doi:10.2533/chimia.2023.39 Chimia 77 (2023) 39–47 © S. Gallarati, P. van Gerwen, A. A. Schoepfer, R. Laplaza, C. Corminboeuf

Genetic Algorithms for the Discovery of Homogeneous Catalysts

Simone Gallarati[a], Puck van Gerwen[ab], Alexandre A. Schoepfer[ab], Ruben Laplaza[ab], and Clemence Corminboeuf*[ab]

Abstract: In this account, we discuss the use of genetic algorithms in the inverse design process of homogeneous catalysts for chemical transformations. We describe the main components of evolutionary experiments, specifically the nature of the fitness function to optimize, the library of molecular fragments from which potential catalysts are assembled, and the settings of the genetic algorithm itself. While not exhaustive, this review summarizes the key challenges and characteristics of our own (*i.e.*, NaviCatGA) and other GAs for the discovery of new catalysts.

Keywords: Catalysis · Discovery · Homogeneous · Machine learning

Simone Gallarati studied Materials Chemistry at the University of St Andrews with a year-long industrial placement at Diamond Light Source, UK. After a summer internship in the group of Prof S. E. Wheeler at UGA, he joined LCMD in 2019 as a PhD student. His research is focused on developing and extending the applicability of state-of-the-art computational methods to organocatalysis using data-driven tools and concepts.

Puck van Gerwen initially studied chemistry at the University of Edinburgh until 2017. She then obtained a masters in physics with a focus on computational methods at Imperial College London in 2019. In 2020, she joined LCMD as a PhD student. Her research focuses on building physics-based machine-learning models to study chemical reactions.

Alexandre A. Schoepfer obtained his bachelor and master degrees in chemistry from the University of Basel in 2019 and 2021 respectively. After an internship in the group of Prof. Michael Nash, he joined both groups of Prof. Jérôme Waser (LCSO) and Prof. Clémence Corminboeuf (LCMD) as a shared PhD student in 2021. His research focuses on developing models for organic synthesis method development, based on laboratory notebook data.

Ruben Laplaza obtained his bachelor degree in chemistry from the University of Zaragoza. In 2020, he completed his PhD in theoretical chemistry at Sorbonne Université, under the supervision of Profs. V. Polo and J. Contreras-Garcia. He has since been a postdoctoral researcher in LCMD. His current work involves using machine learning and automated modelling pipelines to investigate catalytic processes.

Clemence Corminboeuf started her independent career at the EPFL as an assistant professor and Sandoz Family Foundation Chair. She was promoted to associate (2014) and full professor (2019). She was awarded two European Research Council (ERC) grants (2012/2018), received the Werner Prize of the Swiss Chemical Society in 2014, the Theoretical Chemistry Award from the ACS Physical Chemistry Division in 2018. In 2021, she received the Heilbronner-Huckel Lecture Award from the Swiss and German Chemical Societies and the Per-Olov Löwdin (Uppsala) lecture in 2022. Her research on electronic structure theory exploits the interplay of deterministic and statistical approaches applied to the area of homogeneous catalysts and molecular organic materials.

1. Introduction

Homogeneous catalyst landscapes are built from a near-infinite array of plausible ligands, transition metals, functional groups, and substituents,[1–3] making their exploration by 'brute force' (*i.e.*, direct screening with high-throughput experiments or computations) often impractical.[4] Inverse design[5] offers an efficient alternative: given a desired target property, such as high catalyst turnover or product selectivity, a structure yielding the optimal value of that property is searched for. Most inverse design strategies are gradient-based,[6] either by defining gradients from first principles or implicitly learning them using neural networks (NNs).[7–11] In the former case, an initial structure is optimized following the derivative of the target with respect to changes in molecular structure. Alchemical gradients *i.e.*, following the property

*Correspondence: Prof. Dr. C. Corminboeuf, E-mail: clemence.corminboeuf@epfl.ch
[a]Laboratory for Computational Molecular Design, Institut des Sciences et Ingénierie Chimiques, École Polytechnique Fédérale de Lausanne, CH-1015 Lausanne, Switzerland; [b]National Center for Competence in Research-Catalysis (NCCR-Catalysis), École Polytechnique Fédérale de Lausanne, CH-1015 Lausanne, Switzerland

derivative with changes in nuclear charge distribution,[12–14] offer an alternative gradient-based optimization strategy. Deep generative models implicitly learn gradient functions for optimization instead.[7–11] Popular frameworks include variational autoencoders (VAEs),[15–20] recurrent neural networks (RNNs),[21] reinforcement learning (RL),[22–26] and generative adversarial networks (GANs).[27,28]

While generally convenient and efficient, gradient-based inverse design has some limitations: first, the global optimum is difficult to find; second, such property gradients may not be meaningful (*i.e.*, leading to non-physical molecules);[6,29,30] third, the structure–property space may not be continuous.[31,32] The latter point is particularly problematic in catalysis as small modifications (*e.g.*, ligand modification) may result in a sharp drop in performance for non-obvious reasons (activity cliffs).[33] Furthermore, fairly distinct catalysts may be equally efficient for a given reaction, which implies the existence of different optimal regions within the catalyst space. Derivative-free global optimizers circumvent these issues at the cost of an increased number of evaluations. Examples include particle-swarm, polytope, and evolutionary methods.[34] From the latter family, genetic algorithms (GAs)[35–39] stand out for their simplicity coupled with their superior performance in benchmarking studies.[7,39–42] Correspondingly, they are a cornerstone of *de novo* drug design and lead optimization campaigns, with numerous examples on the improvement of drug-like properties (*e.g.*, logP) of medium-sized organic molecules predicted directly from their 2D structure.[43–46] GAs have also found widespread use in other optimization problems in chemistry, ranging from the selection of training instances for molecular models of quantum chemical properties,[47] the identification of low-energy minima on complex potential energy surfaces,[48] structure elucidation,[49] the parametrization of force fields,[50] feature selection in multiple linear regression models,[51] and curve fitting.[52] Within the past few years, several applications of GAs to inverse material and catalyst design have been reported.[53] In 2012, Jensen, Alsber, and co-workers coupled an evolutionary algorithm with an automated molecular builder of olefin metathesis catalysts, scored against a ‘productivity’ fitness function obtained *via* QSAR.[54,55] This algorithm was later renamed *DenoptimGA* and included in the general-purpose software package DENOPTIM.[56] Despite being able to precisely control the metal-coordinating environment, the authors reported that many of the complexes generated by the GA contained undesirable functional groups or were otherwise synthetically inaccessible.[57]

More recently, Seumer and Jensen used a graph-based GA (GB-GA)[39] to discover organocatalysts for the Morita–Baylis–Hillman (MBH) reaction.[58] The optimization was not constrained to a user-defined library of fragments, but rather to the entire chemical space of tertiary amines (from the ZINC database), leading to the discovery of previously unseen catalytic motifs.

Aspuru-Guzik *et al.* have combined high-throughput virtual screening and genetic algorithms *e.g.*, JANUS,[38] for the systematic exploration of chemical space and the discovery of novel materials. They demonstrated that augmenting genetic algorithms with NNs helps increase the diversity of the generated molecules and avoids getting stuck in local minima.[59] Similarly, Kulik *et al*. used a GA coupled with an artificial NN to discover spin-crossover transition-metal (TM) complexes in a space of over 5600 compounds.[60] Using a modified fitness function, they were able to balance the exploration of new species with ML model confidence.

Our group introduced NaviCatGA,[61] a versatile genetic optimization pipeline, and showcased its broad applicability in homogeneous catalysis. NaviCatGA operates using both 2D and 3D catalyst representations with any suitable fitness function and is easily adaptable to various tasks. Here, we give an overview of its (and other GAs’) key components and functionalities, and discuss practical choices to use such algorithms and solve inverse design problems.

2. Genetic Algorithms

A GA performs derivative-free optimizations mirroring the mechanism of biological evolution. Each possible solution to the optimization problem is called a ‘chromosome’ composed of a number of ‘genes’. Fig. 1 shows an illustrative example in which the chromosome corresponds to the organocatalyst in the leftmost bubble, fragmented into three genes (red, blue, and green structural units). Each gene can take a number of values (*i.e.*, acceptable molecular fragments), which is fixed beforehand, so that all possible solutions may be enumerated combinatorially from the pool of possible values per gene and the number of genes in a chromosome. The evolutionary experiment is started with a limited number of chromosomes, which constitutes the initial ‘population’. A fitness value, or score, is computed for every chromosome in the population, which is then sorted by fitness (second bubble; Section 4 discusses how to quickly evaluate the fitness function). The fitness of a given chromosome should not be just the sum of fixed gene contributions, but rather be affected by the interplay between different genes. Otherwise, the optimization problem can be solved through simple sorting.

