doi:10.2533/chimia.2023.7 Chimia 77 (2023) 7–16 © D. Pacheco Gutierrez, L. M. Folkmann, H. Tribukait, L. M. Roch*

# How to Accelerate R&D and Optimize Experiment Planning with Machine Learning and Data Science

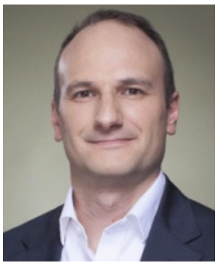
Daniel Pacheco Gutierrez, Linnea M. Folkmann, Hermann Tribukait, and Loïc M. Roch*

*Abstract:* Accelerating R&D is essential to address some of the challenges humanity is currently facing, such as achieving global sustainability goals. Today's Edisonian approach of trial-and-error still prevalent in R&D labs takes up to two decades of fundamental and applied research for new materials to reach the market. Turning around this situation calls for strategies to upgrade R&D and expedite innovation. By conducting smart experiment planning that is data-driven and guided by AI/ML, researchers can more efficiently search through the complex – often constrained – space of possible experiments and find the global optima much faster than with the current approaches. Moreover, with digitized data management, researchers will be able to maximize the utility of their data in the short and long terms with the aid of statistics, ML and visualization tools. In what follows, we present a framework and key technologies to accelerate R&D and optimize experiment planning.

**Keywords**: Artificial intelligence · Autonomous experimentation · Closed-loop optimization · Experiment planning · Machine learning · Materials acceleration platforms · Process optimization · Self-driving labs

***Daniel Pacheco Gutierrez*** is a chemical and software engineer at Atinary Technologies, where he develops cutting edge technology solutions and improvements to Atinary SDLabs platform. Daniel also helps clients to leverage the power of SDLabs platform for their needs. He was born in Colombia, received a M.Sc. degree in Computational Science and Engineering at EPFL and a BSc degree in Chemical Engineering from McGill University in Canada. Daniel is passionate about designing, building and testing digital solutions applied to process optimization in the domains of chemistry and biology.

***Linnea M. Folkmann*** is a data scientist at Atinary Technologies. She develops and tests the machine learning algorithms and data analytics module offered in the SDLabs platform. Linnea also works with clients to address any data questions. Linnea is from Denmark. She has a MSc in physical chemistry from the University of Copenhagen and a BSc in Chemistry from the University of Copenhagen. She carried out her master's project under a MARVEL Master's Fellowship at EPFL.



***Dr. Hermann Tribukait*** is an economist, entrepreneur, and leader in innovation and technology development. He is Co-Founder and CEO of Atinary Technologies based in Silicon Valley, with over 20 years of experience in banking, consulting, and technology development. Prior to launching Atinary, Dr. Tribukait co-authored and led the Clean Energy Materials Innovation Challenge of the global initiative Mission Innovation for 3 years. He holds a BA in Economics from ITAM with highest honors, and a Masters and PhD in Economics from Harvard. Dr. Tribukait is fluent in English, German, and Spanish.

***Dr. Loïc Roch*** is the co-founder and CTO of Atinary Technologies. He leads a multidisciplinary team that pushes the boundaries of ML and innovation to make experiment planning easy. Loïc's expertise includes quantum chemistry, AI, ML, and software design. He is dedicated to accelerating R&D and deploying the self-driving labs at scale. As an entrepreneur, scientist and nature-lover, Dr. Roch believes in a world where science and technology contribute to accelerating the transition to a sustainable planet and a circular economy.

## 1. Introduction

In recent years, the rapid advances in machine learning (ML) and data science have led to a proliferation of new techniques for optimizing experimental design and planning. In this article, we propose a framework for leveraging these techniques to accelerate research and development (R&D) and discovery in a variety of fields. By utilizing artificial intelligence (AI) and ML algorithms to analyze small to large datasets and identify patterns and trends, researchers can more effectively plan experiments, leading to faster and more efficient R&D and discovery. We discuss the benefits of this approach and provide an overview of the key steps involved in implementing it.

