Research Article
doi.org/10.1002/cmtd.202100107

Chemistry Europe
European Chemical Societies Publishing

www.chemistrymethods.org

Genetic Optimization of Homogeneous Catalysts

Ruben Laplaza,[a, b] Simone Gallarati,[a] and Clemence Corminboeuf*[a, b]

We present the NaviCatGA package, a versatile genetic algorithm capable of optimizing molecular catalyst structures using well-suited fitness functions to achieve a set of targeted properties. The flexibility and generality of this tool are validated and demonstrated with two examples: i) Ligand optimization and exploration for Ni-catalyzed aryl-ether cleavage manipulating SMILES and using a fitness function derived from molecular volcano plots, ii) multi-objective (i.e., activity/selectivity) optimization of bipyridine *N,N*-dioxide Lewis basic organocatalysts for the asymmetric propargylation of benzaldehyde from 3D molecular fragments. We show that evolutionary optimization, enabled by NaviCatGA, is an efficient way of accelerating catalyst discovery through bypassing combinatorial scaling issues and incorporating compelling chemical constraints.

Introduction

This work introduces NaviCatGA, a software package capable of optimizing catalysts by exploiting any suitable fitness function that describes their catalytic performance. It manipulates catalyst structures generated in situ from a user-defined library of catalyst components (metal centers, ligands or ligand substituents, scaffolds, etc.); structures can be assembled from the respective components using any molecular representation, including SMILES strings and XYZ coordinates, and evaluated according to any fitness function (e.g., molecular volcano plot descriptors,[1,2] multivariate linear regression expressions[3]). NaviCatGA is a modular part of the broader NaviCat (**Navi**gating **Cat**alysis) platform for catalyst discovery, which includes other utilities and tools (e.g., database constructors, automatic volcano plot builder, etc.).

In the spirit of inverse design,[4–8] NaviCatGA uses a Genetic Algorithm (GA)[9–12] to find optimal catalysts (Figure 1). This pipeline represents a complementary approach to high-throughput screening[13–17] that becomes comparatively more efficient as the dimensionality of the combinatorial space of catalyst components grows. Furthermore, evolutionary experiments with GAs lead to alternative chemical insight into catalyst performance, as demonstrated hereafter. GAs have been shown to be well-suited for molecular optimization,[9,18,19] because they are able to address discontinuities in structure-property space (e.g., activity cliffs)[20,21] and, more importantly, do not require meaningful gradients for the optimization. Nonetheless, flexible and robust implementations of GA algorithms tailored for homogeneous catalysis were lacking.

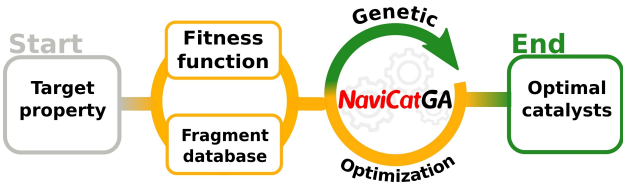


Figure 1. Schematic catalyst optimization pipeline powered by NaviCatGA.

The versatility and efficiency of NaviCatGA are illustrated with two representative applications to transition-metal and organocatalyzed reactions. The goal is to show that closed-loop optimisation with genetic algorithms is an efficient strategy to streamline computer-aided catalyst discovery. The code, documentation, and examples are openly available at https://github.com/lcmd-epfl/NaviCatGA.

Computational Methods

Overview of the NaviCatGA package

NaviCatGA is a lightweight genetic algorithm package that offers a simple, versatile and scalable solution to catalyst optimization problems. Simplicity is given by its Python structure and small number of dependencies, facilitating its adaptation and modification with minimal coding skills. Versatility comes from its modular design, which allows the user to define the optimization problem with utmost flexibility. For scalability, NaviCatGA relies upon the main strengths of genetic algorithms: the ability to tackle a large number of dimensions that are run in parallel. The genetic optimization loop is shown in Figure 2a.

[a] *Dr. R. Laplaza, S. Gallarati, Prof. C. Corminboeuf*
Laboratory for Computational Molecular Design
Institute of Chemical Sciences and Engineering
Ecole Polytechnique Fédérale de Lausanne (EPFL)
1015 Lausanne (Switzerland)
E-mail: clemence.corminboeuf@epfl.ch
[b] *Dr. R. Laplaza, Prof. C. Corminboeuf*
National Center for Competence in Research-Catalysis (NCCR-Catalysis)
Ecole Polytechnique Fédérale de Lausanne (EPFL)
1015 Lausanne (Switzerland)

Supporting information for this article is available on the WWW under https://doi.org/10.1002/cmtd.202100107

Research Article
doi.org/10.1002/cmtd.202100107
Chemistry Europe
European Chemical Societies Publishing

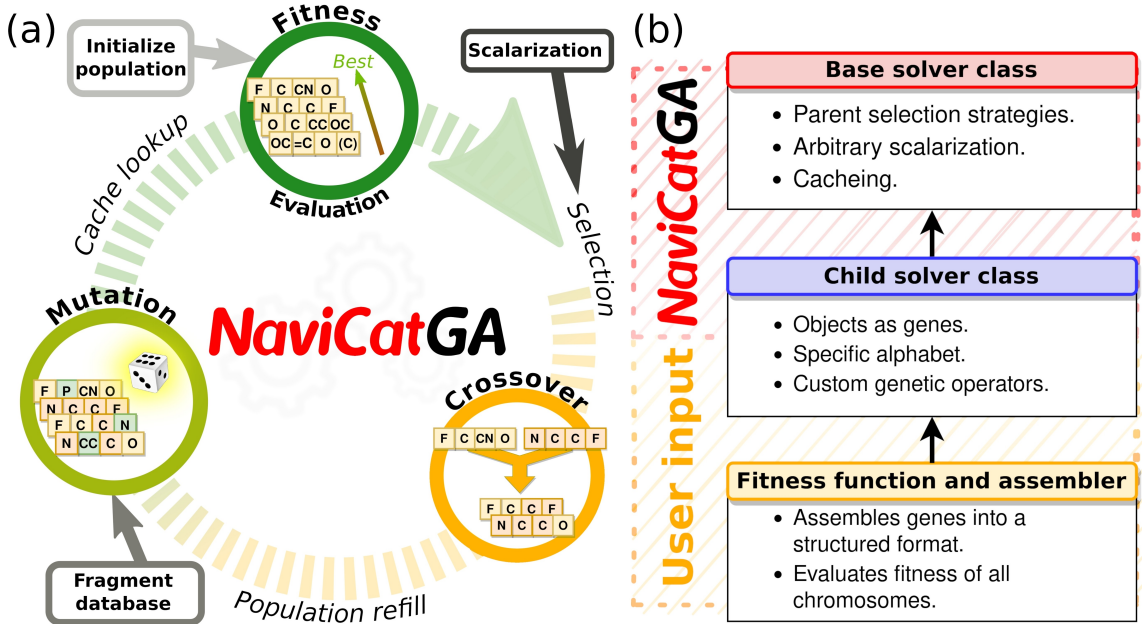


Figure 2. (a) Optimization loop followed by NaviCatGA (b) Schematic representation of the user input and the functionalities implemented.