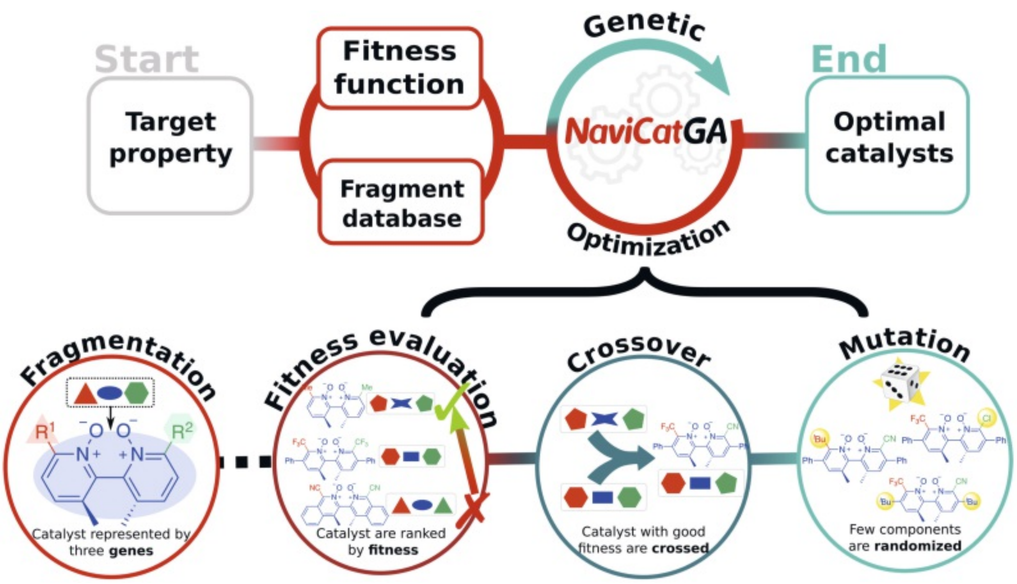

Fig. 1. (Top) Schematic catalyst optimization pipeline enabled by NaviCatGA and other genetic algorithms. (Bottom) Optimization loop mirroring the mechanism of biological evolution.

From the fitness-sorted population, a number of top-ranking chromosomes are selected and the rest discarded. The selected pool is 'crossed-over' by combining their genes to generate new chromosomes, replacing the discarded ones in the population (third bubble). Finally, a few genes are replaced randomly (rightmost bubble). A new population is produced after fitness evaluation, selection, cross-over, and mutation, and evaluated anew to continue the process iteratively. Each iteration is called a 'generation', and the process is stopped after a fixed number of generations has passed, a certain fitness value is reached, or the average fitness plateaus.

An advantage of evolutionary experiments over 'direct' screening is that the genetic optimization process provides insight on the structural modifications that prove to be beneficial for catalyst performance.[61] The appearance or disappearance of certain functional groups or chemical moieties over many generations may be associated with sudden changes in catalyst fitness. Thus, evolutionary experiments help rationalize design principles for further exploration.

2.1 *What is a Chromosome?*

In the context of inverse homogeneous catalyst design, chromosomes are molecules *e.g.*, transition-metal species, ligands for organometallic complexes, or organocatalysts, represented by character strings,[38,62,63] 2D graphs,[39,64–66] or 3D coordinates.[54,56,67,68] The infrastructure of the evolutionary experiments can be adapted depending on how the chromosomes are represented.[61] Genes are fragments of the chosen representation *e.g.*, characters to be concatenated into a larger SMILES string. All possible arrangements of genes must lead to a valid chromosome for which a fitness value may be computed. Furthermore, the mapping from the genetic composition of a chromosome to its fitness should be injective. For this reason, representing molecules with SELFIES[69] is becoming increasingly popular, as they guarantee validity and uniqueness with respect to permutations of their composing characters.

Choosing how to represent a chromosome depends on the availability of suitable fragment libraries and how easy it is to evaluate the fitness function using that representation. SMILES and SELFIES are often preferred given the abundance of string-based chemoinformatics tools[38,69,70] (which also help estimate synthetic accessibility),[71,72] and the possibility of using alphabets of characters as fragments library. Alternatively, 3D-based GAs rely on manipulating libraries of (Cartesian) coordinates of fragments, but the molecular conformation can be controlled precisely.

2.2 *Designing the Evolutionary Experiment*

Several design choices determine the efficiency of an evolutionary experiment. Depending on the problem, the desired output, and the molecular representation used for chromosomes and genes, different settings may be beneficial. We will review some of the key ones.

Selection strategies control the greediness of the optimization. The most common and straightforward selection rule is to simply preserve the top *N* candidates, which will lead to a thorough exploration of the chemical subspace around them, possibly neglecting other regions. An alternative strategy is the roulette wheel method, which randomizes selection while assigning higher probabilities based on fitness, leading to better exploration of areas corresponding to slightly suboptimal candidates.[58] JANUS[38] combines exploration and exploitation by using two selection strategies in a parallel tempering scheme, with one focusing on the target fitness function (for exploration of the global landscape) and one maximizing similarity to current candidates (for local optima exploitation).

The existence of duplicate chromosomes must also be considered when choosing a selection strategy. At any intermediate stage of the genetic run, a given fit individual may be repeatedly found in the population, and crossover of identical individuals is ineffective. However, having a top-ranking individual partake in several different crossovers and propagating its advantageous genes is beneficial. As a rule of thumb, duplicates are increasingly problematic if a deterministic selection strategy is used, especially in settings with small population sizes and a small number of genes per chromosome. The DENOPTIM package[56] automatically prunes duplicate molecules.

Properly defining the crossover operation is the most important step of an evolutionary experiment. Conceptually, crossover is the driving force that steers the optimization towards the best candidates, under the assumption that combinations of good genes may be even better. Therefore, the effectiveness of the crossover operator is critical. Many possibilities that ultimately rely on the structure of chromosomes exist. In the most general case, crossover is achieved by replacing some parts of a parent chromosome with some formally equivalent parts from another parent chromosome. In string-based algorithms where genes are character tokens (*e.g.*, SMILES, SELFIES) this is achieved by splicing two parent strings in two or more fragments, and reshuffling the resulting parts into offspring strings. Graph and 3D structure-based chromosomes require an explicit definition of equivalence (*i.e.*, which fragment can substitute which) to avoid combinations that lead to invalid molecules: this is achieved based on valence rules or belonging to the same fragment subset. Elaborate algorithms for handling 3D fragments as genes, while maintaining molecular validity, have been developed by Jensen and co-workers and combined in the DENOPTIM software.[56,57,67,68]

Finally, it is possible to steer the optimization towards interesting chemical subspaces by starting the run from a population drawn from a selected region of space or from copies of a reference catalyst. The STONED algorithm[62] proposed by Aspuru-Guzik *et al.* is a convenient way of generating pools of similar molecules based on a given SMILES string. In the same spirit, the initialization and the composition of the fragment libraries can be used to enforce the chemical validity and the synthetic accessibility of the resulting species, which is typically a problem of uncontrolled generative models.[71,72]

2.3 *NaviCatGA*

In NaviCatGA, chromosomes are assembled from the corresponding genes using any suitable molecular representation, including SMILES and SELFIES strings and XYZ coordinates through the corresponding child classes (SmilesGenAlgSolver, SelfiesGenAlgSolver, and XYZGenAlgSolver using AaronTools.py geometry objects[73]). The child classes define the data type of the genes and contain all the possible values any gene can take, called an 'alphabet'. Genes with the same alphabet are considered equivalent. Depending on the user's needs, new child solver classes can be easily defined, as the core shared functionalities are kept separately in the base solver class (the core genetic loop, which is data-type agnostic). Different data structures, supported by other libraries (*e.g.*, Molassembler[74] or molSimplify[75]) could be used as alternative back-ends.

Five different selection strategies are provided, including two-by-two, pairwise tournament, random, and roulette wheel method. The latter is the default, however temperature-based schemes (*i.e.*, Boltzmann-weighted) in which the selection becomes increasingly greedy as the experiment progresses are also available. In case of duplicate chromosomes, NaviCatGA supports both options of pruning identical catalyst candidates and preserving them to partake in the crossover of their advantageous genes. It is also possible to lock specific genes, so that they remain unchanged during the optimization procedure. Additionally, our package also implements the STONED[62] algorithm to initialize the search process from a specific neighborhood of the chemical space.

3. Structures and Fragments Libraries

The total combinatorial space explored during the evolutionary experiments is determined by the extent of the database of catalyst components and the scheme chosen to fragment them into genes (*i.e.*, building blocks). Its size typically ranges from 10^4 to 10^6 candidates.[58] Of course, the efficiency of GAs lies in potentially finding the best combinations of fragments in relatively few iterations, rather than evaluating the entire library. Two approaches have been used (Fig. 2): screening user-defined libraries of fragments,[61] or subsets of larger regions of chemical space.[58] In the former, a fragmentation scheme is first defined based on structural patterns observed in a smaller pool of catalysts; fragments are then listed manually and used to build candidates on-the-fly during the evolutionary experiments. The latter relies on having access to bigger databases of compounds, with the molecular sites for mutation and crossover being identified afterwards.[58] Below, we report some of the most popular databases that can be used for this second 'top-down' approach. We then discuss fragment-based strategies enabling the first 'bottom-up' approach.