Accelerating R&D and discovery will also accelerate the transition to a sustainable circular economy.[1,2] The transition to a circular economy is an important goal for our society, as it will create a more sustainable and resilient future.[3] A circular economy is one in which we reduce waste and maximize the use of resources, by designing products and systems that can be used and reused over and over again. This is in contrast to the linear economy[4] that we

*Correspondence: Dr. L. M. Roch, E-Mail: loic@atinary.com
Atinary Technologies Sàrl, Lausanne, VD, Switzerland

currently have, in which we extract resources, use them to create products, and then dispose of those products after a 'one-time' use.

However, in order to make the transition to a circular economy, we need to accelerate R&D and discovery in key areas, such as energy storage and production, recycling, and waste valorization. This is because many of the technologies and systems that we need to support a circular economy either do not exist yet, or are not yet mature enough for widespread adoption.

All in all, accelerating R&D and materials discovery is essential for achieving global sustainability goals, such as those outlined in the United Nations' Sustainable Development Goals (SDGs).[5] Moreover, the current energy crisis in Europe as a result of the war in Ukraine, and other geopolitical conflicts, as well as global warming, make a faster transition to a sustainable circular economy a high priority.

The challenge is that discovering and developing these breakthrough materials needed to accelerate the transition to a sustainable and circular economy is a very slow and expensive process. The current Edisonian approach of trial-and-error that is still prevalent in R&D labs can take up to two decades of fundamental and applied research[6] for new materials to reach the market.[4,7,8] Simply put, we cannot afford to wait decades for these new materials to reach the market.

For the last decades, R&D productivity has been declining, despite the huge advances in science and technology. This decline, first diagnosed by Jack W. Scannell in 2012,[9] was named 'Eroom's law',[10] also referred to as the 'law of diminishing returns for research and development'. It states that the productivity of R&D decreases exponentially over time. However, the exact cause of this decline is not well understood.[79]

One possible explanation for the decline in R&D productivity is the increasing complexity of the challenges that researchers are facing. The low hanging fruits of R&D have already been picked and the R&D challenges are growing in complexity. In addition, as more data is available, we are hitting the limits of the human mind to process and understand large, multi-dimensional data.

To this growing complexity we would add that R&D processes are outdated and are often performed manually, which makes them hard to replicate, slow and expensive.

Turning around this situation calls for strategies to upgrade R&D and expedite innovation. We believe that part of the solution to Eroom's Law is upgrading and modernizing R&D. First, research labs need to digitize R&D to promote and enhance collaboration and information sharing. Second, R&D labs also need to embrace data-centric processes guided by statistics, ML and AI. For example, AI and ML can help speed up R&D and materials discovery by:

1. systematically and automatically screening for new materials that meet the necessary criteria;
2. optimizing the experimental designs and helping operators and technicians navigate the large materials space;
3. shedding light on the data collected and providing insights on the applications at hand.

In the following sections, we will review current technologies used in modern R&D facilities and list the associated challenges with these technologies. We then present how AI and ML can be used to plan experiments and to shed light on the output results to extract the key information from the data collected. By conducting smart experiment planning that is data-driven and guided by AI/ML, researchers can more efficiently search through the space of possible experiments and find the global optima much faster than with the current approaches and techniques, such as Design of Experiment (DoE), one-factor-at-a-time (OFAT) and grid search or high-throughput experimentation (HTE).

## 2. Today's State-of-of-the-art Technologies Used in Modern R&D Facilities

In the very first step for a successful experimental campaign, researchers need to define the goals: What are the optimization objectives? What are the key parameters?

