De Novo Catalyst Discovery: Fast Identification of New Catalyst Candidates

Maria H. Rasmussen¹,², Julius Seumer¹, and Jan H. Jensen¹,²

¹Department of Chemistry, University of Copenhagen, Denmark
²E-mail: mhr@chem.ku.dk, jhjensen@chem.ku.dk, Twitter: @janhjensen

June 30, 2023

Abstract

Very recently our group has demonstrated how a genetic algorithm (GA) starting from random tertiary amines can be used to discover a new and efficient catalyst for the alcohol-mediated Morita-Baylis-Hillman (MBH) reaction. In particular, the discovered catalyst was shown experimentally to be eight times more active than 1,4-diazabicyclo[2.2.2]octane (DABCO), which is commonly used to catalyze the MBH reaction. This represents a breakthrough in using generative models for catalyst optimization. However, the GA-procedure, and hence discovery, relied on two important pieces of information: 1) the knowledge that tertiary amines catalyze the reaction and 2) the mechanism and reaction profile for the catalyzed reaction, in particular the transition state structure of the rate determining step. Thus truly de novo catalyst discovery must also include these steps. Here we present such a method for discovering catalyst candidates for a specific reaction while simultaneously proposing a mechanism for the catalyzed reaction. We use the method to show that tertiary amines and phosphines are potential catalysts for the MBH reaction by screening a selection of 11 molecular templates representing common functional groups. The presented method relies on an automated reaction discovery workflow using meta-dynamics calculations. Combining this method for catalyst candidate discovery with our GA-based catalyst optimization method results in an algorithm for truly de novo catalyst discovery.

1 Introduction

The search for new catalysts is instrumental in addressing some of humanities current challenges. The right catalyst in principle holds the key to selective and sustainable production of target molecules; an important task in all areas of chemical research and industry. However, several factors make catalyst discovery challenging, slow and expensive. Detailed knowledge about the reaction network including side reactions and solvent reactions is needed and these steps still rely heavily on chemical intuition and experimentation.

So far, the main contribution of computational chemistry methods to catalyst discovery has been in establishing a mechanism for catalytic cycles of known catalysts. Current state of the art within computationally driven catalyst discovery is generally focused on improving known catalysts based on the catalytic mechanism. One approach is to screen large libraries of molecules that are structurally similar to the known catalyst. For example Nandy et al. combined machine learning (ML) and density functional theory (DFT) methods to screen a library of 16 million candidates for the catalysis of the methane to methanol oxidation. Another recent example of computationally driven catalyst optimization from Cramer et al. used mechanistic information of three intertwined catalytic cycles for CO₂ conversion to formic acid, formaldehyde and methanol to generate theoretically founded rules for predicting catalytic activity and selectivity. The theoretical framework was used to predict a catalyst that optimizes selectivity towards formaldehyde, which was verified experimentally with an 81% yield. Das et al. mapped catalytic activity in CO₂ hydrogenation
to acid/base properties of frustrated Lewis pairs to predict a specific combination of a Lewis acid and a Lewis base from approx 4000 candidate pairs which was experimentally verified to catalyze the reaction.\(^3\)

An alternative to the screening approach is to use generative models to propose new catalyst candidates. The advantage of this approach is that one is not limited to discovering catalyst candidates already present in a predefined library. One such approach; a graph-based genetic algorithm (GA), was recently employed by our group to discover a new catalyst with a previously untested structural motif for the alcohol-mediated Morita-Baylis-Hillman (MBH) reaction.\(^4\)\(^5\) The proposed catalyst candidate was experimentally verified to be eight times more active than 1,4-diazabicyclo[2.2.2]octane (DABCO), a commonly used catalyst for the reaction.\(^5\)

Vital to these important findings is mechanistic insight into the relevant catalytic cycle. Typically, these catalytic cycles are generated by a combination of experimental work, expert knowledge and quantum chemical calculations, making this step a laborious undertaking. Moreover, someone needs to actually have the idea of testing a specific molecular moiety as a catalyst for a reaction. Thus, a method for truly de novo catalyst discovery must also include a solution for this part of the discovery process.