3.1 *Subsets of Chemical Space*

Pioneering work in the mapping of ligand spaces for organometallic catalysis has been conducted by Fey and *et al*. with the ligand knowledge bases (LKB).[76–81] These include *ca*. 1.3 k mono- and bidentate P, C, N, O, P,P and P,N ligands with associated steric and electronic descriptors. They have showcased the use of Principal Component score plots as 'maps' of ligand space,[82] facilitating the observation of reactivity trends and the optimization of reaction properties. Building on this, Sigman, Aspuru-Guzik and co-workers recently introduced the *kraken* platform for monodentate organophosphorus(III) compounds.[83] Using ~1.5 k ligands from the literature, they built combinatorial libraries of over 300 k species and trained ML models to predict their conformationally relevant physicochemical descriptors. Combined with high-throughput experimentation, *kraken* has now been used to find optimal ligands for several reactions.[84–88]

The Kulik group has conducted extensive work on the exploration of transition-metal space and the development and validation of computational tools to accurately predict its properties.[89,90] Among the datasets curated for these tasks, the octahedral homoleptic ligand database[91] (OHLDB: 11,325 theoretical ligands, *ca*. 700 complexes fully characterized with DFT) stands out for the enumerative strategy behind its construction and its coverage of previously unexplored regions of chemical space (only 71 ligands were previously included in common organic molecule libraries). Extending this approach to fused five- and six-membered rings, they generated 2.8M homoleptic complexes for multi-objective redox flow battery design.[92] Kulik *et al*. showed that enriching ML models with training data from the smaller datasets (*e.g.*, OHLDB) improved ML performance on larger TM-complexes from experimental sets,[93] like the Cambridge Structural Database (CSD).[94,95] This platform likely hosts the most diverse collection of synthesizable molecules. Recently, our group developed the *cell2mol* software[96] to characterize molecular crystals from CSD and retrieve the connectivity, charge, and oxidation state information. *cell2mol* enables the construction of quantum chemistry-ready datasets, such as a library of 31 k TM-complexes and 13 k ligands with incomparable chemical diversity.[96]

The space of 'small' organic molecules *i.e.*, closed-shell up to 10–20 heavy atoms, has been extensively mapped.[97,98] The generated database GDB-17 lists *ca*. 166.4 billion molecules of up to 17 C, N, O, S, and halogen atoms from systematic enumeration following simple rules of chemical stability and synthetic feasibility.[99–102] Other popular datasets for drug discovery are

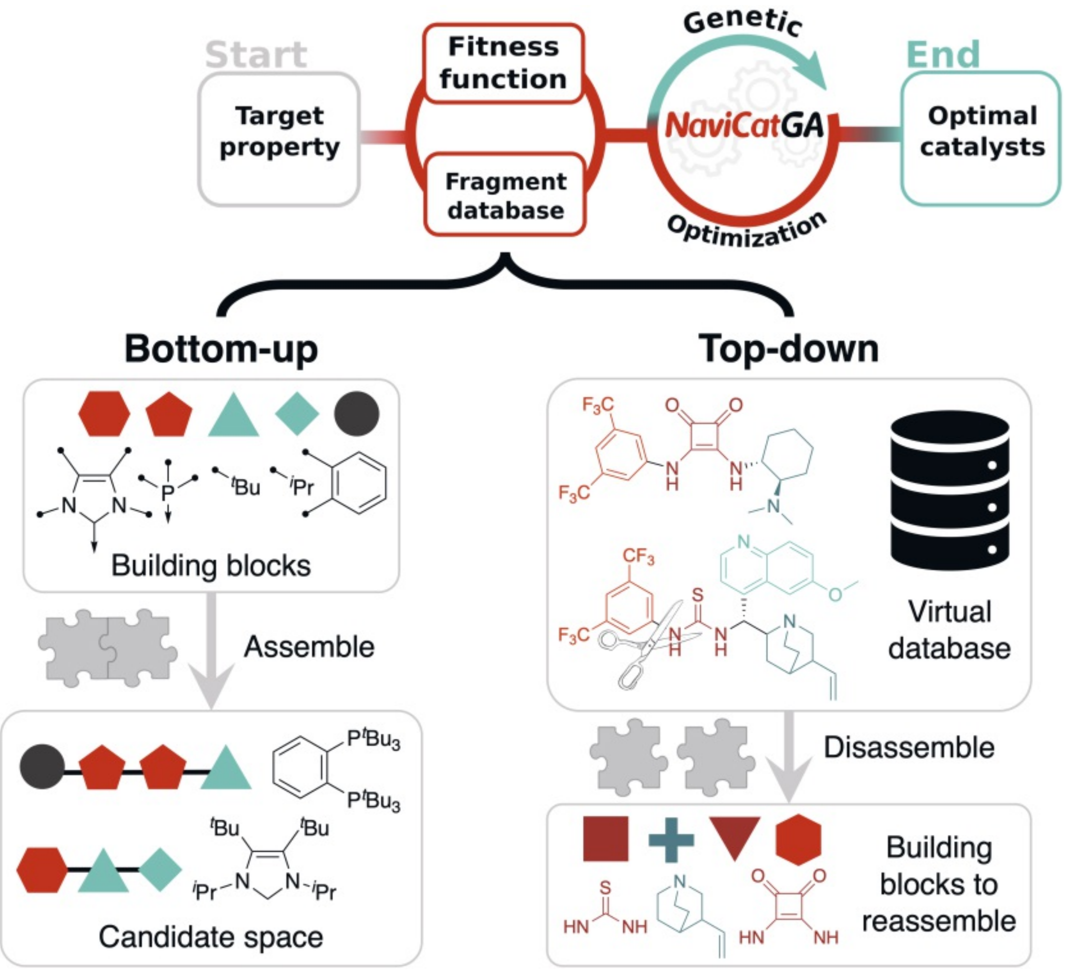

Fig. 2. Bottom-up and top-down strategies to define the total combinatorial space explored during genetic optimization and the molecular fragments used to assemble catalyst candidates.

ChEMBL (>1.6M structures with 14M activity values),[103] ZINC (886M),[104] PubChem50 (101M molecules up to 50 atoms),[105] COCONUT (402k natural products),[106] and DrugBank (>2k FDA-approved drugs).[107] Going beyond druglike molecules, the COMPAS Project was recently curated, consisting of ~34k *cata*-condensed polybenzenoid hydrocarbons, with the aim of enabling the data-driven development of improved organic electronic materials.[108]

3.2 *Fragment-based Strategies*

While the space of organic molecules is continuously being enumerated, its subset of organocatalysts is far less frequently explored. To obviate the lack of data-driven tools that facilitate the exploration of wider regions of organocatalyst space, we introduced OSCAR.[109] This repository of organic molecules leverages the modularity of organocatalysts, offering a route to build datasets up to 1M structures. This was achieved by automatically mining species containing pre-defined function-based fragments from the literature and CSD and re-assembling the building blocks in a combinatorial fashion. The fragment-based nature of organocatalysts was further exploited in combination with activity maps and statistical modelling to suggest structural modifications for activity enhancement.[110]

Transition-metal catalysts are more frequently viewed in a modular fashion as a combination of active metal center and ligands, which are further decomposed into metal-coordinating groups, backbone/bridging units, and substituents. Indeed, pioneering studies on the automated generation of TM-species from fragments relied on this tailored feature.[81,111] Subsequently, Jensen *et al*. introduced the concept of 'organometallic fragment space' as the combination of annotated molecular fragments and connection rules for organometallic species.[57] They harvested building blocks from CSD and used them to automatically generate synthetically accessible TM-complexes with both 2D and 3D representations.[57,67]

3.3 *NaviCatGA*

NaviCatGA combines fragments with the utmost flexibility through a user-defined assembler function. The assembler function takes a given individual (a list of genes of the specified datatype) and assembles it into a potential catalyst. If genes are represented as SMILES strings, assembly is as simple as concatenating their characters, otherwise 3D geometry objects[73] may be handled as well through the XYZGenAlgSolver child class. Depending on the specific optimization problem, any assembler function is supported, enabling the generation of more complex graph structures from the corresponding chromosomes.

4. The Fitness Function

The role of the fitness function is to evaluate how close a catalyst candidate is to achieving optimal performance, which is often exemplified in terms of activity and/or stereo/regio/chemoselectivity. Measures of activity and selectivity can be obtained either from experiments or computations. The product yield, turnover number (TON), and turnover frequency (TOF) are the most commonly used experimental quantities that describe activity.[112] Selectivity is generally reported as product ratio (enantiomeric/diastereomeric ratio) or converted to $\Delta\Delta G^{\ddagger}$ values according to Transition State Theory. Quantum chemical computations are frequently used to estimate both $\Delta\Delta G^{\ddagger}$ and TOF,[113,114] however, predicting the TOF of large libraries of catalysts is highly costly. Linear free energy scaling relationships and volcano plots allow the TOF to be estimated solely from a descriptor variable, such as the relative energy of a catalytic cycle intermediate, which can then be used directly as fitness.[115,116]

Beyond activity and selectivity, the overall performance of a catalyst must satisfy a multitude of other objectives, such as stability, solubility, synthesizability, toxicity, and cost. Therefore, catalyst optimization is a multiobjective problem, where improving an individual requirement often results in the deterioration of another.[86,117] Finding solutions in the Pareto front requires scaling the fitness function appropriately.[118] The relative weights of each requirement may be chosen manually (as done by Jensen with the activation energy of the MBH reaction rate-determining step and a synthetic accessibility measure)[58] or using a scalarizer like Chimera.[119]

Evaluating the fitness function is the bottleneck of evolutionary experiments. Obtaining activity/selectivity measures experimentally is time- and resource-intensive, and only tractable in closed-loop optimizations with robotized HTE methods and self-driving laboratories.[120,121] DFT-based catalyst performance predictions often require computing the complete free energy profile associated with a catalytic cycle, which becomes expensive for more than a handful of systems.[122] To accelerate this process, statistical models are used to predict the candidates' fitness. They may be trained using either experimental or computational data. The first approach is often limited by the small size of the experimental datasets available and by their inherent noise,[123] while the second suffers from the difficulties associated with reproducing difficult-to-compute targets *e.g.*, *e.e.* values.[124] Regardless of the nature of the target, the fast evaluation of the fitness function involves two aspects: the representation used to encode a catalyst's structure/composition, and the statistical model used to make predictions (linear, multilinear, or nonlinear). These two aspects are depicted in Fig. 3 and described in the following sections.