The three distinct levels in which the NaviCatGA package is structured are represented in Figure 2b: the base solver class with all core functionalities, the child solver class (several of which are provided) that defines the problem type (crossover and mutation), and the user input (assembler and fitness function). This structure allows for significantly increased flexibility and adaptability, whereas adapting existing optimization tools could be difficult.[11]

Base Solver Class

The core genetic loop is provided by the GenAlgSolver base class (see Figure 2b). By design, the base class is data-type agnostic, with individuals represented by flexible lists of elements, and contains the solve method, which performs the optimization run (fitness evaluation, crossover, and mutation). Five different selection strategies to decide which individuals to recombine are provided (i.e., two-by-two, roulette wheel, pairwise tournament, Boltzmann-weighted, and random). This choice regulates the greediness of the optimization by defining a number of individuals for cross-over. The number of selected individuals is limited to a percentage of the total population (i.e., the selection rate). Additional features are such as pruning of duplicates in each successive generation, a least-recently used cache of fitness evaluations and in situ scalarization of fitness, are implemented (see Figures 2a and 2b for an overview). It is also possible to lock specific genes, so that they remain unchanged during the optimization procedure.

Implemented solvers

The specificities of the optimization problem are imposed by a child class (Figure 2b), which defines the way mutation and cross-over are performed. Three child solver classes are provided: the SmilesGenAlgSolver, based on SMILES strings,[22] the SelfiesGenAlgSolver, based on SELFIES strings,[23,24] and the XYZGenAlgSolver, which uses AaronTools.py geometry objects,[25] representing a 3D molecular fragment. In these respective solvers, each gene has the corresponding data type. The SMILES and SELFIES solvers are suited for systems that can be readily represented as strings. On the other hand, the XYZ solver allows for detailed 3D control, as each gene contains a set of coordinates. As child classes define the data type of genes, they also contain all the possible values any given gene on an individual can take, which in NaviCatGA parlance is called an "alphabet" (Figure S1 in Supporting Information). Genes with the same alphabet are considered to be equivalent (i.e., they can be replaced and mixed with one another).

In the implemented child solver classes, mutation is defined as substitution of a randomly chosen percentage of genes, or mutation rate, by random elements of the respective alphabets (Figure 2a). In turn, cross-over is achieved by combining the equivalent genes over one or more randomly determined crossover points (single-point cross-over is exemplified in Figure 2a but additional crossover operators could be considered in the future).

Defining new child solver classes is simple, as the core shared functionalities are kept in the base solver class. Different data structures, supported by other libraries (e.g., Molassembler[26] or molSimplify[27]) could be used as alternative back-ends. Additionally, child classes can be inherited to

incorporate additional definitions of mutation and crossover without substantial modifications.

Fragmentation Scheme

The fragmentation scheme and the corresponding alphabets define the total catalyst components combinatorial space to be explored. This step has a twofold goal: Avoiding the consideration of catalysts that are not expected to be stable and/or synthetically accessible,[28] and ensuring the domain of applicability and transferability of the fitness function (see below).

Assembler and Fitness Function

Once an appropriate catalyst space is defined through the fragmentation scheme, the user is required to input the fitness function and the assembler function into the solver (Figure 2). The assembler function takes a given individual (a list of genes of the specified data-type) and assembles them into a potential catalyst. In the case of SMILES, assembly can be as simple as concatenation of characters. In the XYZGenAlgSolver child class, the fragments must be suitably assembled in 3D. The user is free to define any assembler function in order to generate more complex graph structures from the underlying chromosomes.

Finally, the fitness function takes as argument an individual as interpreted by the assembler function and returns a fitness value. By default, NaviCatGA attempts to maximize fitness, although internal scalarizers can be used to change the default (see Example 2 below for a complex demonstration of multi-objective optimization).

Choosing a Fitness Function

The choice of fitness function for catalyst optimization depend on the specific application. In a broad sense, NaviCatGA favors fitness functions that map a candidate catalyst's chemical structure to a measure of its performance in a given reaction.

Molecular volcano plots, which have been favored by us,[2] provide a way to connect a 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 the energy span or TOF).[29] Some of us previously trained kernel-based ML models to predict the volcano descriptor variables for large pool of catalysts, from an approximate intermediate structure.[30,31] As demonstrated in Example 1, this inexpensive mapping between chemical structure and reactivity constitutes a natural fitness function that can be exploited for the GA optimization. An alternative approach to rapidly evaluate the catalytic properties and thus the fitness function consists in fitting Multivariate Linear Regression (MLR) expressions.[3] In Example 2, we fit and use MLR expressions to relate both the activity (i.e., the volcano descriptor) and the selectivity, expressed in terms of $\Delta\Delta G^{\ddagger}$, to an intermediate structure. However, NaviCatGA imposes no constraint on the form of the fitness function and any alternative defined by the user is possible. In general, any ML-based models tailored for the prediction of catalytic properties constitute a powerful alternative.[32,33]

In order to help users defining fitness functions and assemblers conveniently, a number of predefined wrapper functions are provided, built around RDKit[34] and pySCF.[35,36] Frequent descriptors, such as frontier molecular orbital energies or molecular volumes, are provided through wrappers from multiple molecular formats, including SMILES. Coupling any of the solvers to production-level quantum chemical computations is equally possible. Thus, the set of wrappers allows users to define highly customised fitness functions with minimal coding effort.

Results and Discussion

Example 1: Exploration of Ligand Space for Ni-Catalyzed Aryl-Ether Cleavage

One of us recently explored the ligand space for Ni-catalyzed aryl-ether reductive cleavage (Figure 3) relying upon a tandem volcano plot-ML approach to screen over 10^5 Ni catalysts bearing over 140 000 different phosphine or carbene ligands.[31] The volcano peak (maximum activity, see Figure 3) was found to correspond to $\Delta G(\mathbf{4}) = -33$ kcal/mol, where $\Delta G(\mathbf{4})$ is the free energy change associated with the formation of intermediate **4** (see Figure 3), used as a descriptor variable. Interestingly, very few phosphine and carbene ligands lead to high turnover

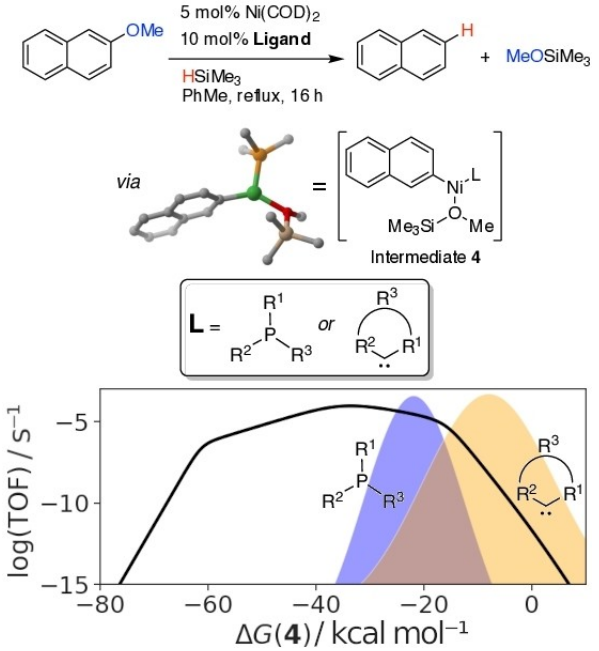


Figure 3. Reductive Ni-catalyzed cleavage of the 2-methoxynaphthalene C(sp^2)–O bond with trimethylsilane. The volcano plot predicts optimal catalytic activity at $\Delta G(\mathbf{4}) = -33$ kcal/mol. The blue and orange curves represent the approximate distribution of phosphine and carbene ligands, respectively (adapted from[31]).

frequencies, as they are spread in two gaussian distributions approximately centered on $\Delta G(\mathbf{4}) = -20$ kcal/mol (blue curve, Figure 3) and $\Delta G(\mathbf{4}) = -5$ kcal/mol (orange curve, Figure 3), respectively.