Once these questions have been answered, the next step is to pick an experiment planner to guide the optimization process. Nowadays, modern R&D labs have adopted a variety of experiment planning technologies. These include popular techniques such as design of experiments (DoE), one-factor-at-a-time (OFAT) and grid search or high-throughput experimentation (HTE). While these approaches are systematic and have shown to be useful, they require a vast amount of evaluations, they have difficulties handling non-numerical data, and they also require important human intervention and tuning to be successful. In Fig. 1 we briefly review these approaches and list their main pitfalls.

### 2.1 *Design of Experiments*

Design of experiment (DoE, Fig. 1, panel A) is a systematic approach to planning and conducting experiments, introduced back in 1935, by Sir Ronald Fisher.[11,12] The goal of DoE is to identify the factors that affect a particular response, and to determine the optimal levels of these factors in order to maximize or minimize the response.[13]

In a DoE, the experimenter first identifies the factors that are expected to affect the response, and then selects the levels at which each factor will be tested. There are almost as many initial designs as chemists, with the response surface methodology (RSM)[13,14] and the Taguchi[15] methods being the most widely used. RSM is a statistical approach that can be used to model the relationship between inputs and outputs in order to identify optimal settings. Taguchi methods are a set of techniques for experimental design and data analysis that aim to minimize variability and maximize quality. DoE allows researchers to draw conclusions about the factors that affect the response, and to make predictions about how the response will behave in different conditions.

One of the main pitfalls of DoE is that it relies heavily on the choice of the initial design. As a consequence, one might miss important aspects and features of the response surface if the prior assumption is violated. Typically, once the design is selected, experiments are conducted following the said design layout, without further refinement. Additionally, the number of experiments to evaluate according to the design layout grows exponentially with the number of parameters or design factors, making it an unsuitable method for most real use cases.

### 2.2 *One-factor-at-a-time*

One-factor-at-a-time (OFAT, Fig. 1, panel B) experimentation[16] is a common approach used in R&D settings. It is one of the first strategies for designing experiments. The goal of OFAT experimentation is to identify the effect of a single factor on a response variable of interest. This is typically accomplished by holding all other factors at some fixed level and systematically varying the level of the factor of interest. While this approach may seem straightforward, there are several potential pitfalls associated with OFAT experimentation.

First, OFAT experimentation can be very time consuming and expensive if a large number of factors need to be considered. Second, because only one factor is varied at a time, it can be difficult to isolate the effect of that factor from other confounding factors. Finally, because each factor is considered in isolation, OFAT experimentation does not provide any information about potential interactions between factors, and thus does not guarantee to locate the global optimum.