4.1 *Representing Catalyst Candidates*

Popular topological descriptors such as Morgan fingerprints[125] (also known as extended connectivity fingerprints, ECFPs)[126] encode a given compound using its 2D structure. This family of 'hashed' fingerprints[127,128] enumerate through the molecule to identify chemical substructures up to a certain radius or number of bonds from central atoms. The location and counts of substructures are converted to feature vectors using a hash function. These fingerprints, particularly the ECPFs, have been widely used for catalytic properties predictions,[129–131] also in combination with evolutionary experiments.[58] Part of their popularity stems from their accessibility: to generate them, only the SMILES string of a molecule is needed.

The field of physics-based ML assumes that molecules should be represented as 3D objects instead.[132–134] The core principle is that the representation replaces the role of the Hamiltonian in the Schrödinger equation, and should therefore require the same information as input (for neutral molecules, atom types and 3D coordinates). Physics-based representations describe interactions between atomic environments in a molecule, typically using either non-linear potentials inspired from the early days of molecular dynamics,[135–139] atom-centered continuous basis functions,[140–144] or cheap estimates of quantum-chemical objects.[145–148] Invariances with respect to molecular symmetry are naturally incorporated, and modified representations exist to handle equivariance.[149] An additional level of complexity can be incorporated in the representation by considering the fact that molecules at finite temperature do not exist in a single conformation. Conformational ensembles may be represented, for example, using Boltzmann-weighted physics-based representations of many conformers[150] or features describing the conformer-averaged occupancy over a 3D grid.[151] Our group recently illustrated the relevance of 'reaction-inspired' representations, whereby considering structural changes from reactant(s) to product(s) is particularly effective for predicting both thermodynamic and kinetic reaction properties.[152,153]

An alternative way of representing molecules or reactions, which involves fewer and generally more intuitive features, is by

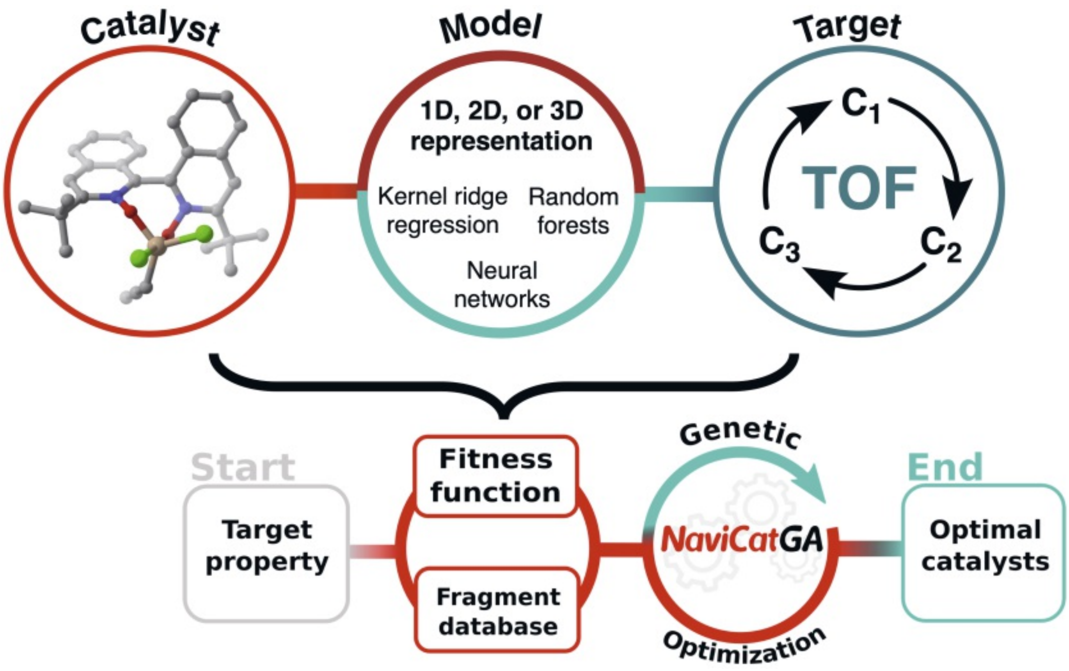

Fig. 3. Components needed to quickly and affordably evaluate the fitness of a catalyst candidate, namely a way of representing its structure (in 1D, 2D, or 3D format) and a statistical model that predicts a target property, such as the turnover frequency, or the reaction selectivity.

their properties.[154] In the early days, these descriptors were derived largely from experiments *e.g.*, the Hammett substituent constants;[155] in recent decades, stereoelectronic parameters obtained *via* quantum chemical computations of low-energy structures or ensembles of conformers have become popular.[156]

4.2 *Statistical Models for Fitness Evaluation*

The choice of the ML model used in combination with molecular fingerprints varies from support vector machines (SVM),[129,157] random forests (RF),[130,158–160] to neural networks.[161–164] Typically, physics-based representations are used in combination with kernel ridge regression (KRR) models, which rely on a (dis) similarity metric between molecules.[165] Such machine learning models have been successfully trained to predict catalytic properties.[166] Models combining electronic and steric descriptors with interpretable multivariate linear regression analysis (MLR) have been extensively developed for reaction outcome predictions, especially $\Delta\Delta G^{\ddagger}$.[167–169] Yet, such models, which depend on DFT computations of relatively expensive properties (*e.g.,* vibrational frequencies and intensities, polarizabilities)[170] are not adapted to the purpose of fast (GA) optimization for which bypassing the DFT bottleneck is key.

4.3 *NaviCatGA*

The choice of fitness function depends on the specific application. NaviCatGA favors fitness functions that map a candidate catalyst's chemical structure to a measure of its performance in a given reaction. In this sense, molecular volcano plots[115] are ideally suited as they provide a way of connecting the descriptor variable, typically the energy change associated with a step of the reaction mechanism (x-axis), to the overall catalytic performance (y-axis, expressed in terms of energy span or TOF). This inexpensive mapping between structure and reactivity constitutes a natural fitness function to be exploited in close-loop optimizations. The descriptor variable is easily evaluated using quantum chemical computations or predicted using ML models.[122,166] Alternatively, MLR expressions for activity (*i.e.*, the volcano descriptor) and selectivity ($\Delta\Delta G^{\ddagger}$) with inexpensive steric and electronic parameters have also been used within NaviCatGA. Our package imposes no constraints on the form of the fitness function and any alternative defined by the user is possible.

5. Conclusions and Outlook

Genetic algorithms are a suitable strategy to explore large and complex homogeneous catalyst landscapes and find good candidates within a few iterations in the absence of analytical gradients. To ensure their successful application to inverse catalyst design tasks, three aspects have to be carefully considered: (1) the optimization problem must be robustly defined, especially when multiple catalytic properties should be improved simultaneously; (2) if surrogate statistical models are used to predict a candidate's fitness, they must be fast and affordable, including the representation used as input; (3) the nature of the search space must be thoroughly addressed, particularly if candidates are built from fragment libraries. Currently, a lot of effort is placed in improving the last two aspects, but not often in the context of evolutionary experiments, leading to suboptimal performance when they are coupled to genetic algorithms. With NaviCatGA, we aim at considering these three aspects simultaneously for their successful implementation in a closed-loop optimization pipeline.

Regarding the size and diversity of the search space, generating candidates from user-defined libraries of molecular fragments may introduce a bias in the experiment and limit the discovery of entirely new chemical motifs. A possible workaround is to include more diverse fragments in the database, however the statistical model may not extrapolate well beyond the training set, meaning that selecting diverse compounds should be done before the model is trained. Alternatively, active learning approaches, like Bayesian optimization,[171,172] enable a model to adapt as it navigates the search space, balancing the exploration of areas of high uncertainty with the exploitation of available data.

While implementing genetic algorithms and other generative models has become more routine, developing affordable models to accurately and quickly predict complex catalytic properties still remains a challenge. This is partially due to the scarcity of large experimental datasets in machine-readable formats, and to the difficulties associated with generating reliable reactivity data with *ab initio* methods, a topic of active research in our group.[124,173]

Finally, improvements in the decision-making protocol in multiobjective scenarios will be beneficial to the catalyst design process. One approach would be to use complex fitness functions that take into account more than just the reaction yield and selectivity, such as the previously reported Asymmetric Catalyst Efficiency (ACE) metric.[174] Alternatively, the definition of performance, rather than being fixed prior to the optimization experiment, can be dynamic and able to respond to new knowledge generated on-the-fly, such as the unforeseen stability and reactivity of novel compounds.[117] These are some of the foreseen directions for future applications of NaviCatGA.

Acknowledgements

The work highlighted in this article has been supported by the EPFL, the National Center of Competence in Research (NCCR) "Sustainable chemical process through catalysis (Catalysis)" of the Swiss National Science Foundation (SNSF, grant number 180544) and the European Research Council (ERC, Grant Agreement No. 817977) within the framework of the European Union's H2020.