Based on the aforementioned exhaustive screening, we validate the capability of NaviCatGA by identifying the best phosphine ligands for the Ni catalyst with minimal computational cost. Additionally, we demonstrate how evolutionary experiments provide additional chemical insight and how they can be used to purge the pool of bad candidates from the database prior to further exploration. We finally demonstrate the versatility of the assembler function in exploiting the same procedure for the carbene ligands which, unlike phosphines, are composed of a backbone and two side groups.

Problem Definition

In this example, chromosomes are composed of three genes, accounting for the three different substituents in the phosphine ligands, all represented by SMILES strings using the SmilesGenAlgSolver class. The assembler is a function that generates the complete SMILES of intermediate **4** (see Figure 3) from the chromosome information. The combinatorial space, which was taken from[31] (see it listed in the Supporting Information), comprises a set of 68 possible substituents for the phosphine ligands, as well as 77 ring and 30 backbone substituents for carbene ligands. Note that these numbers could further increase by including more exotic ligands or by decomposing the fragments into smaller components. Yet, this extension would potentially compromise both the experimental relevance of the generated intermediates **4**, a typical flaw of generative models, and the accuracy of our fitness function (see below).

Fitness Function

Following our previous work,[31] a kernel ridge regression model is trained to predict $\Delta G(\mathbf{4})$ from the approximate 3D structure of intermediate **4** using the same database of 1473 catalysts. The trained model has a cross-validated MAE of < 4 kcal/mol. Details of the ML model can be found in the Supporting Information. For prediction, the SMILES in the GA is embedded to 3D coordinates using RDKit, then its SLATM representation[37] is obtained, which leads to its predicted $\Delta G(\mathbf{4})$ through the trained regression coefficients and kernel. For a candidate *i* final fitness score f_i is obtained by evaluating its $\Delta G(\mathbf{4})_i$ value compared to a normalized gaussian distribution centered in the target value *x*, $f_i = exp\left(-\frac{1}{2}\left(\frac{\Delta G(\mathbf{4})_i - x}{\sigma}\right)^2\right)$ where $\sigma = |x|/2$.

Optimization

The genetic optimization is initiated with a population of 10 randomized ligands (individuals) and a mutation rate of 10% for 50 generations. The maximum number of evaluations, 500, is infinitesimal w.r.t. the combinatorial search space of 3×10^5 (68^3). The first run is set up with a target value of $x = -33$ kcal/mol, the peak of the volcano (maximum activity). Results are shown in Figure 4a (blue curves and frame). The GA is able to identify top candidates, lying exactly on the volcano top, within the first 30 iterations – under 300 total evaluations. The top candidate contains a bis(pentachlorophenyl)phosphine ligand, in agreement with our previous screening,[31] in which the pentachlorophenyl (**65**) substituent was identified as one of the best options. The overall increase in fitness coinciding with the selection of the pentachlorophenyl substituent by the optimizer occurs in generation 20, as illustrated by the sharp increase in the best fitness curve in Figure 4a. It is important to stress that the GA takes three orders of magnitude less evaluations than our previous screening approach to identify it.

Given the low computational cost of the run, ablation evolutionary experiments are performed to obtain additional insight and explore different possible local fitness maxima. First, the pentachlorophenyl (**65**) substituent is removed from the database and the optimization is run again. This run leads to the identification of isopropylamino (**48**) as a good substituent, shown in Figure 4b as the green curve and frame, again in agreement with our previous work. Removing the aforementioned substituent and re-running leads to an increasingly difficult start for the optimization run, as less good options are available, but nevertheless ultimately identifying the 2,2,2-trichloro-1-hydroxyethyl substituent (**66**) as a good candidate in less than 20 iterations (Figure 4b, red curve and frame). Overall, the three best substituents that had previously been identified (diisopropylamino, pentachlorophenyl, and 2,2,2-trichloro-1-hydroxyethyl) are correctly and systematically located by NaviCatGA in less than 600 evaluations.

A similar optimization run is performed for a target of $\Delta G(\mathbf{4}) = -10$ kcal/mol. This value, which corresponds to the right-hand-side of the volcano plot results in negligible catalytic activity. The GA identify ligands with a predicted $\Delta G(\mathbf{4})$ close to the targeted value, which leads to the identification of the least optimum substituents for the phosphine ligands, in this case N-containing heterocycles (Figure 4c). Both good and poor candidates are identified with the same setup.

Finally, we optimize a *N*-heterocyclic carbene ligand using the same parameters with a target of $x = -33$ kcal/mol. The flexibility of NaviCatGA facilitates alternative definition of the fragment combinatorial space (in this case, the N-atom substituents, see Supporting Information for details). The results, shown in Figure 4d, capture a key observation in line with previous work: unlike phosphine ligands, *N*-heterocyclic carbene ligands are generally unable to reach the top of the volcano. The optimization problem thus becomes harder as illustrated by the significantly lower fitness scores. Nevertheless, the genetic algorithm finds the best possible candidates within the combinatorial space, achieving a remarkably close value to the top using diisopropylamino substituents.[31] This optimization procedure provides a traceable evolution for every fit candidate and for the relative preference of the different substituents.

Research Article
doi.org/10.1002/cmtd.202100107
Chemistry Europe
European Chemical Societies Publishing