Since catalytic activity can be extracted from a reaction network including the reactants and potential catalyst candidate, many groups (including us) have worked to develop methods for automated reaction discovery for the generation and exploration of reaction networks.\(^6\)\(^7\)\(^8\)\(^9\)\(^10\)\(^11\)\(^12\)\(^13\)\(^14\)\(^15\)\(^16\)\(^17\)\(^18\) This in principle provides a way of discovering completely novel catalysts for reactions with no prior knowledge of catalytic activity.

Some proof-of-principle papers have been presented using these methods to discover the mechanism of a catalytic cycle. Specifically the cobalt-catalyzed alkene hydroformylation has been a popular example, since the system consists of only 18 atoms.\(^19\)\(^20\)\(^21\)\(^22\)\(^23\)\(^24\)\(^25\)\(^26\) A more widespread application of these reaction discovery methods is mostly hindered by the vast computational cost of freely growing several reaction networks. In order to move from hypothesis-testing to discovery, the methods need to be efficient enough that screening of different potential catalyst candidates is possible.

Our method for exploring reaction networks is based on the meta-molecular dynamics (meta-MD) approach proposed by Grimme.\(^23\) As we have demonstrated previously, an automated workflow that tracks the reactions occurring during the meta-MD simulations can be used as an efficient way of predicting which low-barrier (defined as <30 kcal/mol) reactions are possible.\(^22\)\(^23\) The combination of using a method that focuses on low-barrier reactions while relying on a fast semi-empirical quantum chemistry method (GFN2-xTB)\(^23\) means that we can grow reaction networks quite fast even for larger molecular systems, making screening applications possible.

In this work, we present a method for discovering that a certain catalyst-template such as a tertiary amine can be used to catalyze a certain reaction such as the MBH reaction. We build reaction networks for the reactants of the MBH reaction with 11 different catalyst-templates and use them to detect the presence/absence of catalytic activity. As part of the method we also map out a proposed mechanism for the catalyzed reaction. We demonstrate the method by re-discovering tertiary amines and phosphines as catalysts for the MBH reaction. From the found reaction profile we extract a transition state (TS) template for the rate-determining step. With the TS-template available, the genetic algorithm (GA) applied by Seumer et al. can be used to optimize the catalytic activity. This represents an end-to-end approach for de novo catalyst discovery starting from no prior knowledge about how a reaction can be catalyzed and ending with a range of possible catalysts optimized for said reaction (Figure 1).

Within drug discovery, two of the crucial steps are lead identification and lead optimization. Lead identification is the task of finding a compound that is active against a specific drug target. Once a lead
compound is identified, lead optimization represents the process of making structural changes that improves activity and selectivity while reducing toxicity and other unwanted side effects. The analogy to catalyst discovery is clear and while current state-of-the-art computationally driven catalyst discovery has been focused on catalyst optimization (i.e. lead optimization), this work represents a step forward in filling in the gap for lead identification.

In section 2 the computational methodology is described, in section 3 we present results for the found reaction networks as well as the subsequent GA search and we validate one of the suggested catalysts with density functional theory (DFT) calculations and in section 4 we summarize our findings and present an outlook for the method.

2 Computational Methodology

We compute a reaction network for each of several possible catalyst candidates. Our method for building reaction networks relies on extracting possible single-step reactions from a number of meta-MD runs. Below we give a short summary of the method but details can be found in 12, 11 and 23. Unlike more systematic approaches such as graph enumeration of products or reaction coordinates, 10, 8, 6, 11, 25 where the search for intermediates is done in a (semi-) exhaustive manner, our goal is not to exhaustively find all possible intermediates. Rather, the goal is to find all low barrier reactions (<30 kcal/mol). While using a method based on molecular dynamics simulations makes the search for one intermediate more expensive than e.g. graph enumeration of products, we typically generate several orders of magnitude fewer intermediates to be analyzed. Therefore, we can afford a much more accurate screening procedure of the found intermediates. Furthermore, locating the intermediates without finding the barrier simultaneously (as some other methods do, see e.g. 16 and 20) allows us to efficiently perform an initial screen based on the energy of the intermediates.
Our workflow for generating reaction networks can be summarized in four steps:

