a given reaction should be evaluated, a certain margin of error is permitted. Depending on the estimated reliability, a barrier may exceed the exploration threshold by a specific margin without ruling out the considered reaction. Therefore, if the previously evaluated equivalent reaction was linked to an exceptionally large barrier, the reaction considered will be discarded. If that barrier was only slightly above the exploration threshold, it might not be sufficient to eliminate the considered reaction. The tolerance thresholds used in this work are presented and explained in the subsection 2.1.1 of the ESI<sup>†</sup>.

### 3.4 Complexation isomers, conformers and tautomers

In non-covalent adducts, the catalyst can move from one Lewis basic complexation site to another, which is explicitly considered in our approach. The complexation isomers are chosen to sample the whole range of possible activations by the catalyst, as measured through the  $R$  descriptor (see section 2.2 of the ESI<sup>†</sup> for the complete methodology). Conformers of covalent adducts are also sampled to find geometries that might exhibit a different reactivity.

In addition to reaction possibilities, proton transfers are considered possible and are explicitly considered. However, in nearly all reaction conditions, there is no strong acid or base, and so the reactant is considered to be the species allowing proton transfers. The proton transfer barriers are approximated through the Gibbs free energy of the transient protonated/deprotonated species.



The relative Gibbs free energy of this species is calculated by coupling the (de)protonation process with the inverse process involving the starting material. This models the net reaction of a proton transfer between the intermediate and a molecule of starting material. Water clusters have also been considered as acid and base, as indicated in the text. The technical approach is described in detail in the section 2.2 of the ESI†

### 3.5 Network Exploration Cycles

The network exploration is carried out in a series of cycles, during which a set of relevant intermediates are considered as reactants of possible new reactions. At the start of the exploration, the complexation isomers of the starting materials are generated (see section 2.2 of the ESI†) and considered as the reactive intermediates of the first cycle. All their reaction possibilities are generated and evaluated during this cycle. Tautomerizations are also considered for these intermediates. Once all those reactions have been evaluated, the updated reaction network undergoes a neophile kinetic simulation which will determine the new set of intermediates to consider in the following cycle. In addition to reactions and tautomerizations, complexation isomerisation is evaluated for all intermediates who aren't complexation isomers themselves.

The exploration stops when there is no new reactive intermediate to consider according to the neophile kinetic simulation. The product distribution is calculated using a normal kinetic simulation.

### Author contributions

R.R. conceived the project, lead the methodology development, obtained literature data, developed the software implementation, developed the visualization approach and wrote the original draft of the manuscript. C.Y.L. obtained funding and computational resources, contributed to the methodology development and provided supervision. Both authors co-revised and co-edited the manuscript.

### Conflicts of interest

R.R. is developing a commercial service based on the described methodology. All queries regarding this service should be sent to raphael@hexalence.com.

### Data availability

The data supporting this article have been included as part of the Supplementary Information, which includes all final reaction networks (including molecular structures, Gibbs free energy and original labels). They can be explored interactively using the provided static web pages, which were also used to produce the network illustrations of this paper. Additionally, the same raw data is provided in JSON format. The results of the preliminary theory level benchmark are also provided.

The specific software implementation used in this work is not available due to its importance in the commercial endeavour of one of the authors. Pseudocode can also be found in the Supplementary Information, and an archived version on Zenodo

(10.5281/zenodo.18635813).

### Acknowledgements

C.Y.L. and R.R. thank the Natural Science and Engineering Research Council (NSERC) of Canada, the Canada Foundation for Innovation (CFI), the FRQNT Centre in Green Chemistry and Catalysis (CGCC), and the Université de Sherbrooke for their support. Computational resources were provided by Calcul Québec and the Digital Research Alliance of Canada. R.R. is grateful to FRQNT for a Graduate Scholarship (B1X/B2X), Hydro-Québec and Université de Sherbrooke for a doctoral scholarship, and NSERC for a Canada Graduate Scholarship (CGS M) as well as a Vanier Canada Graduate Scholarship.