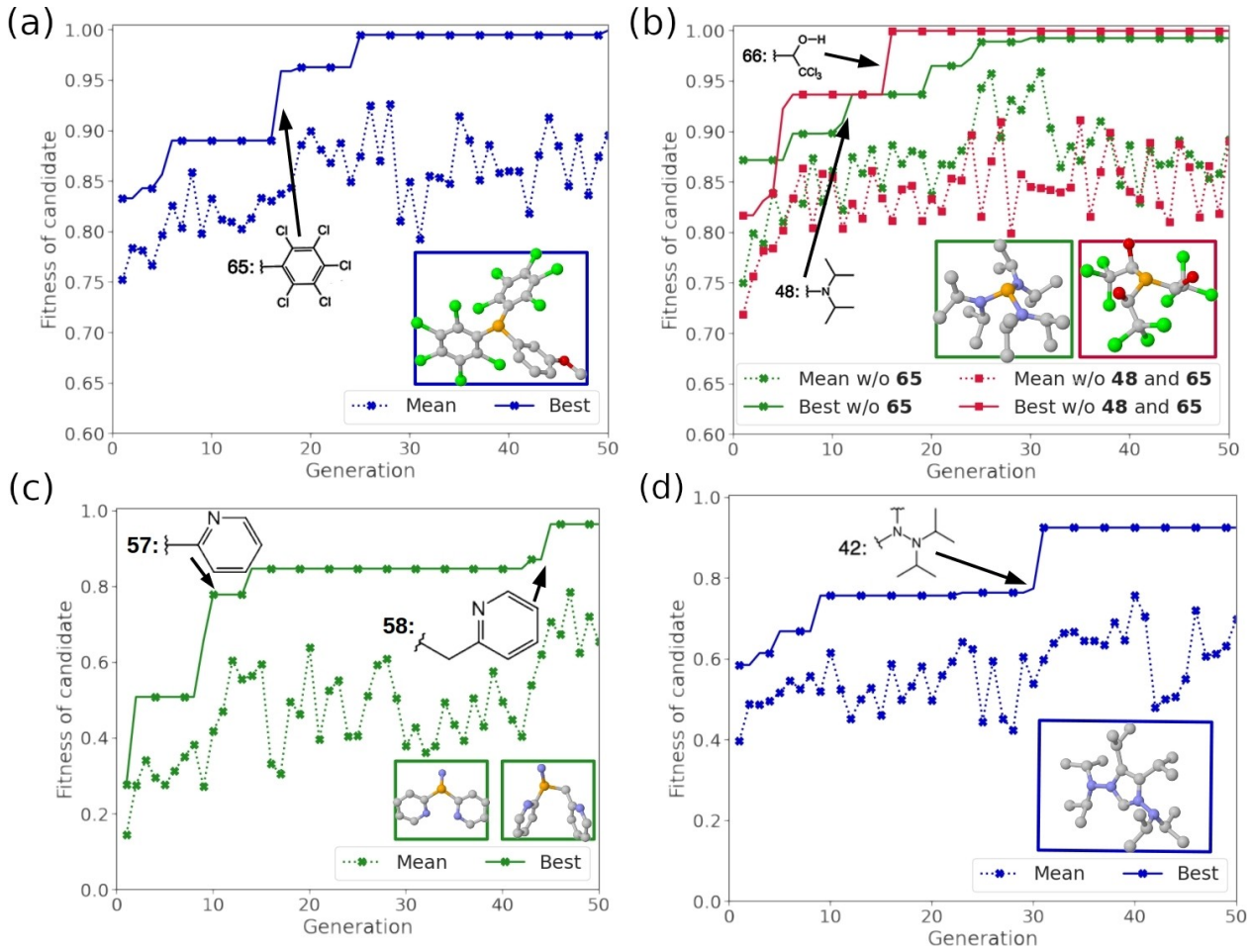


Figure 4. Evolution of mean population fitness and best candidate fitness over the optimization runs. Fitness is defined as $f_i = exp\left(-\left(\frac{\Delta G(\mathbf{4})_i - x}{|x|}\right)^2\right)$. The most fit ligands from each run are highlighted in the corresponding boxes (H atoms omitted for clarity). (a) Complete run over the whole combinatorial space with $x = -33$ kcal/mol. (b) Ablation experiments in which fragments with $x = -33$ kcal/mol are removed from the combinatorial space; removal of **65** is represented with green lines, removal of both **48** and **65** is represented in red. (c) Complete run over the whole combinatorial space with $x = -10$ kcal/mol. (d) Complete run over the carbene combinatorial space with $x = -33$ kcal/mol.

Example 2: Achieving the Activity/Selectivity Trade-Off with Enantioselective Organocatalysts

While Example 1 focuses on validation and comparison with high-throughput screening, this second example is chosen to illustrate the convenience of NaviCatGA to explore a large combinatorial space and optimize several properties simultaneously. Whenever several properties are to be optimized, there is often a trade-off between two or more targets preventing the existence of optimum solutions. In such cases, a large number of solutions to the optimization problem, the so-called Pareto front, can be identified depending on the criteria selected by a decision maker.

In catalyst design, a classic example of multi-objective optimization is the activity versus selectivity conundrum, where increased activity of a catalyst generally leads to decreased selectivity. A good catalyst should be both as active and as selective as possible. A pragmatic way to decide over this particular Pareto front is to search specifically for catalysts that retain noticeable activity while prioritizing selectivity, as opposed to compromising selectivity for increased activity, or reducing activity to a negligible level in search of perfect selectivity.

Given the flexible structure of NaviCatGA, the user imposes selected criteria on the optimization problem by assigning weights to different properties (e.g., the final fitness is defined 60% by selectivity and 40% by activity), or using step functions to define hard boundaries (e.g., give a fitness of 0 whenever selectivity drops under some value). However, translating human criteria into mathematical functions is difficult. NaviCatGA thus supports fitness functions which return several values, which are then processed by a scalarizer to derive the final, singular fitness value. Although any internal scalarizer object can be used, we recommend the achievement scalarizing function Chimera[38] to process multi-objective fitness functions within the optimization run. Chimera requires a priority ranking and a degradation threshold to be assigned for each optimization objective and generates a score for each candidate by assessing its relative performance in the population (for further details, we refer the reader to the original publication[38]).

Chimera's versatility matches NaviCatGA's and allows for the effortless formalization of complex human criteria.

To demonstrate conflicting multi-objective optimization, this example exploits a Chimera scalarizer to find optimal Lewis base organocatalysts for the enantioselective propargylation of benzaldehyde (Figure 5).[32,39–41]

Problem Definition

In this case, chromosomes are composed of three genes: a chiral scaffold (e.g., the parent scaffold **S1** is a (*S*)-2,2-bipyridine *N,N*-dioxide, **S2** and **S3** include additional Ph or ^{t}Bu substituents at the 6,6-positions, **S4** is a (*S*)-8,8-disubstituted 2,2-biquinoline *N,N*-dioxide, etc.) and two substituents at the 6,6-positions (see the Supporting Information for the full list of scaffolds **S1**-**S14**). The 3D coordinates of all substituents and scaffolds are obtained from DFT computations, and thus the XYZGenAlgSolver class is used. The assembler in this case is a function capable of building the 3D structure of intermediate **1** (Figure 5a) from a given chromosome by substituting the 3D structures of the two substituents in the 6,6-positions of the scaffold ($R^1$and R^2 in Figure 5a), with no reoptimization necessary. The combinatorial space is given by 14 *N,N*-dioxide scaffolds and 34 different substituents (16 184 combinations, see Supporting Information for details). Note that, to increase the size of the combinatorial space, catalysts with different 6 and 6′substituents, in addition to the more synthetically accessible symmetrically substituted ones, are considered.

Fitness Function

Based on previous work,[32,41] reference energies of intermediates **1**–**3** and of **TS1** are computed at the PCM(dichloromethane)/B97-D/TZV(2p,2d) level for 78 different organocatalysts using structures optimized at the same level of theory. Relative energies (i.e., electronic energies plus solvation free energies) at this level were found to be more robust to reproduce experimental results for this reaction.[40,41] A volcano plot is constructed for the propargylation of benzaldehyde with allenyltrichlorosilane (Figure 5a), leading to the identification of the descriptor variable ΔE(**1**) and of the the region of maximum activity (ΔE(**1**)$\approx$3 kcal/mol). Enantioselectivity is calculated as a function of $\Delta\Delta E^{\ddagger}$, which is defined as the difference between the (*R*)- and (*S*)-Boltzmann-weighted activation energies of the **2**→**TS1** reaction step, relative to the lowest-lying (*R*)- or (*S*)-ligand arrangement of **2** (see Supporting Information for further details).