1. **Predict new intermediates** To predict possible single-step reactions from a reactant system, we run 100 meta-MD simulations for every stereoisomer of the reactant system as described in references \[12\] and \[11\]. The simulation is stopped when a reaction involving a change in bonding is observed. We only keep intermediates/products that are found more than five times when combining reactions from meta-MD simulations across the different stereoisomers.

2. **Generate tautomers and proton-transfer intermediates** Solvent mediates proton transfer are often important low-barrier steps in the mechanism. While these in principle can be modeled with meta-MD it is more efficient to generate them systematically with tools like RDKit.\[27\] We consider the following two kinds of proton transfer reactions for every saved intermediate found in step (1). We first generate all tautomers using the TautomerEnumerator() function in RDKit.\[27\] Second, for every tautomer we consider proton transfers between equivalent heteroatoms. For example, if both an alkoxy group an a hydroxy group is present in the tautomer, we consider the intermediate formed after a proton transfer from the hydroxy group to the alkoxy group (Figure 2). Similarly if e.g. both a protonated amine and a neutral amine is present, we consider proton transfer between the amine groups.

3. **Screen intermediates from (1) and (2) based on reaction energies** In this study we screen the found intermediates based on their Gibbs free energies relative to the original reactant system being <30 kcal/mol. The Gibbs free energies are calculated for isolated molecules by the following steps: First we embed 3+3N_{rot} conformers (N_{rot} is the number of rotatable bonds in the molecule) using ETKDGv3 in RDKit.\[28;27\] Each conformer is then optimized at the GFN2-xTB level of theory where methanol is modelled as an implicit solvent using the generalized Born and surface area model (GBSA) through the xTB program.\[24;29\] Following Seumer et al.\[5\] the conformer with lowest energy at the GFN2-xTB level of theory is optimized at the B3LYP-D3/6-31+G(d,p) level of theory with methanol modelled using the SMD continuum solvent model in Gaussian 16.\[30;31;32;33;34;35;36\] If the optimization fails to converge or a change in bonding is observed during the optimization, the intermediate is discarded. In case any of the intermediates generated in step (2) are no longer connected to the main reaction network after screening of the intermediates, these are also discarded.

4. **Take a new step in the reaction network** All saved intermediates that have not previously gone through step (1) now serve as reactant systems in the next iteration of generating the reaction network. In case a reactant system consists of more than two molecular entities, we continue with the two largest molecular species. For these new reaction systems we repeat step (1) through (4).

Our method relies on Grimme’s meta-MD method to generate relevant single-step intermediates/products\[23\]. In short this is a molecular dynamics simulation where additional biasing potentials are added to the GFN2-xTB potential to speed up the occurrence of reactive events. The biasing potentials are added every 1 ps...
and work by penalizing previously visited structures. After \( n \) ps the combined biasing potential will have the following form:

\[
E_{\text{RMSD bias}} = \sum_{i}^{n} k_{\text{push}} \cdot e^{-\alpha \Delta_{i}^{2}}
\]  

(1)

\( \Delta_{i} \) is the RMSD between the current structure and the structure saved after \( i \) ps of the simulation. In this way the meta-MD simulation is “pushed” towards unexplored molecular configurations e.g. different conformations and chemical reactions. The magnitude of this pushing is controlled by the parameter \( k_{\text{push}} \). The width of these Gaussian biasing potentials is controlled by the parameter \( \alpha \). In previous studies we have found the combination of \( k_{\text{push}} = 0.05 \cdot E_{\text{h}} \) and \( \alpha = 0.3 \) Bohr\(^{-2} \) to generally work well for finding low-barrier reactions.\(^{11,12} \) The molecular system is encapsulated in a spherical cavity for which a scaling parameter \( s \) is used to control the size. We have previously found it necessary to gradually decrease this cavity for bimolecular reactions. This is done by decreasing \( s \) with 0.02 every 5 ps in case no reaction has occurred.\(^{12} \) We use \( s = 1.0 \) as the initial scaling factor for the cavity.