Received: November 9, 2022

[1] P. Kirkpatrick, C. Ellis, *Nature* **2004**, *432*, 823, https://doi.org/10.1038/432823a.
[2] A. Mullard, *Nature* **2017**, *549*, 445, https://doi.org/10.1038/549445a.
[3] C. W. Coley, *Trends Chem*. **2021**, *3*, 133, https://doi.org/10.1016/j.trechm.2020.11.004.
[4] G. dos Passos Gomes, R. Pollice, A. Aspuru-Guzik, *Trends Chem*. **2021**, *3*, 96, https://doi.org/10.1016/j.trechm.2020.12.006.
[5] J. Noh, J. Kim, H. S. Stein, B. Sanchez-Lengeling, J. M. Gregoire, A. Aspuru-Guzik, Y. Jung, *Matter* **2019**, *1*, 1370, https://doi.org/10.1016/j.matt.2019.08.017.
[6] J. G. Freeze, H. R. Kelly, V. S. Batista, *Chem. Rev*. **2019**, *119*, 6595, https://doi.org/10.1021/acs.chemrev.8b00759.
[7] A. Nigam, R. Pollice, G. Tom, K. Jorner, L. A. Thiede, A. Kundaje, A. Aspuru-Guzik, *arXiv:2209.12487* **2022**, https://doi.org/10.48550/ARXIV.2209.12487.
[8] B. Sanchez-Lengeling, A. Aspuru-Guzik, *Science* **2018**, *361*, 360, https://doi.org/10.1126/science.aat2663.
[9] R. Pollice, G. dos Passos Gomes, M. Aldeghi, R. J. Hickman, M. Krenn, C. Lavigne, M. Lindner-D'Addario, A. Nigam, C. T. Ser, Z. Yao, A. Aspuru-Guzik, *Acc. Chem. Res*. **2021**, *54*, 849, https://doi.org/10.1021/acs.accounts.0c00785.
[10] D. Schwalbe-Koda, R. Gómez-Bombarelli, 'Generative Models for Automatic Chemical Design', **2020**, https://doi.org/10.1007/978-3-030-40245-7_21.
[11] O. Schilter, F. Zipoli, A. C. Vaucher, P. Schwaller, T. Laino, 'Deep learning assisted Suzuki cross coupling catalyst design', ACS Fall, **2022**.
[12] O. A. von Lilienfeld, R. D. Lins, U. Rothlisberger, *Phys. Rev. Lett*. **2005**, *95*, 153002, https://doi.org/10.1103/PhysRevLett.95.153002.
[13] O. A. von Lilienfeld, M. E. Tuckerman, *J. Chem. Phys*. **2006**, *125*, 154104, https://doi.org/10.1063/1.2338537.
[14] O. A. von Lilienfeld, *J. Chem. Phys*. **2009**, *131*, 164102, https://doi.org/10.1063/1.3249969.
[15] D. P. Kingma, M. Welling, *arXiv:1312.6114* **2013**, https://doi.org/10.48550/ARXIV.1312.6114.
[16] D. K. Duvenaud, D. Maclaurin, J. Iparraguirre, R. Bombarell, T. Hirzel, A. Aspuru-Guzik, R. P. Adams, in 'Advances in Neural Information Processing Systems', Vol. 28, Eds. C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, R. Garnett, Curran Associates, Inc., **2015**.
[17] R. Gómez-Bombarelli, J. N. Wei, D. Duvenaud, J. M. Hernández-Lobato, B. Sánchez-Lengeling, D. Sheberla, J. Aguilera-Iparraguirre, T. D. Hirzel, R. P. Adams, A. Aspuru-Guzik, *ACS Cent. Sci*. **2018**, *4*, 268, https://doi.org/10.1021/acscentsci.7b00572.
[18] R. Winter, F. Montanari, A. Steffen, H. Briem, F. Noé, D.-A. Clevert, *Chem. Sci*. **2019**, *10*, 8016, https://doi.org/10.1039/C9SC01928F.
[19] R.-R. Griffiths, J. M. Hernández-Lobato, *Chem. Sci*. **2020**, *11*, 577, https://doi.org/10.1039/C9SC04026A.
[20] R. Tempke, T. Musho, *Commun. Chem*. **2022**, *5*, 40, https://doi.org/10.1038/s42004-022-00647-x.
[21] A. Graves, *arXiv:1308.0850* **2013**, https://doi.org/10.48550/ARXIV.1308.0850.
[22] V. Mnih, K. Kavukcuoglu, D. Silver, A. Graves, I. Antonoglou, D. Wierstra, M. Riedmiller, *arXiv:1312.5602* **2013**, https://doi.org/10.48550/ARXIV.1312.5602.
[23] D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. van den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot, S. Dieleman, D. Grewe, J. Nham, N. Kalchbrenner, I. Sutskever, T. Lillicrap, M. Leach, K. Kavukcuoglu, T. Graepel, D. Hassabis, *Nature* **2016**, *529*, 484, https://doi.org/10.1038/nature16961.
[24] X. Yang, J. Zhang, K. Yoshizoe, K. Terayama, K. Tsuda, *Sci. Technol. Adv. Mater*. **2017**, *18*, 972, https://doi.org/10.1080/14686996.2017.1401424.
[25] M. Olivecrona, T. Blaschke, O. Engkvist, H. Chen, *J. Cheminform*. **2017**, *9*, 48, https://doi.org/10.1186/s13321-017-0235-x.
[26] M. Popova, O. Isayev, A. Tropsha, *Sci. Adv*. **2018**, *4*, eaap7885, https://doi.org/10.1126/sciadv.aap7885.
[27] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, *arXiv:1406.2661* **2014**, https://doi.org/10.48550/ARXIV.1406.2661.
[28] G. L. Guimaraes, B. Sanchez-Lengeling, C. Outeiral, P. L. C. Farias, A. Aspuru-Guzik, *arXiv:1705.1084* **2017**, https://doi.org/10.48550/ARXIV.1705.10843.
[29] G. Domenichini, O. A. von Lilienfeld, *J. Chem. Phys*. **2022**, *156*, 184801, https://doi.org/10.1063/5.0085817.
[30] O. A. von Lilienfeld, *Int. J. Quantum Chem*. **2013**, *113*, 1676, https://doi.org/10.1002/qua.24375.
[31] M. Aldeghi, D. E. Graff, N. Frey, J. A. Morrone, E. O. Pyzer-Knapp, K. E. Jordan, C. W. Coley, *J. Chem. Inf. Model*. **2022**, *62*, 4660, https://doi.org/10.1021/acs.jcim.2c00903.
[32] X. Y. See, X. Wen, T. A. Wheeler, C. K. Klein, J. D. Goodpaster, B. R. Reiner, I. A. Tonks, *ACS Catal*. **2020**, *10*, 13504, https://doi.org/10.1021/acscatal.0c03939.
[33] S. H. Newman-Stonebraker, S. R. Smith, J. E. Borowski, E. Peters, T. Gensch, H. C. Johnson, M. S. Sigman, A. G. Doyle, *Science* **2021**, *374*, 301, https://doi.org/10.1126/science.abj4213.
[34] L. M. Rios, N. V. Sahinidis, *J. Glob. Optim*. **2013**, *56*, 1247, https://doi.org/10.1007/s10898-012-9951-y.
[35] R. Leardi, *J. Chemometrics* **2001**, *15*, 559, https://doi.org/10.1002/cem.651.
[36] H. A. Bashir, R. S. Neville, in '2014 IEEE International Conference on Systems, Man, and Cybernetics (SMC)', **2014**, pp. 100, https://doi.org/10.1109/SMC.2014.6973891.
[37] G. Schneider, U. Fechner, *Nat. Rev. Drug Discov*. **2005**, *4*, 649, https://doi.org/10.1038/nrd1799.
[38] A. Nigam, R. Pollice, A. Aspuru-Guzik, *Digital Discovery* **2022**, *1*, 390, https://doi.org/10.1039/D2DD00003B.
[39] J. H. Jensen, *Chem. Sci*. **2019**, *10*, 3567, https://doi.org/10.1039/C8SC05372C.
[40] N. Brown, M. Fiscato, M. H. S. Segler, A. C. Vaucher, *J. Chem. Inf. Model*. **2019**, *59*, 1096, https://doi.org/10.1021/acs.jcim.8b00839.
[41] W. Gao, T. Fu, J. Sun, C. W. Coley, *arXiv:2206.12411* **2022**, https://doi.org/10.48550/ARXIV.2206.12411.
[42] E. S. Henault, M. H. Rasmussen, J. H. Jensen, *PeerJ Phys. Chem*. **2020**, *2*, e11, https://doi.org/10.7717/peerj-pchem.11.
[43] T. Slater, D. Timms, *J. Mol. Graph*. **1993**, *11*, 248, https://doi.org/10.1016/0263-7855(93)80005-C.
[44] D. R. Westhead, D. E. Clark, D. Frenkel, J. Li, C. W. Murray, B. Robson, B. Waszkowycz, *J. Comput.-Aided Mol. Des*. **1995**, *9*, 139, https://doi.org/10.1007/BF00124404.
[45] R. C. Glen, A. W. R. Payne, *J. Comput.-Aided Mol. Des*. **1995**, *9*, 181, https://doi.org/10.1007/BF00124408.
[46] C. Steinmann, J. H. Jensen, *PeerJ Phys. Chem*. **2021**, *3*, e18, https://doi.org/10.7717/peerj-pchem.18.
[47] N. J. Browning, R. Ramakrishnan, O. A. von Lilienfeld, U. Roethlisberger, *J. Phys. Chem. Lett*. **2017**, *8*, 1351, https://doi.org/10.1021/acs.jpclett.7b00038.
[48] J. Villard, M. Kılıç, U. Rothlisberger, *ChemRxiv* **2022**, https://doi.org/10.26434/chemrxiv-2022-cq1qm.
[49] A. H. C. van Kampen, L. M. C. Buydens, *Chemom. Intell. Lab. Syst*. **1997**, *36*, 141, https://doi.org/10.1016/S0169-7439(97)00016-6.
[50] J. Hunger, G. Huttner, *J. Comput. Chem*. **1999**, *20*, 455, https://doi.org/10.1002/(SICI)1096-987X(199903)20:4<455::AID-JCC6>3.0.CO;2-1.
[51] J. Trejos, M. A. Villalobos-Arias, J. L. Espinoza, in 'Handbook of Research on Modern Optimization Algorithms and Applications in Engineering and Economics', Eds. P. Vasant, G.-W. Weber, V. N. Dieu, IGI Global, Hershey, PA, USA, **2016**, pp. 133, https://doi.org/10.4018/978-1-4666-9644-0.ch005.
[52] A. P. De Weijer, C. B. Lucasius, L. Buydens, G. Kateman, H. M. Heuvel, H. Mannee, *Anal. Chem*. **1994**, *66*, 23, https://doi.org/10.1021/ac00073a006.
[53] J. E. Kreutz, A. Shukhaev, W. Du, S. Druskin, O. Daugulis, R. F. Ismagilov, *J. Am. Chem. Soc*. **2010**, *132*, 3128, https://doi.org/10.1021/ja909853x.
[54] Y. Chu, W. Heyndrickx, G. Occhipinti, V. R. Jensen, B. K. Alsberg, *J. Am. Chem. Soc*. **2012**, *134*, 8885, https://doi.org/10.1021/ja300865u.
[55] G. Occhipinti, H.-R. Bjørsvik, V. R. Jensen, *J. Am. Chem. Soc*. **2006**, *128*, 6952, https://doi.org/10.1021/ja060832i.
[56] M. Foscato, V. Venkatraman, V. R. Jensen, *J. Chem. Inf. Model*. **2019**, *59*, 4077, https://doi.org/10.1021/acs.jcim.9b00516.
[57] M. Foscato, G. Occhipinti, V. Venkatraman, B. K. Alsberg, V. R. Jensen, *J. Chem. Inf. Model*. **2014**, *54*, 767, https://doi.org/10.1021/ci4007497.
[58] J. Seumer, J. H. Jensen, *ChemRxiv* **2022**, https://doi.org/10.26434/chemrxiv-2022-ngwvt.
[59] A. Nigam, P. Friederich, M. Krenn, A. Aspuru-Guzik, *arXiv:1909.11655* **2020**, https://doi.org/10.48550/arXiv.1909.11655.
[60] J. P. Janet, L. Chan, H. J. Kulik, *J. Phys. Chem. Lett*. **2018**, *9*, 1064, https://doi.org/10.1021/acs.jpclett.8b00170.