Two Multivariate Linear Regression (MLR) expressions are then parametrized to predict $\Delta E(1)$ and the $\Delta\Delta E^{\ddagger}$ from the unoptimized 3D structure of intermediate **1** assembled by the genetic algorithm (Figure 5), using as parameters five dihedral angles, the Sterimol B5 and L values of the 6, 6′-substituents, and E_{LUMO} (see Supporting Information for details). The parametrized MLR expressions lead to RMSE values of 1.65 kcal/mol and 0.25 kcal/mol for ΔE(**1**) and $\Delta\Delta E^{\ddagger}$, respectively. Details and cross-validation of the MLR models are given in the Supporting Information.

Using the two MLR models, activity is gauged by the proximity of ΔE(**1**)$\approx$3 kcal/mol (plateau of maximum activity, see Figure 5b) and selectivity is defined as proportional to $\Delta\Delta E^{\ddagger}$. While the explicit MLR equations give a rough idea of the balance between different parameters, an explicit criterion has to be used to narrow down the Pareto front. Several

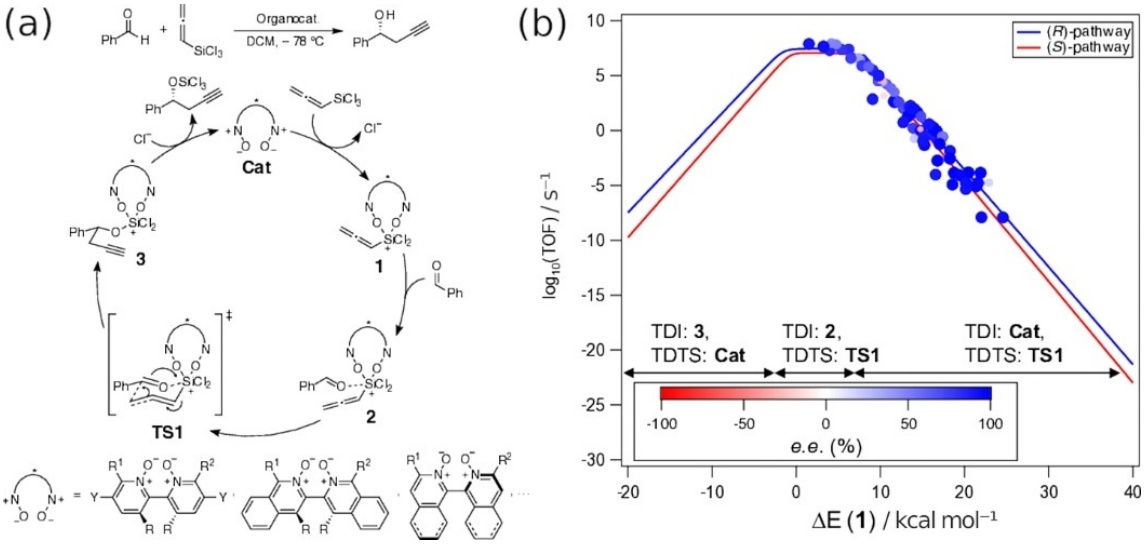


Figure 5. (a) Catalytic cycle for the bipyridine N,N-dioxide-catalyzed enantioselective propargylation of benzaldehyde (R = H or Me).[40,41] (b) Enantioselectivity TOF-molecular volcano plot for the 78 test set organocatalysts depicted in Figure S12. Larger and darker blue spheres indicate catalysts with higher ee values favouring (R)-product formation, smaller and red spheres indicate catalysts favouring (S)-product formation. The different slopes of the volcano correspond to different TOF-determining intermediates (TDI) and transition states (TDTS).

Research Article
doi.org/10.1002/cmtd.202100107
Chemistry Europe
European Chemical Societies Publishing

options are explored (see below) to showcase the importance of proper multi-objective criteria.

Optimization

Three GA runs are started with an initial population of 25 randomized individuals, consisting of a bipyridine *N,N*-dioxide scaffold and two $R^1$and R^2 substituents each, a mutation rate of 5% and a selection rate of 25%. All optimizations are run for 15 generations leading to a maximum of 375 evaluations out of the $> 10^4$ combinatorial possibilities. The fitness functions are all based on the aforementioned MLR expressions but scalarized differently using Chimera: for the first run, a minimum absolute $\Delta\Delta E^{\ddagger} = 1.5$ kcal/mol is imposed while $\Delta E(\mathbf{1})$ is minimized with a 25% degradation threshold, due to the flatness of the activity plateau around $\Delta E(\mathbf{1}) = 0$ kcal/mol. This exemplifies a standard situation in which enantioselectivity is to be guaranteed and only subsequently activity has to be optimized. After the optimization procedure (Figure 6a), several good candidates are found with predicted $\Delta\Delta E^{\ddagger}$ of ≈ 2 kcal/mol and $\Delta E(\mathbf{1})$ of 1 kcal/mol, with the top candidate having the (*S*)-2,2-bipyridine *N,N*-dioxide scaffold with Ph substituents at the 5,5-positions, R = Me, and R^1 = 3,5-Me-4-F–Ph and R^2 = 2,4,6-^{t}Bu–Ph. NaviCat-GA, driven by the scalarizer, is able to explore activity and selectivity and find a good compromise between both. The