Note that in this work we choose not to screen the found reactions based on barriers but only reaction energies. These choices are possible since the number of intermediates generated by our meta-MD procedure is low enough that we can grow the reaction network quite a bit with screening based only on reaction free energies (in this case three cycles of step 1-4 above) before doing another cycle would become computationally unfeasible (months on our local cluster). Subsequently we can define the parts of the reaction networks found that seem most interesting. In this case the application is to look for potential catalysts, therefore we look for places in the reaction network where the catalyst is regenerated. Meanwhile the reaction energy for the catalyzed reaction should not be very endothermic, the threshold will depend on the accuracy of the DFT method used for calculating the reaction energies.

3 Results and Discussion

To test the meta-MD based reaction discovery method described above we set out to rediscover tertiary amines and phosphines as catalysts for the MBH reaction. The MBH reaction is represented by the reaction of methyl acrylate (MA) with \( p \)-nitrobenzaldehyde (\( p \)NBA) following Seumer et al.\(^{5} \) The task is then to find a catalyst template that can catalyze the reaction between MA and \( p \)NBA. To demonstrate this, we initially provide a list of eight possible catalyst templates (Round 1), for which reaction networks with the reactants (MA and \( p \)NBA) are grown following step 1-4 in Section 2. Based on an analysis of the most promising reaction networks grown in Round 1, three new possible catalyst templates are tested (Round 2).

3.1 Generating reaction networks for the potential catalyst templates

3.1.1 Round 1

We build reaction networks for eight possible templates (Figure 3, Round 1) with a variety of functionality. In particular, we include several different kinds of nucleophiles; ethene (representing a double bond), hydroxide (nucleophile and strong base), ammonia (nucleophile and weak base) and phosphine. We also include entries that act as acids; the hydronium ion (strong acid, conjugate base weak nucleophile), hydrogen sulfide (weak acid, conjugate base strong nucleophile) and formic acid (representing a carboxylic acid). Formaldehyde is included as an example of an electrophile.

In the first step of generating reaction networks for the eight template structures in Figure 3 we include MA and each of the catalyst templates in the meta-MD run. Before doing another step of the reaction discovery workflow, we add \( p \)NBA to the reaction mixture. Since we only consider elementary reactions between two molecules, in case more than two molecules are present in an intermediate, we only search for reactions between the two largest molecules.
The eight reaction networks vary a lot in size after three iterations of the procedure. For formic acid (C8), the network consists of only 13 intermediates (Figure S1a) while for phosphine (C5) the network consists of 298 intermediates (Figure S1b). For some of the catalyst templates (C3, C5 and C6), the inclusion of tautomers adds a vast number of new intermediates; notice the number of yellow edges (proton transfer reactions) for phosphine (Figure S1b).

Of the eight catalyst templates tested, only the reaction network for ammonia (C3) shows clear signs of catalytic activity, defined by the presence of a node with the catalyst regenerated and reaction free energy < 10 kcal/mol. For the remaining seven reaction networks, some show no intermediate/product with the template regenerated (C2, C6 and C7) while some have 1-3 intermediates with the catalyst regenerated but all with reaction Gibbs free energies >10 kcal/mol (C1, C4, C5 and C8). Phosphine (C5) is thus not recognized as a potential catalyst from the reaction network grown after three iterations. Below, we will analyze the behavior of phosphine in a bit more detail by comparing with the reaction network found for ammonia.