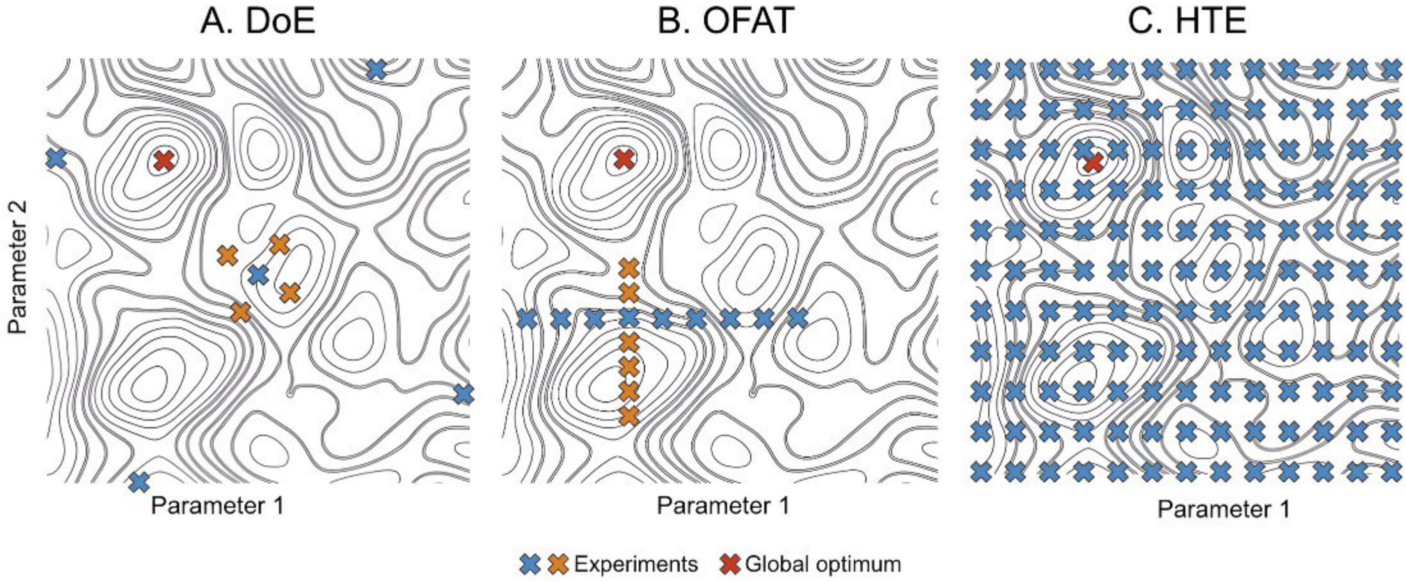


Fig. 1. Illustrations of today's state-of-the-art methodologies to plan experiments used in modern R&D facilities across life science, chemistry, materials science, biotechnology and pharma. These methodologies include: design of experiments (DoE, panel A), one-factor-at-a-time (OFAT, panel B) and high-throughput experiment (HTE, panel C). The red cross highlights the global optimum, the blue crosses shows the initial set of experiments and the orange crosses the first refinement of experiment (relevant only for DoE and OFAT).

### 2.3 *High-throughput Experimentation*

High-throughput experimentation (HTE, Fig. 1, panel C), also referred to as grid search, is a popular and widely used technique in R&D for finding the best combination of parameters for optimizing a given model. It works by systematically evaluating combinations of parameters over a grid of possible values, allowing for an exhaustive exploration of all possible parameter combinations.

The main pitfall of HTE is that it can be experimentally expensive as the number of combinations increases exponentially with the number of parameters. Often, this is referred to as the curse of dimensionality. Additionally, the results of grid search can be sensitive to the choice of parameter values, and it is difficult to determine the best parameter values without a thorough understanding of the problem.

## 3. Reversing Eroom's Law: ML as Next-generation Experiment Planning

One of the ways to achieve faster development cycles is by leveraging adaptive methods that can learn from experiments in real time, such as those offered by ML for materials science.[17,18] Compared to traditional approaches that follow a 'blind recipe' ML-driven methods attempt to perform experiments either where there is little information (exploration) or where better results are likely to occur (exploitation). Hereafter, we describe the requirements for ML frameworks to allow the generation of new designs based on existing data whereby a multi-objective function is optimized with input variables that can be continuous, discrete, and/or categorical (Fig. 2).