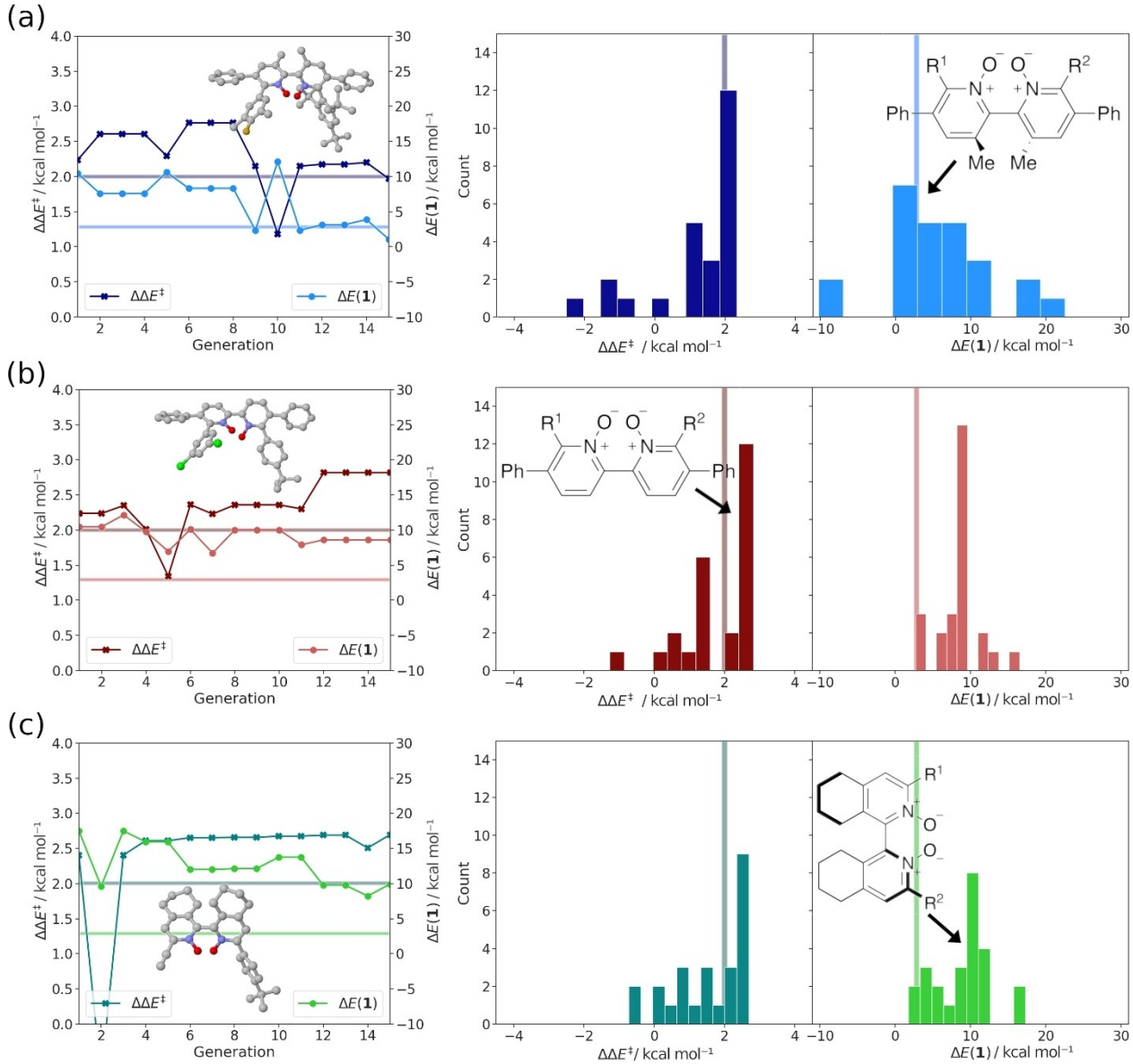


Figure 6. Evolution of maximum selectivity and activity over optimization runs with three different scalarization setups. The best catalyst candidate from each run is highlighted (H atoms omitted for clarity). Shaded lines indicate the optimal activity region of $\Delta E(\mathbf{1})$ (light hue) and a minimum $\Delta\Delta E^{\ddagger}$ threshold for guaranteed enantioselectivity (dark hue). The distribution of $\Delta\Delta E^{\ddagger}$ and $\Delta E(\mathbf{1})$ for the final populations of each run are shown right. (a) First setup with minimum $\Delta\Delta E^{\ddagger} = 1.5$ kcal/mol and 25% compromise on minimizing $\Delta E(\mathbf{1})$ (b) Second setup with maximum $\Delta E(\mathbf{1}) = 10$ kcal/mol and 25% compromise on maximizing $\Delta\Delta E^{\ddagger}$ (c) Third setup with minimum $\Delta\Delta E^{\ddagger} = 2.5$ kcal/mol and 50% compromise on minimizing $\Delta E(\mathbf{1})$.

distribution of values in the final population shows how it is enriched with high $\Delta\Delta E^{\ddagger}$ and low ΔE(**1**) candidates after 15 generations: a rightmost bump in the distribution of $\Delta\Delta E^{\ddagger}$ and a bump in the region between 0 and 10 in the distribution of ΔE(**1**) (Figure 6a).

For the second run, a maximum value of ΔE(**1**) = 10 kcal/mol is imposed, while $\Delta\Delta E^{\ddagger}$ is maximized with a 25% degradation threshold. This represents the opposite setting, in which good activity is guaranteed (the estimated TOF for ΔE(**1**) = 10 is ≈50 000 s^{-1}, see Figure 5b) and selectivity comes as a second priority. By inverting the priorities, the optimization problem becomes noticeably more difficult. For the first 10 generations, the top candidate found with this setup is stuck at the ΔE(**1**) = 10 kcal/mol mark, having $\Delta\Delta E^{\ddagger}$ slightly over 2 kcal/mol (scaffold = (*S*)-1,1-disubstituted 3,3-biisoquinoline *N,N*-dioxide, R = H, R^1 = I, R^2 = 4-^{t}Bu–Ph). In this case, the scalarizer setup leads to a very steep local optimum after a few exploratory generations, and evolution is hindered due to the relatively tight 25% degradation margin. The final population thus shows a very large percentage of nearly identical candidates. However, through mutation, the optimizer finds an optimal candidate with high selectivity and acceptable activity in the last four generations, depicted in Figure 6b. Here, the scaffold is (*S*)-2,2-bipyridine *N,N*-dioxide with 5,5-Ph substituents, R = H, R^1 = 3,5-Cl–Ph, and R^2 = 4-^{t}Bu–Ph. Some common trends are evident comparing the top performer of this optimization run with the results from the previous one, particularly the presence of a ^{t}Bu-substituted phenyl group in the R^2 position and of halogen-containing groups as R^1, as well as the similar (*S*)-2,2-bipyridine *N,N*-dioxide scaffold. The small change in scaffold (R = Me in the first run, R = H in the second one), which is associated to reduced activity in the second evolutionary experiment, exemplifies the difficulties associated with activity cliffs in catalyst design.

In the third run, we exemplify a more flexible setup requiring a minimum $\Delta\Delta E^{\ddagger}$ value of 2.5 kcal/mol while attempting to reach the top of the volcano as before, but accepting a 50% degradation of the latter to enforce the former, which provides much more flexibility than in the previous examples. In this case, the top candidates quickly present significant selectivity, but no compromise is achieved with respect to activity, and thus ΔE(**1**) is barely improved over the run and remains over 10 kcal/mol, in spite of the noticeable trade-off exploration in the early generations (with even a generation exploring structures that would lead to (*S*)-product formation in search of improved activity), which is afforded by the increased degradation margin. The final population of the run, shown in Figure 6c, excels in selectivity but is worse than the first two runs in terms of activity, with the distribution heavily centered around the 10 kcal/mol mark. The top candidate has a (*S*)-H_8-[1,1′biisoquinoline] 2,2-dioxide backbone with R^1 = CCH, R^2 = 4-^{t}Bu–Ph; this scaffold is shown to be associated with improved selectivity because of its dominating presence in the final population.