For ammonia, the primary reaction path found with meta-MD as the first step is nucleophilic attack at the $\beta$-carbon of MA to form intermediate 13 (Figure 4a). Another major reaction is transfer of a proton from ammonia to oxygen simultaneously with the nucleophilic attack forming intermediate 16. For phosphine, we see another primary reaction path for the first step of meta-MD generated products which is the addition of phosphine to the double bond forming intermediate 12 (Figure 4b). While the second most observed reaction is nucleophilic attack of phosphine to the $\beta$-carbon of MA forming intermediate 25, the Gibbs free energy at 27 kcal/mol is much higher and in fact very close to our cutoff of 30 kcal/mol. In ammonia’s case we do find the next step of the MBH reaction when reacting intermediate 13 with pNBA forming intermediate 209. However, when trying to optimize this intermediate at the DFT level, we observe a proton transfer between nitrogen and oxygen forming intermediate 138 (Figure 4a). Intermediate 430 is found within the same iteration as a tautomer to intermediate 138. For phosphine on the other hand we observe another intermediate where a five-membered ring is formed from phosphor and oxygen binding (intermediate 151) which is possible due to phosphors ability to form 5 bonds. We do find the expected MBH intermediate (intermediate 371) in the next meta-MD iteration and with it intermediate 1752 from a proton transfer reaction which is the equivalent to intermediate 430 for ammonia. Thus, the reason we are not observing the MBH product for phosphine is that the extra intermediate observed on the path (intermediate 151) means that it would take another, fourth, iteration of our reaction discovery method to get there.

Based on the initial analysis of reaction networks for the eight potential catalyst candidates (Figure 3 red box) ammonia is the most promising being the only candidate that shows catalytic activity after three iterations of the reaction discovery procedure. For both ammonia and phosphine we see that the possibility of a proton transfer from the nitrogen/phosphor atom creates a vast amount of reaction channels (Figures S2 and S1b). A way of hindering those reactions would be to exchange the hydrogens in ammonia and phosphine with methyl groups forming trimethylamine (TMA) and trimethylphosphine (TMP), respectively.
Figure 4: Highlighted reaction paths towards the MBH product for (a) C3: ammonia, (b) C5: phoshine, (c) C9: TMA and (d) C11: TMP.

For completeness, we also try dimethyl sulfide (DMS) formed by changing the hydrogens in hydrogen sulfide to methyl groups. Thus, in the second round of testing possible catalyst templates, we include TMA (C9), DMS (C10) and TMP (C11) (Figure 3, Round 2).

3.1.2 Round 2

Figure 5 shows the reaction network grown for TMA (C9) after three iterations of the meta-MD procedure. Importantly, the MBH reaction is discovered as a four-step reaction (highlighted in green). From the initial
Figure 5: The reaction network grown for TMA (C9). Structures of all found intermediates/products not involving a reaction at the nitro group of pNBA are shown. Red arrows represent reactions found with meta-MD, while orange arrows represent the proton-transfer reactions described in step 2 (section 2). The transparency of the red arrows indicate how many times a reaction is found in the meta-MD simulations. A fully colored arrow is found at least 30 times. Only reactions found at least five times in the meta-MD simulations are included in the network. The MBH reaction path found as part of the network is highlighted in green.

reactant system (TMA+MA, intermediate 76 in the reaction network), the only elementary reaction found with meta-MD is the attack of TMA on the \( \beta \)-carbon (resulting in intermediate 77), which is indeed the first step of the MBH reaction. From this point we observe four different reaction paths, all involving a nucleophilic attack of the enolate anion. Attack at the carbonyl carbon of pNBA results in the second intermediate of the MBH reaction (intermediate 166). For the additional three elementary reactions originating at intermediate 77, the enolate anion attacks at either a benzene carbon ortho to the nitro group (intermediate 150 and intermediate 176) or at an oxygen of the nitro group (intermediate 150 and intermediate 176). The third step of the MBH reaction is found as a proton-transfer reaction from intermediate 166 (generating intermediate 446) and from here on, the only reaction found by the meta-MD approach is the final step of the MBH reaction to form the product (intermediate).