R.R. thanks Johannes Steinmetzer for providing help and prompt bug fixes for Pysisyphus.

### Notes and references

- 1 S. Grimme and P. R. Schreiner, *Angewandte Chemie International Edition*, 2018, **57**, 4170–4176.
- 2 C. A. Dvorak and V. H. Rawal, *Chemical Communications*, 1997, 2381–2382.
- 3 M. Alaboalirat, L. Qi, K. J. Arrington, S. Qian, J. K. Keum, H. Mei, K. C. Littrell, B. G. Sumpter, J.-M. Y. Carrillo, R. Verdusco and J. B. Matson, *Macromolecules*, 2019, **52**, 465–476.
- 4 K. K. Irikura and R. D. Johnson, *The Journal of Physical Chemistry A*, 2000, **104**, 2191–2194.
- 5 P. M. Zimmerman, *Molecular Simulation*, 2015, **41**, 43–54.
- 6 M. Jafari and P. M. Zimmerman, *Physical Chemistry Chemical Physics*, 2018, **20**, 7721–7729.
- 7 A. L. Dewyer, A. J. Argüelles and P. M. Zimmerman, *WIREs Computational Molecular Science*, 2018, **8**, e1354.
- 8 S. Maeda, Y. Harabuchi and K. Saita, *Molecular Technology*, John Wiley & Sons, Ltd, 2018, ch. 8, pp. 173–199.
- 9 Q. Zhao and B. M. Savoie, *Nature Computational Science*, 2021, **1**, 479–490.
- 10 A. Hashemi, S. Bougueroua, M.-P. Gageot and E. A. Pidko, *Journal of Chemical Theory and Computation*, 2022, **18**, 7470–7482.
- 11 A. Stan, B. von der Esch and C. Ochsenfeld, *Journal of Chemical Theory and Computation*, 2022, **18**, 6700–6712.
- 12 J. P. Unsleber, S. A. Grimm and M. Reiher, *Journal of Chemical Theory and Computation*, 2022, **18**, 5393–5409.
- 13 S. Maeda, Y. Harabuchi, H. Hayashi and T. Mita, *Annual Review of Physical Chemistry*, 2023, **74**, 287–311.
- 14 P. Ramos-Sánchez, J. N. Harvey and J. A. Gámez, *Journal of Computational Chemistry*, 2023, **44**, 27–42.
- 15 H. Hayashi, S. Maeda and T. Mita, *Chemical Science*, 2023, **14**, 11601–11616.
- 16 M. Bensberg and M. Reiher, *The Journal of Physical Chemistry A*, 2024, **128**, 4532–4547.
- 17 I. M. Pendleton, M. H. Perez-Temprano, M. S. Sanford and P. M. Zimmerman, *Journal of the American Chemical Society*, 2016, **138**, 6049–6060.
- 18 A. L. Dewyer and P. M. Zimmerman, *ACS Catalysis*, 2017, **7**,