[61] R. Laplaza, S. Gallarati, C. Corminboeuf, *Chem. Methods* **2022**, e202100107, https://doi.org/10.1002/cmtd.202100107.

[62] A. Nigam, R. Pollice, M. Krenn, G. dos P. Gomes, A. Aspuru-Guzik, *Chem. Sci*. **2021**, *12*, 7079, https://doi.org/10.1039/D1SC00231G.

[63] Y. Kwon, J. Lee, *J. Cheminform.* **2021**, *13*, 24, https://doi.org/10.1186/s13321-021-00501-7.

[64] N. Brown, B. McKay, F. Gilardoni, J. Gasteiger, *J. Chem. Inf. Comput. Sci*. **2004**, *44*, 1079, https://doi.org/10.1021/ci034290p.

[65] J. Verhellen, *Chem. Sci*. **2022**, *13*, 7526, https://doi.org/10.1039/D2SC00821A.

[66] J. Verhellen, J. Van den Abeele, *Chem. Sci*. **2020**, *11*, 11485, https://doi.org/10.1039/D0SC03544K.

[67] M. Foscato, V. Venkatraman, G. Occhipinti, B. K. Alsberg, V. R. Jensen, *J. Chem. Inf. Model*. **2014**, *54*, 1919, https://doi.org/10.1021/ci5003153.

[68] M. Foscato, B. J. Houghton, G. Occhipinti, R. J. Deeth, V. R. Jensen, *J. Chem. Inf. Model.* **2015**, *55*, 1844, https://doi.org/10.1021/acs.jcim.5b00424.

[69] M. Krenn, F. Häse, A. Nigam, P. Friederich, A. Aspuru-Guzik, *Mach. Learn.: Sci. Technol*. **2020**, *1*, 045024, https://doi.org/10.1088/2632-2153/aba947.

[70] M. Krenn, Q. Ai, S. Barthel, N. Carson, A. Frei, N. C. Frey, P. Friederich, T. Gaudin, A. A. Gayle, K. M. Jablonka, R. F. Lameiro, D. Lemm, A. Lo, S. M. Moosavi, J. M. Nápoles-Duarte, A. Nigam, R. Pollice, K. Rajan, U. Schatzschneider, P. Schwaller, M. Skreta, B. Smit, F. Strieth-Kalthoff, C. Sun, G. Tom, G. Falk von Rudorff, A. Wang, A. D. White, A. Young, R. Yu, A. Aspuru-Guzik, *Patterns* **2022**, *3*, 100588, https://doi.org/10.1016/j.patter.2022.100588.

[71] C. W. Coley, L. Rogers, W. H. Green, K. F. Jensen, *J. Chem. Inf. Model*. **2018**, *58*, 252, https://doi.org/10.1021/acs.jcim.7b00622.

[72] W. Gao, C. W. Coley, *J. Chem. Inf. Model*. **2020**, *60*, 5714, https://doi.org/10.1021/acs.jcim.0c00174.

[73] V. M. Ingman, A. J. Schaefer, L. R. Andreola, S. E. Wheeler, *WIREs Comput. Mol. Sci.* **2021**, *11*, e1510, https://doi.org/https://doi.org/10.1002/wcms.1510.

[74] J.-G. Sobez, M. Reiher*, J. Chem. Inf. Model*. **2020**, *60*, 3884, https://doi.org/10.1021/acs.jcim.0c00503.

[75] E. I. Ioannidis, T. Z. H. Gani, H. J. Kulik, *J. Comput. Chem*. **2016**, *37*, 2106, https://doi.org/10.1002/jcc.24437.

[76] D. J. Durand, N. Fey, *Acc. Chem. Res*. **2021**, *54*, 837, https://doi.org/10.1021/acs.accounts.0c00807.

[77] N. Fey, A. C. Tsipis, S. E. Harris, J. N. Harvey, A. G. Orpen, R. A. Mansson, *Chem. Eur. J.* **2006**, *12*, 291, https://doi.org/https://doi.org/10.1002/chem.200500891.

[78] J. Jover, N. Fey, J. N. Harvey, G. C. Lloyd-Jones, A. G. Orpen, G. J. J. Owen-Smith, P. Murray, D. R. J. Hose, R. Osborne, M. Purdie, *Organometallics* **2010**, *29*, 6245, https://doi.org/10.1021/om100648v.

[79] N. Fey, J. N. Harvey, G. C. Lloyd-Jones, P. Murray, A. G. Orpen, R. Osborne, M. Purdie, *Organometallics* **2008**, *27*, 1372, https://doi.org/10.1021/om700840h.

[80] D. J. Durand, N. Fey, *Chem. Rev*. **2019**, *119*, 6561, https://doi.org/10.1021/acs.chemrev.8b00588.

[81] J. Jover, N. Fey, *Dalton Trans*. **2013**, *42*, 172, https://doi.org/10.1039/C2DT32099A.

[82] N. Fey, *Chem. Cent. J*. **2015**, *9*, 38, https://doi.org/10.1186/s13065-015-0104-5.

[83] T. Gensch, G. dos Passos Gomes, P. Friederich, E. Peters, T. Gaudin, R. Pollice, K. Jorner, A. Nigam, M. Lindner-D'Addario, M. S. Sigman, A. Aspuru-Guzik, *J. Am. Chem. Soc*. **2022**, *144*, 1205, https://doi.org/10.1021/jacs.1c09718.

[84] D. Zell, C. Kingston, J. Jermaks, S. R. Smith, N. Seeger, J. Wassmer, L. E. Sirois, C. Han, H. Zhang, M. S. Sigman, F. Gosselin, *J. Am. Chem. Soc*. **2021**, *143*, 19078, https://doi.org/10.1021/jacs.1c08399.

[85] T. Gensch, S. R. Smith, T. J. Colacot, Y. N. Timsina, G. Xu, B. W. Glasspoole, M. S. Sigman, *ACS Catal*. **2022**, *12*, 7773, https://doi.org/10.1021/acscatal.2c01970.

[86] J. Dotson, L. van Dijk, J. Timmerman, S. Grosslight, R. Walroth, K. Püntener, F. Gosselin, K. Mack, M. S. Sigman, *ChemRxiv* **2020**, https://doi.org/10.26434/chemrxiv-2022-qqxd1.

[87] J. M. Crawford, T. Gensch, M. S. Sigman, J. M. Elward, J. E. Steves, *Org. Process Res. Dev*. **2022**, *26*, 1115, https://doi.org/10.1021/acs.oprd.1c00357.

[88] M. Christensen, L. P. E. Yunker, F. Adedeji, F. Häse, L. M. Roch, T. Gensch, G. dos Passos Gomes, T. Zepel, M. S. Sigman, A. Aspuru-Guzik, J. E. Hein, *Commun. Chem*. **2021**, *4*, 112, https://doi.org/10.1038/s42004-021-00550-x.