TMP (C11), the phosphor-equivalent to TMA, generally behaves similarly with key features of the reaction networks being identical (Figure S3). Importantly, the MBH reaction path is also identified: intermediate 74 \( \rightarrow \) intermediate 75 \( \rightarrow \) intermediate 178 \( \rightarrow \) intermediate 456 \( \rightarrow \) intermediate 558. The reaction network for (C10) indicates no sign of catalytic activity.
For both TMA (Figure 4c) and TMP (Figure 4d) we find only a single elementary reaction for MA + TMA/TMP which is indeed the expected first step of the MBH mechanism (intermediates 77 and 75, respectively). This is contrary to what we observed for ammonia and phosphine, where competing paths where found already from the first step. Also, the intermediate free energies are significantly lower for TMA/TMP. We generally find far fewer reaction paths for TMA and TMP compared to ammonia and phosphine resulting in the much simpler reaction networks (Figures 5 and S3). For both TMA and TMP we find the expected MBH mechanism (Figures 4c and 4d). Without considering any barriers but solely based on Gibbs free energies of the intermediates, we would expect TMA and TMP to be better starting points for catalyst optimization. Here, we focus on TMA as a starting point for catalyst optimization using a genetic algorithm as done in [5].

3.2 Catalyst optimization using a genetic algorithm based on the catalyst template

3.2.1 Validation of the catalyst candidate

Having obtained a possible catalyst template (TMA) and reaction mechanism from the reaction networks grown, we need to (1) validate the mechanism by calculating transition states (TSs) for all steps in the proposed mechanism and (2) consider the side-reactions suggested by the reaction network. In particular, having obtained a barrier for the rate-determining step in (1), we can evaluate competing reactions on their barrier heights being higher/lower than the rate-determining step.

![Figure 6: Calculated reaction profile for the mechanism shown in Figure 4c for TMA and DABCO. The TS structures are optimized at the B3LYP/6-31+G(d,p)/SMD(methanol) level of theory using Gaussian 16.](https://doi.org/10.26434/chemrxiv-2023-lchmv)

Finding TSs for testing possible mechanisms (typically suggested by experimental chemists) is an important part of computational chemistry. While many promising methods for automating this process have been proposed in the last couple of decades [37;38;39;40;41], much work regarding finding TSs is still done manually. For a non-screening application like this one, where a mechanism is proposed and a handful of TSs need to be found, there would likely be some degree of manual adjustment/evaluation. Numerous computational studies have demonstrated this kind of work for the MBH mechanism [42;43;44]. The TS-structures for the TMA catalytic cycle (steps highlighted with a full green arrow in Figure 5) are also found manually here. The full
reaction profile, with the proton transfer from intermediate 166 to intermediate 446 conducted by methanol, is shown in Figure 6. The energies and barriers for equivalent steps using DABCO as the catalyst are shown for comparison. As expected, TMA is calculated to be a worse catalyst than DABCO. With a calculated activation energy 1.3 kcal/mol higher than DABCO, the reaction catalyzed by TMA is expected to be roughly 9 times slower. The catalyst templates chosen need to be rather generic in order to reduce the number of templates to be investigated by growing their reaction networks. Therefore, we generally do not expect to find good catalysts among the templates. Rather, we hope to find a starting point for further optimization.

The activation energy for TMA is found to be 23.4 kcal/mol. Side reactions having activation energies that are lower than or close to this value must thus be considered. This can be done by incorporating knowledge of side reactions in the scoring function of the genetic algorithm e.g. a given catalyst is optimized to both lower the activation energy for the target reaction while simultaneously raising it for side reactions.