The comparison between the three runs highlights how the same optimization setup, guided by slightly different human input, ends up exploring very different areas of the combinatorial space and finds diverse solutions in the Pareto front. Hence, the use of scalarization and careful problem definition is recommended in order to navigate multi-objective optimization. For typical bipyridine *N,N*-dioxide-derived organocatalysts, selectivity is believed to arise from favorable electrostatic interactions between the formyl C–H of benzaldehyde and the nearby Cl ligand in the lowest-lying transition state structure leading to the (*R*)-alcohol.[41] Activity is largely a function of the organocatalyst's Lewis basicity, with better electron-donors able to more efficiently activate the allenyltrichlorosilane substrate (and hence being located closer to the volcano plateau), while catalysts bearing strongly electron-withdrawing substituents are less active and found lower on the right slope of the volcano. The evolutionary experiments highlight how changes in the scaffold and in the nature of $R^1$and R^2 affect this selectivity-activity interplay and reveal the unique role played by aromatic substituents. Ph groups at the 6 or 6-position with electron-donating alkyl substituents are clearly important for enhanced activity, although additional ^{t}Bu substituents (at the *ortho*-positions) cause unfavorable steric interactions with the formyl C–H, overwhelming the stabilizing effect from favorable C–H…Cl interactions and hence reducing selectivity (this is the case of the first run, Figure 6a). When placed at the 5,5-positions, the Ph groups lead to additional π-stacking interactions favoring the (*R*)-pathway (benzene trimer-like interactions involving benzaldehyde and two Ph substituents)[42] and offsetting otherwise unfavorable π-stacking and CH/π interactions that stabilize the (*S*)-pathway.[41,43] Thus, in the second run (Figure 6b), the presence of less electron-rich substituents (including hydrogen atoms instead of methyl groups at the R position) results in a slight loss of activity, but ensures favorable noncovalent interactions that yield very high selectivity. In line with recent experimental results,[43] the presence of aliphatic substituents (instead of aromatic ones) is associated with reduced activity (as in the third run, Figure 6c), however the ethynyl group as R^1 helps improve selectivity, since it leads to a more favorable electrostatic environment for the formyl C–H in the (*R*)-pathway (partially positively charged C–H interacting with the π-bonds in CCH).[41]

Conclusions

We presented NaviCatGA, a tool capable of optimizing the structure of homogeneous catalysts to find top candidates with tailored properties for a given reaction. Using evolutionary techniques, it is possible to perform the optimization task with the possibility of tracing the origin of favorable catalyst components (e.g., ligand substituents, catalyst scaffolds or side groups) during the evolutionary experiments and pinpoint their influence on different aspects of a catalyst's performance (e.g., activity, selectivity).

From a technical perspective, NaviCatGA is versatile, flexible and thus applicable to a variety of catalytic problems. Thanks to its hierarchical structure, it is compatible with diverse structural representations (e.g., SMILES, 3D structures), genetic operations and fitness functions. Additional functionalities, including ML-

based acceleration,[44–50] can also be conveniently deployed for the fitness evaluation. While NaviCatGA, as presented here, is a core component of inverse design efforts in catalysis, it also constitutes a powerful stand-alone program for general optimization problems.

In order to further streamline the inverse design workflow, it is desirable to automate the elucidation of the fitness function as well as of other eventual quantum chemical tasks. Within this context, NaviCatGA is integrated into the broader the NaviCat platform (https://github.com/lcmd-epfl/NaviCat), collecting an ensemble of tools for computational catalysis. This set of utility tools, which include, for instance, automated construction of volcano plots (https://github.com/lcmd-epfl/volcanic), can be used independently and/or in combination with each other. Overall, these efforts represent a complementary addition to alternative programs such as those addressing automated mechanistic studies[25,51–54] and structure generation.[26,27,55]

Acknowledgements

This publication was created as part of NCCR Catalysis (grant number 180544), a National Centre of Competence in Research funded by the Swiss National Science Foundation. C.C. and S.G. acknowledge funding from the European Research Council (ERC, Grant Agreement No. 817977) within the framework of European Union's H2020.

Conflict of Interest

The authors declare no conflict of interest.

Data Availability Statement

The NaviCatGA package with documentation and examples is available at https://github.com/lcmd-epfl/NaviCatGA, and as a part of the NaviCat project at https://github.com/lcmd-epfl/NaviCat. Chemical structures and minimal working examples of the optimization experiments are available at https://doi.org/10.24435/materialscloud:fz-sw.