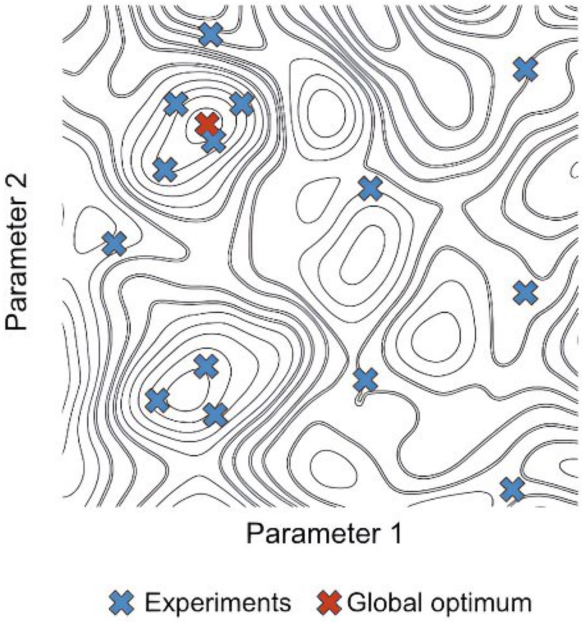


Fig. 2. Illustration of ML-driven methodologies for next-generation experiment planning. The red cross highlights the global optimum and the blue crosses shows the experiments suggested by the ML algorithm based on the information gathered throughout the optimization process. Some experiments are performed in unknown regions of the search space (exploration), but a higher density is present in the identified optimal regions of the search space (exploitation) illustrating the tradeoff used by ML-driven methods.

### 3.1 *Global Optimization with ML*

As with all optimization problems, one also needs a strategy on how to avoid getting stuck in local minima or maxima, and how to explore the design space efficiently to find the global optima. Computationally involved methods such as particle swarm,[19,20] support vector machine, simulated annealing,[21,22] or evolutionary strategies[23,24] can account for the results from recent experiments to make informed decisions about the next conditions. Bayesian optimization (BO)[25–27] recently gained traction as a promising alternative. BO is known to reduce redundancy in the proposed conditions, thereby maximizing domain knowledge while minimizing the number of objective evaluations. BO is a method often applied to black-box problems for global optimization.

BO optimization algorithms suggest the most promising experiments to run next to achieve optimization objectives according to various models. Typically, BO is very well suited to accelerate the traditional *design – build – test – learn* cycle, referred to as closed-loop optimization as shown in Fig. 3.

BO is a powerful technique for optimizing complex functions, and has been shown to be particularly effective for optimizing functions with many local optima. BO is a derivative-free optimi-

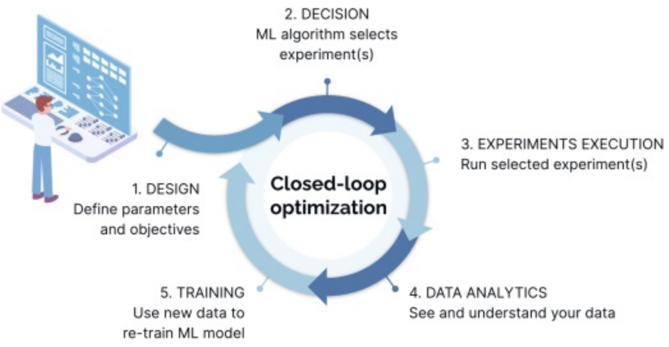

2. DECISION
ML algorithm selects experiment(s)
3. EXPERIMENTS EXECUTION
Run selected experiment(s)
1. DESIGN
Define parameters and objectives
Closed-loop optimization
5. TRAINING
Use new data to re-train ML model
4. DATA ANALYTICS
See and understand your data

Fig. 3. Representation of the data-driven closed-loop optimization for ML driven experiment planning.

zation method that uses a probabilistic model to guide the search for the global optimum. The key idea is to construct a probabilistic model to the past observations (surrogate model), and then use a selection policy (acquisition function) to decide which point to evaluate next. The acquisition function determines the tradeoff exploration and exploitation with respect to the surrogate model. The new evaluated points are then used to update the model, and the process is repeated until convergence.

The success of BO was reported in a variety of applications across chemistry, materials science and life science. For example, BO was used to accelerate the screening of molecules for organic photovoltaics,[28] discover non-fullerene acceptor candidates for organic photovoltaics,[29] discover new drug-related materials,[30] speed up synthesis planning and reaction optimization,[31] to design new photocatalyst[32] or to discover new battery electrolyte[33] to name but a few.

Recently, Atinary Technologies[34] launched its flagship BO algorithm: Falcon. Falcon is a global optimization strategy that can solve optimization problems including continuous, discrete and/or categorical variables with or without physicochemical descriptors, as well as batch-constrained optimization. Falcon comes with three flavors of surrogate models and acquisition functions (AF).