![Figure 7: Possibly relevant side reactions extracted from the TMA reaction network in Figure 5.](https://doi.org/10.26434/chemrxiv-2023-lchmv)

From the reaction network of TMA, we see that any side reaction needs to go through intermediate 150, 176, 216 or 553 (Figure 5). The primary path to intermediates 150, 176 and 216 originates from intermediate 77 while intermediate 553 is formed from intermediate 166 (Figure 7). We find neither new meta-MD nor proton-transfer reactions from intermediate 216 - only the back-reaction to intermediate 77 is observed. With a Gibbs free energy 14 kcal/mol higher than the reactants, the system will also not get stuck here meaning that this side-reaction is deemed unimportant. Both intermediates 150 and 176 have high Gibbs free energies close to the activation energy of the MBH reaction; 23 and 20 kcal/mol, respectively. However, if the barriers are low enough they could act as “gateways” to products with lower reaction energies than the MBH product (Figure 5). Thus we use a previously published method for finding TS guess structures based on the same kind of biasing potentials used in the meta-MD to find TSs for the reaction between intermediates 77 and 150 and between intermediates 77 and 176 (23;40;11;12). For the 77 → 176 reaction we find a barrier of 29 kcal/mol. When searching for the 77 → 150 TS, we instead find the TS for the 176 → 150 reaction with a barrier of 25 kcal/mol. Looking at intermediates 150 and 177, it makes sense that the 77 → 150 goes through intermediate 176. Thus, we expect a 29 kcal/mol barrier for getting to either intermediates. Compared to a barrier of 23 kcal/mol for the MBH reaction we do not deem these side-reactions important and ignore them in the GA scoring function. Finally, intermediate 553 has a Gibbs free energy 24 kcal/mol higher than the reactant system. Since the intermediate Gibbs free energy is already
higher than the MBH activation energy, this side-reaction is also deemed unimportant.

### 3.2.2 Results of the Genetic algorithm optimization

In this case, no side-reactions of the network are deemed important and we score the catalyst solely based on the reactant to TS3 GFN2-xTB electronic energy barrier as done in the original study\(^5\). The five GA searches done by Seumer et al. are now repeated with the same starting populations (100 generations, a population size of 100 and a mutation rate of 50%, details can be found in \(^5\)). The only difference is, that while the original study scored the catalysts based on a GFN2-xTB TS template from a known catalyst (DABCO) we score the catalysts based on a GFN2-xTB TS template from the simple and generic catalyst template, TMA. As the GA optimization has several stochastic elements, we do not expect to find the exact same final populations. However, molecules containing an azetidine ring are still dominating the final populations and in fact 20 of the catalysts are present in both this and the original study (Figure S4). We calculate the TS and barrier for the rate-determining step (TS3) for three of them (m5, m6 and m10) at the DFT level and find an activation energy ≈2 kcal/mol lower than DABCO (≈20 kcal/mol vs. 22.1 kcal/mol) for all of them. The reductions are similar to what was found by Seumer et al. for the two molecules calculated at the DFT level (1.7 and 2.4 kcal/mol\(^5\)). For m10 we compute barriers for the remaining steps in the reaction path and confirm that TS3 still corresponds to the rate-determining step (Figure S). This shows that the GA catalyst optimization procedure proposed by Seumer et al. is not dependent on the availability of an already known catalyst (e.g. DABCO). Rather, one can get to catalysts performing equally well on the DFT level starting from a bad catalyst (e.g. TMA).

![Diagram](attachment:diagram.png)

Figure 8: Calculated reaction profile for the MBH mechanism for TMA, DABCO and m10. The TSs are optimized at the B3LYP/6-31+G(d,p)/SMD(methanol) level of theory using Gaussian 16.\(^{10}\)

### 4 Conclusions and outlook

Full de novo catalyst discovery requires both automated lead identification and optimization. We have previously shown that the latter step can be done efficiently using a graph-based genetic algorithm, given
the TS structure of the rate determining step in the mechanism. Here, we present a method for the first step.