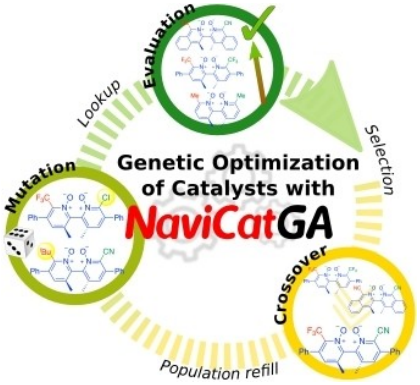
[1] M. Busch, M. D. Wodrich, C. Corminboeuf, *Chem. Sci.* **2015**, *6*, 6754.
[2] M. D. Wodrich, B. Sawatlon, M. Busch, C. Corminboeuf, *Acc. Chem. Res.* **2021**, *54*, 1107.
[3] M. S. Sigman, K. C. Harper, E. N. Bess, A. Milo, *Acc. Chem. Res.* **2016**, *49*, 1292.
[4] B. Sanchez-Lengeling, A. Aspuru-Guzik, *Science* **2018**, *361*, 360.
[5] M. B. Patrascu, J. Pottel, S. Pinus, M. Bezanson, P.-O. Norrby, N. Moitessier, *Nat. Catal.* **2020**, *3*, 574.
[6] G. dos Passos Gomes, R. Pollice, A. Aspuru-Guzik, *Trends Chem.* **2021**, *3*, 96.
[7] 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.
[8] T. Gensch, G. dos Passos Gomes, P. Friederich, E. Peters, T. Gaudin, R. Pollice, K. Jorner, A. Nigam, M. L. DAddario, M. S. Sigman, A. Aspuru-Guzik, **2021**, ChemRxiv preprint DOI 10.26434/chemrxiv.12996665.v1.
[9] F. Clerc, M. Lengliz, D. Farrusseng, C. Mirodatos, S. R. M. Pereira, R. Rakotomalala, *Rev. Sci. Instrum.* **2005**, *76*, 062208.
[10] M. Foscato, V. Venkatraman, V. R. Jensen, *J. Chem. Inf. Model.* **2019**, *59*, 4077.
[11] M. Foscato, V. R. Jensen, *ACS Catal.* **2020**, *10*, 2354.
[12] 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.
[13] M. Renom-Carrasco, L. Lefort, *Chem. Soc. Rev.* **2018**, *47*, 5038.
[14] A. R. Rosales, J. Wahlers, E. Limé, R. E. Meadows, K. W. Leslie, R. Savin, F. Bell, E. Hansen, P. Helquist, R. H. Munday, O. Wiest, P.-O. Norrby, *Nat. Catal.* **2018**, *2*, 41.
[15] T. Gensch, S. Smith, T. Colacot, Y. Timsina, G. Xu, B. Glasspoole, M. Sigman, **2021**, ChemRxiv preprint DOI 10.33774/chemrxiv-2021-fgm7v.
[16] A. Soyemi, T. Szilvási, *Dalton Trans.* **2021**, *50*, 10325.
[17] 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.
[18] J. H. Jensen, *Chem. Sci.* **2019**, *10*, 3567.
[19] J. P. Janet, L. Chan, H. J. Kulik, *J. Phys. Chem. Lett.* **2018**, *9*, 1064.
[20] D. Stumpfe, H. Hu, J. Bajorath, *ACS Omega* **2019**, *4*, 14360.
[21] S. Newman-Stonebraker, S. Smith, J. Borowski, E. Peters, T. Gensch, H. Johnson, M. Sigman, A. Doyle, **2021**, ChemRxiv preprint DOI 10.26434/chemrxiv.14388557.v1.
[22] D. Weininger, *J. Chem. Inf. Model.* **1988**, *28*, 31.
[23] M. Krenn, F. Häse, A. Nigam, P. Friederich, A. Aspuru-Guzik, *Mach. Learn.: Sci. Technol.* **2020**, *1*, 045024.
[24] A. Nigam, R. Pollice, M. Krenn, G. dos Passos Gomes, A. Aspuru-Guzik, *Chem. Sci.* **2021**, *12*, 7079.
[25] V. M. Ingman, A. J. Schaefer, L. R. Andreola, S. E. Wheeler, *WIREs Comput. Mol. Sci.* **2020**, *11*, e1510.
[26] J.-G. Sobez, M. Reiher, *J. Chem. Inf. Model.* **2020**, *60*, 3884.
[27] E. I. Ioannidis, T. Z. H. Gani, H. J. Kulik, *J. Comput. Chem.* **2016**, *37*, 2106.
[28] W. Gao, C. W. Coley, *J. Chem. Inf. Model.* **2020**, *60*, 5714.
[29] M. D. Wodrich, B. Sawatlon, E. Solel, S. Kozuch, C. Corminboeuf, *ACS Catal.* **2019**, *9*, 5716.
[30] B. Meyer, B. Sawatlon, S. Heinen, O. A. von Lilienfeld, C. Corminboeuf, *Chem. Sci.* **2018**, *9*, 7069.
[31] M. Cordova, M. D. Wodrich, B. Meyer, B. Sawatlon, C. Corminboeuf, *ACS Catal.* **2020**, *10*, 7021.
[32] S. Gallarati, R. Fabregat, R. Laplaza, S. Bhattacharjee, M. D. Wodrich, C. Corminboeuf, *Chem. Sci.* **2021**, *12*, 6879.
[33] S. Heinen, G. F. von Rudorff, O. A. von Lilienfeld, *J. Chem. Phys.* **2021**, *155*, 064105.
[34] "RDKit: Open-source cheminformatics," can be found under http://www.rdkit.org.
[35] Q. Sun, T. C. Berkelbach, N. S. Blunt, G. H. Booth, S. Guo, Z. Li, J. Liu, J. D. McClain, E. R. Sayfutyarova, S. Sharma, S. Wouters, G. K.-L. Chan, *WIREs Comput. Mol. Sci.* **2017**, *8*, e1340.
[36] Q. Sun, X. Zhang, S. Banerjee, P. Bao, M. Barbry, N. S. Blunt, N. A. Bogdanov, G. H. Booth, J. Chen, Z.-H. Cui, J. J. Eriksen, Y. Gao, S. Guo, J. Hermann, M. R. Hermes, K. Koh, P. Koval, S. Lehtola, Z. Li, J. Liu, N. Mardirossian, J. D. McClain, M. Motta, B. Mussard, H. Q. Pham, A. Pulkin, W. Purwanto, P. J. Robinson, E. Ronca, E. R. Sayfutyarova, M. Scheurer, H. F. Schurkus, J. E. T. Smith, C. Sun, S.-N. Sun, S. Upadhyay, L. K. Wagner, X. Wang, A. White, J. D. Whitfield, M. J. Williamson, S. Wouters, J. Yang, J. M. Yu, T. Zhu, T. C. Berkelbach, S. Sharma, A. Yu, Sokolov, G. K.-L. Chan, *J. Chem. Phys.* **2020**, *153*, 024109.
[37] B. Huang, O. A. von Lilienfeld, *Nat. Chem.* **2020**, *12*, 945.
[38] F. Häse, L. M. Roch, A. Aspuru-Guzik, *Chem. Sci.* **2018**, *9*, 7642.
[39] C.-H. Ding, X.-L. Hou, *Chem. Rev.* **2011**, *111*, 1914.
[40] D. Sepúlveda, T. Lu, S. E. Wheeler, *Org. Biomol. Chem.* **2014**, *12*, 8346.
[41] A. C. Doney, B. J. Rooks, T. Lu, S. E. Wheeler, *ACS Catal.* **2016**, *6*, 7948.
[42] T. P. Tauer, C. D. Sherrill, *J. Phys. Chem. A* **2005**, *109*, 10475.
[43] V. Yu, Vaganov, Y. Fukazawa, N. S. Kondratyev, S. A. Shipilovskikh, S. E. Wheeler, A. E. Rubtsov, A. V. Malkov, *Adv. Synth. Catal.* **2020**, *362*, 5467.
[44] L. A. Thiede, M. Krenn, A. Nigam, A. Aspuru-Guzik, **2020**, arXiv preprint arXiv:2012.11293v1 [cs.LG].
[45] A. Nigam, P. Friederich, M. Krenn, A. Aspuru-Guzik, **2020**, arXiv preprint arXiv:1909.11655v4 [cs.NE].
[46] Q. Vanhaelen, Y.-C. Lin, A. Zhavoronkov, *ACS Med. Chem. Lett.* **2020**, *11*, 1496.
[47] T. T. Le, W. Fu, J. H. Moore, *Bioinformatics* **2020**, *36*, 250.
[48] J. A. Hueffel, T. Sperger, I. Funes-Ardoiz, J. S. Ward, K. Rissanen, F. Schoenebeck, *Science* **2021**, *374*, 1134.

[49] L.-C. Xu, S.-Q. Zhang, X. Li, M.-J. Tang, P.-P. Xie, X. Hong, *Angew. Chem. Int. Ed.* **2021**, *60*, 22804.
[50] K. Jorner, A. Tomberg, C. Bauer, C. Sköld, P.-O. Norrby, *Nat. Chem. Rev.* **2021**, *5*, 240.
[51] Y. V. Suleimanov, W. H. Green, *J. Chem. Theory Comput.* **2015**, *11*, 4248.
[52] G. N. Simm, M. Reiher, *J. Chem. Theory Comput.* **2017**, *13*, 6108.
[53] G. N. Simm, A. C. Vaucher, M. Reiher, *J. Phys. Chem. A* **2018**, *123*, 385.
[54] T. A. Young, J. J. Silcock, A. J. Sterling, F. Duarte, *Angew. Chem. Int. Ed.* **2020**, *60*, 4266.
[55] Y. Guan, V. M. Ingman, B. J. Rooks, S. E. Wheeler, *J. Chem. Theory Comput.* **2018**, *14*, 5249.

Manuscript received: December 16, 2021

RESEARCH ARTICLE

NaviCatGA is a genetic algorithm designed to optimize molecular catalysts towards target properties. The proposed strategy is versatile and robust, as demonstrated on both transition metal and organo-catalysis applications. Evolutionary optimization using fragment databases bypasses combinatorial scaling through the incorporation of chemical constraints.



*Dr. R. Laplaza, S. Gallarati, Prof. C. Corminboeuf**

1 – 11

Genetic Optimization of Homogeneous Catalysts