- 5466–5477.
- 19 M. Bensberg and M. Reiher, *Israel Journal of Chemistry*, 2023, **63**, e202200123.
- 20 C. W. Gao, J. W. Allen, W. H. Green and R. H. West, *Computer Physics Communications*, 2016, **203**, 212–225.
- 21 M. Liu, A. Grinberg Dana, M. S. Johnson, M. J. Goldman, A. Jocher, A. M. Payne, C. A. Grambow, K. Han, N. W. Yee, E. J. Mazeau, K. Blondal, R. H. West, C. F. Goldsmith and W. H. Green, *Journal of Chemical Information and Modeling*, 2021, **61**, 2686–2696.
- 22 J. P. Unsleber, *Journal of Chemical Information and Modeling*, 2023, **63**, 3392–3403.
- 23 D. Rappoport, C. J. Galvin, D. Y. Zubarev and A. Aspuru-Guzik, *Journal of Chemical Theory and Computation*, 2014, **10**, 897–907.
- 24 T. Klucznik, L.-D. Syntrivanis, S. Baś, B. Mikulak-Klucznik, M. Moskal, S. Szymkuć, J. Mlynarski, L. Gadina, W. Beker, M. D. Burke, K. Tiefenbacher and B. A. Grzybowski, *Nature*, 2024, **625**, 508–515.
- 25 P. M. Zimmerman, *Journal of Computational Chemistry*, 2013, **34**, 1385–1392.
- 26 Q. Zhao and B. M. Savoie, *Angewandte Chemie International Edition*, 2022, **134**, e202210693.
- 27 B. M. Trost and G. Dong, *Nature*, 2008, **456**, 485–488.
- 28 S. A. Grimmel and M. Reiher, *CHIMIA*, 2021, **75**, 311–311.
- 29 M. Bergeler, G. N. Simm, J. Proppe and M. Reiher, *Journal of Chemical Theory and Computation*, 2015, **11**, 5712–5722.
- 30 M. S. Johnson, H.-W. Pang, M. Liu and W. H. Green, *International Journal of Chemical Kinetics*, 2025, **57**, 93–107.
- 31 M. Woulfe and B. M. Savoie, *Journal of Chemical Theory and Computation*, 2025, **21**, 1276–1291.
- 32 N. Mézailles, L. Ricard and F. Gagosz, *Organic Letters*, 2005, **7**, 4133–4136.
- 33 C. Nieto-Oberhuber, M. P. Muñoz, E. Buñuel, C. Nevado, D. J. Cárdenas and A. M. Echavarren, *Angewandte Chemie International Edition*, 2004, **43**, 2402–2406.
- 34 Y. T. Lee, Y. K. Kang and Y. K. Chung, *The Journal of Organic Chemistry*, 2009, **74**, 7922–7934.
- 35 A. Simonneau, Y. Harrak, L. Jeanne-Julien, G. Lemièrre, V. Mouriès-Mansuy, J.-P. Goddard, M. Malacria and L. Fensterbank, *ChemCatChem*, 2013, **5**, 1096–1099.
- 36 M. R. Luzung, J. P. Markham and F. D. Toste, *Journal of the American Chemical Society*, 2004, **126**, 10858–10859.
- 37 B. D. Sherry and F. D. Toste, *Journal of the American Chemical Society*, 2004, **126**, 15978–15979.
- 38 X. Li, J. Zhao, X. Xie and Y. Liu, *Organic & Biomolecular Chemistry*, 2017, **15**, 8119–8133.
- 39 N. D. Shapiro and F. D. Toste, *Journal of the American Chemical Society*, 2007, **129**, 4160–4161.
- 40 V. Mamane, T. Gress, H. Krause and A. Fürstner, *Journal of the American Chemical Society*, 2004, **126**, 8654–8655.
- 41 D. Garayalde, E. Gómez-Bengoia, X. Huang, A. Goeke and C. Nevado, *Journal of the American Chemical Society*, 2010, **132**, 4720–4730.
- 42 S. Wang and L. Zhang, *Journal of the American Chemical Society*, 2006, **128**, 8414–8415.
- 43 Y.-Q. Gu, P.-P. Zhang, J.-K. Fu, S. Liu, Y. Lan, J.-X. Gong and Z. Yang, *Advanced Synthesis & Catalysis*, 2016, **358**, 1392–1397.
- 44 C. Wang, Q. Cui, Z. Zhang, Z.-J. Yao, S. Wang and Z.-X. Yu, *Chemistry – A European Journal*, 2019, **25**, 9821–9826.
- 45 Y.-P. Xiao, X.-Y. Liu and C.-M. Che, *Angewandte Chemie International Edition*, 2011, **50**, 4937–4941.
- 46 C. Zhang, H. Jiang and S. Zhu, *Chemical Communications*, 2017, **53**, 2677–2680.