We use the meta-MD method developed by Grimme to automatically determine reaction networks for possible catalyst candidates. From the reaction networks we extract the presence/absence of catalytic activity as well as a mechanism for the catalytic cycle. We find that the TS of the rate determining step for the found catalyst candidate can be used for catalyst optimization (lead optimization).

We demonstrate the method by using it to rediscover tertiary amines and phosphines as catalysts for the MBH reaction. Building reaction networks for 11 possible catalysts candidates representing different functional groups (amines, phosphines, sulfides, alkenes, acids, bases and carbonyl groups), tertiary amines and phosphines clearly stand out as the most promising catalyst candidates. Furthermore, the reaction networks are used to extract information about possible side reactions; in this case we identified no relevant side-reactions that are competitive with the desired catalytic mechanism. The tertiary amine template, TMA, identified by the screening of possible catalyst candidates is used as a starting point for the GA proposed by Seumer et al., and found that the optimized catalyst candidates performed similarly to the experimentally validated M10 catalyst suggested by Seumer et al. at the B3LYP/6-31+G(d,p) level of theory with methanol modelled as a continuum solvent using the SMD model.

A method for catalyst candidate identification requires that screening of possible candidates is practically possible, meaning that growing the reaction networks must be fast. We achieve this by focusing on screening intermediate Gibbs free energies rather than searching for TSs. Furthermore, the meta-MD simulations are performed at the fast semi-empirical level of theory GFN2-xTB.

The combination of this method for catalyst candidate discovery with the GA-based catalyst optimization method by Seumer et al. results in an algorithm for truly de novo catalyst discovery.

We note that the MBH catalytic cycle is in many ways an ideal case for the proposed method. First of all, the family of molecules catalyzing the reaction (tertiary amines and phosphines) is quite simple in terms of functionality. One can easily imagine reactions requiring a much more complicated functionality of the catalyst. This affects the size and nature of the library of possible catalyst candidates, we need to test. It is very likely that some kind of iterative procedure in terms of updating the library will be necessary to implement in order to find something good enough that the genetic algorithm can take over. Second, this catalytic cycle is relatively simple (four steps in the mechanism) and no additives, acids, bases etc. is needed. Clearly, testing several conditions for each catalyst template will increase the cost significantly. However, the cost is unlikely to become prohibitive and even several of months of computing is an acceptable investment of effort for discovering a novel catalyst candidate.

Another thing to note is that organometallic catalysts are very important in the field of catalysis. How this method works for transition metal compounds is an important yet still unanswered question that we will continue working on providing an answer for.

Thus while this method by no means represents the final word for method development within de novo catalyst discovery, it is an important step in the right direction.

5 Acknowledgements

This work was supported by Novo Nordisk Fonden via grant number NNF20OC0064104.
References


Supporting Information

The code and data resulting from this study can be found here [https://github.com/jensengroup/MBH_CatalystDiscovery](https://github.com/jensengroup/MBH_CatalystDiscovery) and [https://sid.erda.dk/sharelink/C4RLJdhC5](https://sid.erda.dk/sharelink/C4RLJdhC5), respectively.

Figure S1: Examples of reaction networks. (a) The smallest (13 intermediates) reaction network grown in three iterations for the first iteration of potential catalyst candidates is for formic acid (C8). (b) The largest reaction network (298 intermediates) grown in three iterations is for phosphine (C5).

Figure S2: Reaction network grown for ammonia (C3) after three iterations of the reaction discovery method.
Figure S3: The reaction network grown for trimethylphosphine (C11). Red arrows represent reactions found with meta-MD, while orange arrows represent the proton-transfer reactions described in step 2 (section 2). The transparency of the red arrows indicate how many times a reaction is found in the meta-MD simulations. A fully colored arrow is found at least 30 times. Only reactions found at least five times in the meta-MD simulations are included in the network. The MBH reaction found as part of the network is highlighted in green.
Figure S4: 20 molecules found in both the final populations from the GA searches in this work as well as in the work by Seumer et. al.⁵