1. Falcon Light and its proprietary surrogate and AF which allows an explicit trade-off between exploration and exploitation.
2. Falcon GPBO, which uses Gaussian Process[35,36] Bayesian Optimization as a surrogate model. Typically, GPBO is well suited for optimization problems that can potentially be solved with a relatively small number of experiments. However, GPBO scales cubically with the number of experiments. Thus, the computational cost can potentially be very high if used in complex simulation cases.
3. Falcon DNGO (Deep Network for Global Optimization[37] as a surrogate model) maintains desirable properties of the Gaussian Processes (*e.g.* management of uncertainty) while improving its scalability. Specifically, unlike a standard Gaussian process, DNGO scales linearly with the number of evaluations or experiments. Falcon DNGO creates a robust, scalable, and effective Bayesian optimization system that generalizes across many global optimization problems, for a suitable set of design choices.

### 3.2 *Quantifying Performance of Global ML Algorithms*

Below we report a concise benchmark study on two real-case experiments that feature numerical and categorical parameters. Fig. 4 compares the convergence of different optimization algorithms on simulated chemical reactions, referred to as *digital twins*. These digital twins include one cross-coupling reaction (Reizman-Suzuki) and one nucleophilic aromatic substitution reaction. The digital twins were developed by Felton *et al.* in 2021. The Falcon algorithms are compared against GPyOpt[38] and Dragonfly,[39] which are popular open-source Bayesian optimization libraries for optimizing black-box functions.

The two benchmarks (Benchmark A and B, Fig. 4) are representations of typical chemical reaction optimizations carried out in R&D laboratories where the goal is to find the experimental conditions that maximize properties (or measurements) such as the yield, selectivity or conversion.

Benchmark A (Fig. 4, top panel) corresponds to a simulated nucleophilic aromatic substitution reaction performed in a plug-flow reactor. The model predicts the space-time-yield (kg m$^{-3}$ h$^{-1}$) as a function of the residence time, inlet concentration of 2,4-dinitrofluorobenzene, reactor temperature and equivalents of pyrrolidine. It is a mechanistic model that uses experimental reaction kinetics and mass balance differential equations as the starting point.

Benchmark B (Fig. 4, bottom panel) corresponds to a Reizman-Suzuki cross-coupling reaction simulation. The model predicts the yield as a function of the catalyst type (8 options), the reactor residence time, the temperature and the catalyst loading. The model is a regression function fit to experimental data as described by Felton *et al.*[40]

Benchmarks A and B show that adaptive strategies perform much better than Random Search. We also observe that FalconGPBO and FalconDNGO converge quicker than other competing methods. For both benchmark A and B, only FalconGPBO reaches the global optimum on average within the allocated budget of 100 iterations. For benchmark A, FalconGPBO requires