[89] J. P. Janet, C. Duan, A. Nandy, F. Liu, H. J. Kulik, *Acc. Chem. Res*. **2021**, *54*, 532, https://doi.org/10.1021/acs.accounts.0c00686.

[90] A. Nandy, C. Duan, M. G. Taylor, F. Liu, A. H. Steeves, H. J. Kulik, *Chem. Rev*. **2021**, *121*, 9927, https://doi.org/10.1021/acs.chemrev.1c00347.

[91] S. Gugler, J. P. Janet, H. J. Kulik, *Mol. Syst. Des. Eng***. 2020**, *5*, 139, https://doi.org/10.1039/C9ME00069K.

[92] J. P. Janet, S. Ramesh, C. Duan, H. J. Kulik, *ACS Cent. Sci*. **2020**, *6*, 513, https://doi.org/10.1021/acscentsci.0c00026.

[93] J. P. Janet, C. Duan, T. Yang, A. Nandy, H. J. Kulik, *Chem. Sci*. **2019**, *10*, 7913, https://doi.org/10.1039/C9SC02298H.

[94] C. R. Groom, I. J. Bruno, M. P. Lightfoot, S. C. Ward, *Acta Crystallogr. B* **2016**, *72*, 171, https://doi.org/10.1107/S2052520616003954.

[95] C. R. Groom, F. H. Allen, *Angew. Chem. Int. Ed*. **2014**, *53*, 662, https://doi.org/https://doi.org/10.1002/anie.201306438.

[96] S. Vela, R. Laplaza, Y. Cho, C. Corminboeuf, *npj Comput. Mater*. **2022**, *8*, 188, https://doi.org/10.1038/s41524-022-00874-9.

[97] Z. Peng, *Drug Discov. Today Technol*. **2013**, *10*, e387, https://doi.org/10.1016/j.ddtec.2013.01.004.

[98] J.-L. Reymond, *Acc. Chem. Res*. **2015**, *48*, 722, https://doi.org/10.1021/ar500432k.

[99] L. Ruddigkeit, R. van Deursen, L. C. Blum, J.-L. Reymond, *J. Chem. Inf. Model*. **2012**, *52*, 2864, https://doi.org/10.1021/ci300415d.

[100] L. C. Blum, J.-L. Reymond, *J. Am. Chem. Soc*. **2009**, *131*, 8732, https://doi.org/10.1021/ja902302h.

[101] K. Meier, S. Bühlmann, J. Arús-Pous, J.-L. Reymond, *CHIMIA* **2020**, *74*, 241, https://doi.org/10.2533/chimia.2020.241.

[102] T. Fink, H. Bruggesser, J.-L. Reymond, *Angew. Chem. Int. Ed*. **2005**, *44*, 1504, https://doi.org/10.1002/anie.200462457.

[103] A. Gaulton, A. Hersey, M. Nowotka, A. P. Bento, J. Chambers, D. Mendez, P. Mutowo, F. Atkinson, L. J. Bellis, E. Cibrián-Uhalte, M. Davies, N. Dedman, A. Karlsson, M. P. Magariños, J. P. Overington, G. Papadatos, I. Smit, A. R. Leach, *Nucleic Acids Res*. **2017**, *45*, D945, https://doi.org/10.1093/nar/gkw1074.

[104] T. Sterling, J. J. Irwin, *J. Chem. Inf. Model*. **2015**, *55*, 2324, https://doi.org/10.1021/acs.jcim.5b00559.

[105] S. Kim, J. Chen, T. Cheng, A. Gindulyte, J. He, S. He, Q. Li, B. A. Shoemaker, P. A. Thiessen, B. Yu, L. Zaslavsky, J. Zhang, E. E. Bolton, *Nucleic Acids Res.* **2019**, *47*, D1102, https://doi.org/10.1093/nar/gky1033.

[106] M. Sorokina, P. Merseburger, K. Rajan, M. A. Yirik, C. Steinbeck, *J. Cheminform*. **2021**, *13*, 2, https://doi.org/10.1186/s13321-020-00478-9.

[107] D. S. Wishart, Y. D. Feunang, A. C. Guo, E. J. Lo, A. Marcu, J. R. Grant, T. Sajed, D. Johnson, C. Li, Z. Sayeeda, N. Assempour, I. Iynkkaran, Y. Liu, A. Maciejewski, N. Gale, A. Wilson, L. Chin, R. Cummings, D. Le, A. Pon, C. Knox, M. Wilson, *Nucleic Acids Res*. **2018**, *46*, D1074, https://doi.org/10.1093/nar/gkx1037.

[108] A. Wahab, L. Pfuderer, E. Paenurk, R. Gershoni-Poranne, *J. Chem. Inf. Model*. **2022**, *62*, 3704, https://doi.org/10.1021/acs.jcim.2c00503.

[109] S. Gallarati, P. van Gerwen, R. Laplaza, A. Fabrizio, S. Vela, C. Corminboeuf, *Chem. Sci*. **2022**, https://doi.org/10.1039/D2SC04251G.

[110] S. Gallarati, R. Laplaza, C. Corminboeuf, *Org. Chem. Front*. **2022**, *9*, 4041, https://doi.org/10.1039/D2QO00550F.

[111] J. A. Hageman, J. A. Westerhuis, H.-W. Frühauf, G. Rothenberg, *Adv. Synth. Catal*. **2006**, *348*, 361, https://doi.org/10.1002/adsc.200505299.

[112] S. Kozuch, J. M. L. Martin, *ACS Catal*. **2012**, *2*, 2787, https://doi.org/10.1021/cs3005264.

[113] Q. Peng, F. Duarte, R. S. Paton, *Chem. Soc. Rev*. **2016**, *45*, 6093, https://doi.org/10.1039/C6CS00573J.

[114] S. Kozuch, S. Shaik, *Acc. Chem. Res*. **2011**, *44*, 101, https://doi.org/10.1021/ar1000956.

[115] M. D. Wodrich, B. Sawatlon, M. Busch, C. Corminboeuf, *Acc. Chem. Res*. **2021**, *54*, 1107, https://doi.org/10.1021/acs.accounts.0c00857.

[116] M. D. Wodrich, B. Sawatlon, E. Solel, S. Kozuch, C. Corminboeuf, *ACS Catal*. **2019**, *9*, 5716, https://doi.org/10.1021/acscatal.9b00717.

[117] M. Foscato, V. R. Jensen, *ACS Catal.* **2020**, *10*, 2354, https://doi.org/10.1021/acscatal.9b04952.

[118] J. A. G. Torres, S. H. Lau, P. Anchuri, J. M. Stevens, J. E. Tabora, J. Li, A. Borovika, R. P. Adams, A. G. Doyle, *J. Am. Chem. Soc*. **2022**, *144*, 19999, https://doi.org/10.1021/jacs.2c08592.

[119] F. Häse, L. M. Roch, A. Aspuru-Guzik, *Chem. Sci.* **2018**, *9*, 7642, https://doi.org/10.1039/C8SC02239A.

[120] Y. Shen, J. E. Borowski, M. A. Hardy, R. Sarpong, A. G. Doyle, T. Cernak, *Nat. Rev. Methods Primers* **2021**, *1*, 23, https://doi.org/10.1038/s43586-021-00022-5.

[121] C. W. Coley, D. A. Thomas, J. A. M. Lummiss, J. N. Jaworski, C. P. Breen, V. Schultz, T. Hart, J. S. Fishman, L. Rogers, H. Gao, R. W. Hicklin, P. P. Plehiers, J. Byington, J. S. Piotti, W. H. Green, A. J. Hart, T. F. Jamison, K. F. Jensen, *Science* **2019**, *365*, eaax1566, https://doi.org/10.1126/science.aax1566.

[122] M. Cordova, M. D. Wodrich, B. Meyer, B. Sawatlon, C. Corminboeuf, *ACS Catal*. **2020**, *10*, 7021, https://doi.org/10.1021/acscatal.0c00774.

[123] D. M. Lustosa, A. Milo, *ACS Catal*. **2022**, *12*, 7886, https://doi.org/10.1021/acscatal.2c01741.

[124] R. Laplaza, J.-G. Sobez, M. D. Wodrich, M. Reiher, C. Corminboeuf, *Chem. Sci*. **2022**, *13*, 6858, https://doi.org/10.1039/D2SC01714H.

[125] H. L. Morgan, *J. Chem. Doc*. **1965**, *5*, 107, https://doi.org/10.1021/c160017a018.

[126] D. Rogers, M. Hahn, *J. Chem. Inf. Model*. **2010**, *50*, 742, https://doi.org/10.1021/ci100050t.

[127] A. Capecchi, D. Probst, J.-L. Reymond, *J. Cheminform*. **2020**, *12*, 43, https://doi.org/10.1186/s13321-020-00445-4.

[128] D. Probst, J.-L. Reymond, *J. Cheminform*. **2018**, *10*, 66, https://doi.org/10.1186/s13321-018-0321-8.

[129] N. Tsuji, P. Sidorov, C. Zhu, Y. Nagata, T. Gimadiev, A. Varnek, B. List, *ChemRxiv* **2022**.

[130] F. Sandfort, F. Strieth-Kalthoff, M. Kühnemund, C. Beecks, F. Glorius, *Chem* **2020**, *6*, 1379, https://doi.org/10.1016/j.chempr.2020.02.017.

[131] L. Pattanaik, C. W. Coley, *Chem* **2020**, *6*, 1204, https://doi.org/10.1016/j.chempr.2020.05.002.