- 47 S. Dhillon and A. L. L. East, *International Journal of Quantum Chemistry*, 2018, **118**, e25703.
- 48 M.-C. Cordonnier, A. Blanc and P. Pale, *Organic Letters*, 2008, **10**, 1569–1572.
- 49 J. M. Fernández-García, H. A. Garro, L. Fernández-García, P. García-García, M. A. Fernández-Rodríguez, I. Merino and E. Aguilar, *Advanced Synthesis & Catalysis*, 2017, **359**, 3035–3051.
- 50 M. Brandstätter, N. Huwyler and E. M. Carreira, *Chemical Science*, 2019, **10**, 8219–8223.
- 51 A. S. K. Hashmi, *Accounts of Chemical Research*, 2014, **47**, 864–876.
- 52 R. Robidas and C. Y. Legault, *Chemistry – A European Journal*, 2023, **29**, e202301510.
- 53 J. Shi, S. Bi, Y.-Y. Jiang, Y. Liu, B. Ling and X.-A. Yuan, *Journal of Organometallic Chemistry*, 2018, **864**, 136–142.
- 54 Z. Lu, G. B. Hammond and B. Xu, *Accounts of Chemical Research*, 2019, **52**, 1275–1288.
- 55 W. Zi and F. D. Toste, *Chemical Society Reviews*, 2016, **45**, 4567–4589.
- 56 O. N. Faza and C. S. López, *Homogeneous Gold Catalysis*, Springer International Publishing, Cham, 2015, pp. 213–283.
- 57 C. Bannwarth, S. Ehlert and S. Grimme, *Journal of Chemical Theory and Computation*, 2019, **15**, 1652–1671.
- 58 H. Neugebauer, B. Bädorf, S. Ehlert, A. Hansen and S. Grimme, *Journal of Computational Chemistry*, 2023, **44**, 2120–2129.
- 59 C. Bannwarth, E. Caldeweyher, S. Ehlert, A. Hansen, P. Pracht, J. Seibert, S. Spicher and S. Grimme, *WIREs Computational Molecular Science*, 2021, **11**, e1493.
- 60 S. Ehlert, M. Stahn, S. Spicher and S. Grimme, *Journal of Chemical Theory and Computation*, 2021, **17**, 4250–4261.
- 61 A. Karton and L. Goerigk, *Journal of Computational Chemistry*, 2015, **36**, 622–632.
- 62 L.-J. Yu, F. Sarrami, R. J. O'Reilly and A. Karton, *Chemical Physics*, 2015, **458**, 1–8.
- 63 A. Karton, *The Journal of Physical Chemistry A*, 2019, **123**, 6720–6732.
- 64 S. Grimme, *Chemistry – A European Journal*, 2012, **18**, 9955–9964.
- 65 H. Kruse and S. Grimme, *The Journal of Chemical Physics*, 2012, **136**, 154101.
- 66 B. Peters, A. Heyden, A. T. Bell and A. Chakraborty, *The Jour-*



- nal of Chemical Physics*, 2004, **120**, 7877–7886.
- 67 P. M. Zimmerman, *The Journal of Chemical Physics*, 2013, **138**, 184102.
- 68 P. M. Zimmerman, *Journal of Chemical Theory and Computation*, 2013, **9**, 3043–3050.
- 69 J. Steinmetzer, S. Kupfer and S. Gräfe, *International Journal of Quantum Chemistry*, 2021, **121**, e26390.



## Data availability statement of “Distilling System Complexity to Enable Unbiased and Predictive Computational Reaction Investigations”

The data supporting this article have been included as part of the Supplementary Information, which includes all final reaction networks (including molecular structures, Gibbs free energy and original labels). They can be explored interactively using the provided static web pages, which were also used to produce the network illustrations of this paper. Additionally, the same raw data is provided in JSON format. The results of the preliminary theory level benchmark are also provided.

The specific software implementation used in this work is not available due to its importance in the commercial endeavor of one of the authors. Pseudocode can also be found in the Supplementary Information, and an archived version on Zenodo (<https://doi.org/10.5281/zenodo.18635813>).