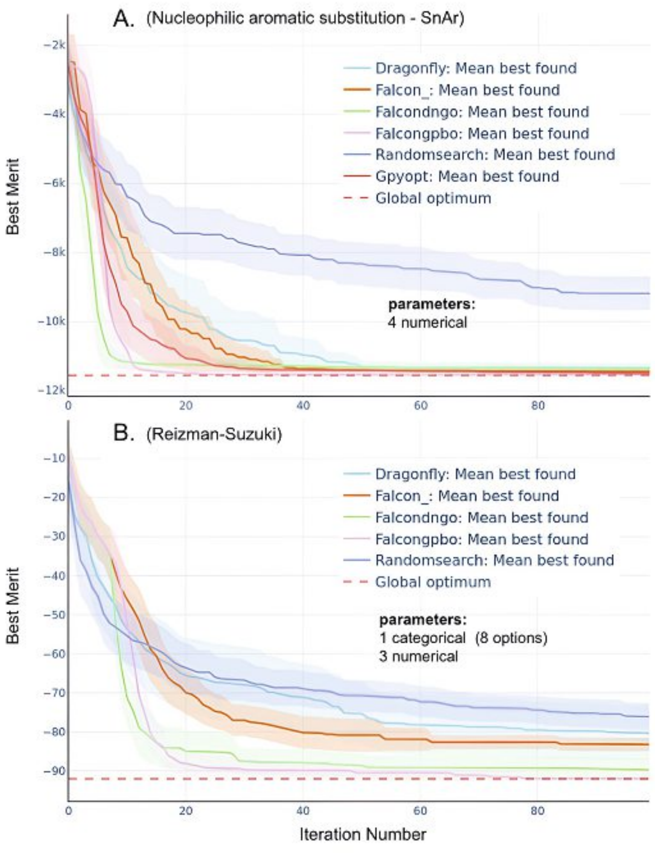


Fig. 4. Chemical reaction digital twin optimization benchmarks: Each figure reports the best merit over 30 campaigns (different starting conditions) according to six different experiment planning strategies: Random search, Gpyopt, Dragonfly, FalconGPBO, FalconDNGO and Falcon. Benchmark A only has numerical parameters and benchmark B has categorical parameters, which are not readily-supported by GPyOpt. The digital twins were initially developed by Felton *et al.*[40]

about 16 iterations on average to converge and reach the optimum set of parameters. In comparison, FalconDNGO requires about 10 iterations to converge, but the mean-best found does not get to reach the global optimum. GPyOpt is the third best, requiring about 25 iterations to converge. In contrast, Dragonfly converges after 50 iterations. A similar ranking is observed for benchmark B except that Falcon-Light occupies the third place. Apart from the convergence rate, it is also possible to compare the strategies with respect to their overall best-found value after 100 iterations: FalconGPBO, FalconDNGO and FalconLight are the closest to benchmark B's global optimum respectively. As benchmark B features categorical variables, GPyOpt was not included as it does not readily support this parameter type. Categorical variables are a challenge for ML algorithms as they lack order and cannot be represented by numerical values. This makes it difficult for ML algorithms to interpret and use categorical data effectively.

Recently, we reported a comparison of DoE versus BO in the optimization of perovskite solar cells' efficiency.[41] In this project, Atinary's proprietary BO algorithm, Falcon, needed to explore less than 10% of the total number of possible combinations to hit the optimal conditions. Falcon maximized the solar cells' efficiency in only a fraction of the experiments and increased productivity by 45%, compared to DoE.

Another important aspect for R&D to fully embrace ML algorithms is to ensure the algorithm's robustness towards noise in the experiments.[42,43] Robustness significantly affects the location of the optima and the exploration of the search space. For illustration purposes, below we report another succinct benchmark study demonstrating the impact of noise on four selected ML optimizers offered on Atinary SDLabs: Random Search, HTE, FalconDNGO and FalconGPBO.

In the examples that follow, we use Dejong[44] and Levy[45] (Fig. 5) analytical functions to demonstrate the robustness of state-of-the-art technologies used in modern R&D labs and BO. The noisy function is constructed by adding a random noise drawn from a Gaussian distribution to each measurement (evaluated Benchmark function value). The Gaussian distribution has mean equal to zero and a standard deviation (scale) of σ % of the noiseless function value, where σ is {0, 5, 10, 15, 30}. The noise level in Fig. 6 below refers to this standard deviation of this Gaussian distribution. For each surface, Fig. 6 shows the minimum noise-free merit (best found) after 50 noisy function evaluations.

Not surprisingly, the performance of grid search or HTE and Random Search is independent of the noise level, as they both follow a random sampling strategy. In both cases, the optimizer is not dependent on the value of prior observations. In sharp contrast, Fig. 6 shows that the performance of the BO ML algorithms – FalconGPBO and FalconDNGO – varies with the level of noise in the measurements. Both BO optimizers learn from previous experiments. However, the more noisy the measurements are, the more challenging it is for these optimizers to find a parameter region near the optimal function value.