[132] O. A. von Lilienfeld, *Angew. Chem. Int. Ed.* **2018**, *57*, 4164, https://doi.org/10.1002/anie.201709686.

[133] F. Musil, A. Grisafi, A. P. Bartók, C. Ortner, G. Csányi, M. Ceriotti, *Chem. Rev*. **2021**, *121*, 9759, https://doi.org/10.1021/acs.chemrev.1c00021.

[134] V. L. Deringer, A. P. Bartók, N. Bernstein, D. M. Wilkins, M. Ceriotti, G. Csányi, *Chem. Rev.* **2021**, *121*, 10073, https://doi.org/10.1021/acs.chemrev.1c00022.

[135] M. Rupp, A. Tkatchenko, K.-R. Müller, O. A. von Lilienfeld, *Phys. Rev. Lett*. **2012**, *108*, 058301, https://doi.org/10.1103/PhysRevLett.108.058301.

[136] B. Huang, O. A. von Lilienfeld, *Nat. Chem*. **2020**, *12*, 945, https://doi.org/10.1038/s41557-020-0527-z.

[137] F. A. Faber, A. S. Christensen, B. Huang, O. A. von Lilienfeld, *J. Chem. Phys*. **2018**, *148*, 241717, https://doi.org/10.1063/1.5020710.

[138] A. S. Christensen, L. A. Bratholm, F. A. Faber, O. Anatole von Lilienfeld, *J. Chem. Phys*. **2020**, *152*, 044107, https://doi.org/10.1063/1.5126701.

[139] A. Grisafi, M. Ceriotti, *J. Chem. Phys*. **2019**, *151*, 204105, https://doi.org/10.1063/1.5128375.

[140] H. Huo, M. Rupp, *arXiv:1704.06439* **2017**, https://doi.org/10.48550/ARXIV.1704.06439.

[141] J. Behler, *J. Chem. Phys*. **2011**, *134*, 074106, https://doi.org/10.1063/1.3553717.

[142] R. Drautz, *Phys. Rev. B* **2019**, *99*, 014104, https://doi.org/10.1103/PhysRevB.99.014104.

[143] A. P. Bartók, R. Kondor, G. Csányi, *Phys. Rev. B* **2013**, *87*, 184115, https://doi.org/10.1103/PhysRevB.87.184115.

[144] L. Zhu, M. Amsler, T. Fuhrer, B. Schaefer, S. Faraji, S. Rostami, S. A. Ghasemi, A. Sadeghi, M. Grauzinyte, C. Wolverton, S. Goedecker, *J. Chem. Phys*. **2016**, *144*, 034203, https://doi.org/10.1063/1.4940026.

[145] J. Nigam, S. Pozdnyakov, M. Ceriotti, *J. Chem. Phys.* **2020**, *153*, 121101, https://doi.org/10.1063/5.0021116.

[146] A. Fabrizio, K. R. Briling, C. Corminboeuf, *Digital Discovery* **2022**, *1*, 286, https://doi.org/10.1039/D1DD00050K.

[147] Z. Qiao, M. Welborn, A. Anandkumar, F. R. Manby, T. F. Miller, *J. Chem. Phys*. **2020**, *153*, 124111, https://doi.org/10.1063/5.0021955.

[148] A. S. Christensen, S. K. Sirumalla, Z. Qiao, M. B. O'Connor, D. G. A. Smith, F. Ding, P. J. Bygrave, A. Anandkumar, M. Welborn, F. R. Manby, T. F. Miller, *J. Chem. Phys*. **2021**, *155*, 204103, https://doi.org/10.1063/5.0061990.

[149] A. Grisafi, D. M. Wilkins, G. Csányi, M. Ceriotti, *Phys. Rev. Lett*. **2018**, *120*, 036002, https://doi.org/10.1103/PhysRevLett.120.036002.

[150] J. Weinreich, N. J. Browning, O. A. von Lilienfeld, *J. Chem. Phys*. **2021**, *154*, 134113, https://doi.org/10.1063/5.0041548.

[151] A. F. Zahrt, J. J. Henle, B. T. Rose, Y. Wang, W. T. Darrow, S. E. Denmark, *Science* **2019**, *363*, eaau5631, https://doi.org/10.1126/science.aau5631.

[152] S. Gallarati, R. Fabregat, R. Laplaza, S. Bhattacharjee, M. D. Wodrich, C. Corminboeuf, *Chem. Sci.* **2021**, *12*, 6879, https://doi.org/10.1039/D1SC00482D.

[153] P. van Gerwen, A. Fabrizio, M. D. Wodrich, C. Corminboeuf, *Mach. Learn.: Sci. Technol*. **2022**, *3*, 045005, https://doi.org/10.1088/2632-2153/ac8f1a.

[154] L. C. Gallegos, G. Luchini, P. C. St. John, S. Kim, R. S. Paton, *Acc. Chem. Res*. **2021**, *54*, 827, https://doi.org/10.1021/acs.accounts.0c00745.

[155] L. P. Hammett, *J. Am. Chem. Soc*. **1937**, *59*, 96, https://doi.org/10.1021/ja01280a022.

[156] W. L. Williams, L. Zeng, T. Gensch, M. S. Sigman, A. G. Doyle, E. V. Anslyn, *ACS Cent. Sci*. **2021**, *7*, 1622, https://doi.org/10.1021/acscentsci.1c00535.

[157] X. H. Liu, H. Y. Song, X. H. Ma, M. J. Lear, Y. Z. Chen, *J. Mol. Catal. A: Chem*. **2010**, *319*, 114, https://doi.org/10.1016/j.molcata.2009.12.008.

[158] S. Singh, M. Pareek, A. Changotra, S. Banerjee, B. Bhaskararao, P. Balamurugan, R. B. Sunoj, *Proc. Natl. Acad. Sci. USA* **2020**, *117*, 1339, https://doi.org/10.1073/pnas.1916392117.

[159] X. Li, S.-Q. Zhang, L.-C. Xu, X. Hong, *Angew. Chem. Int. Ed*. **2020**, *59*, 13253, https://doi.org/10.1002/anie.202000959.

[160] S. M. Maley, D.-H. Kwon, N. Rollins, J. C. Stanley, O. L. Sydora, S. M. Bischof, D. H. Ess, *Chem. Sci*. **2020**, *11*, 9665, https://doi.org/10.1039/D0SC03552A.

[161] J. Chen, W. Jiwu, L. Mingzong, T. You, *J. Mol. Catal. A: Chem*. **2006**, *258*, 191, https://doi.org/https://doi.org/10.1016/j.molcata.2006.05.020.

[162] J. Qiu, J. Xie, S. Su, Y. Gao, H. Meng, Y. Yang, K. Liao, *Chem* **2022**, https://doi.org/10.1016/j.chempr.2022.08.015.

[163] L.-C. Xu, S.-Q. Zhang, X. Li, M.-J. Tang, P.-P. Xie, X. Hong, *Angew. Chem. Int. Ed.* **2021**, *60*, 22804, https://doi.org/10.1002/anie.202106880.

[164] Y. Guan, C. W. Coley, H. Wu, D. Ranasinghe, E. Heid, T. J. Struble, L. Pattanaik, W. H. Green, K. F. Jensen, *Chem. Sci.* **2021**, *12*, 2198, https://doi.org/10.1039/D0SC04823B.

[165] R. Fabregat, P. van Gerwen, M. Haeberle, F. Eisenbrand, C. Corminboeuf, *Mach. Learn.: Sci. Technol*. **2022**, *3*, 035015, https://doi.org/10.1088/2632-2153/ac8e4f.

[166] B. Meyer, B. Sawatlon, S. Heinen, O. A. von Lilienfeld, C. Corminboeuf, *Chem. Sci*. **2018**, *9*, 7069, https://doi.org/10.1039/C8SC01949E.

[167] M. S. Sigman, K. C. Harper, E. N. Bess, A. Milo, *Acc. Chem. Res*. **2016**, *49*, 1292, https://doi.org/10.1021/acs.accounts.6b00194.

[168] J. M. Crawford, C. Kingston, F. D. Toste, M. S. Sigman, *Acc. Chem. Res*. **2021**, *54*, 3136, https://doi.org/10.1021/acs.accounts.1c00285.

[169] C. B. Santiago, J.-Y. Guo, M. S. Sigman, *Chem. Sci*. **2018**, *9*, 2398, https://doi.org/10.1039/C7SC04679K.

[170] A. Milo, E. N. Bess, M. S. Sigman, *Nature* **2014**, *507*, 210, https://doi.org/10.1038/nature13019.

[171] B. J. Shields, J. Stevens, J. Li, M. Parasram, F. Damani, J. I. M. Alvarado, J. M. Janey, R. P. Adams, A. G. Doyle, *Nature* **2021**, *590*, 89, https://doi.org/10.1038/s41586-021-03213-y.

[172] R. J. Hickman, M. Aldeghi, F. Häse, A. Aspuru-Guzik, *Digital Discovery* **2022**, *1*, 732, https://doi.org/10.1039/D2DD00028H.

[173] M. D. Wodrich, M. Chang, S. Gallarati, Ł. Woźniak, N. Cramer, C. Corminboeuf, *Chem. Eur. J.* **2022**, *28*, e202200399, https://doi.org/https://doi.org/10.1002/chem.202200399.

[174] S. El-Fayyoumy, M. Todd, C. Richards, *Beilstein J. Org. Chem*. **2009**, *5*, https://doi.org/10.3762/bjoc.5.67.

License and Terms