Even at a 30% noise level in the Dejong function, both FalconGPBO and FalconDNGO still perform significantly better than HTE and Random Search. Using the more complicated surface with the Levy function, the mean performance of FalconDNGO is constant with increasing noise, despite the noise dependance. Moreover, the performance of the ML optimizer FalconDNGO increases going from 0% to 10% noise level. This artifact has already been reported in earlier studies: adding random Gaussian noise to the desired signal during training of deep neural networks – like the surrogate model of FalconDNGO – helps avoid overfitting to the observations and improves the optimization perfor-

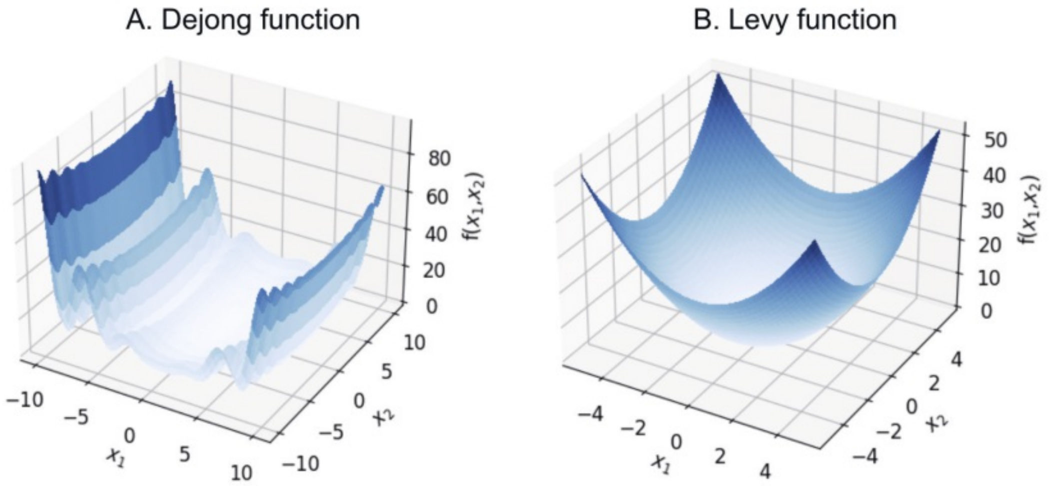


Fig. 5. Illustration of the 2D Dejong (left) and Levy (right) benchmark functions. Both functions have a global minimum at $f(x_1,x_2) = f(0;0) = 0$.

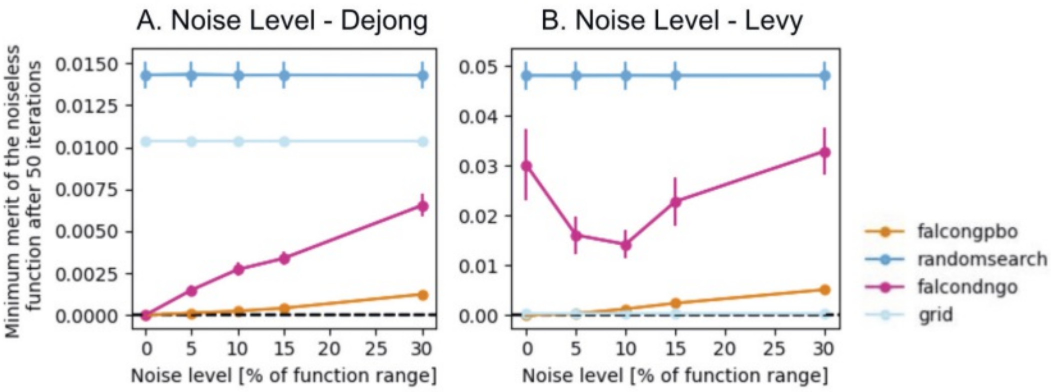


Fig. 6. Minimum merit of the noiseless function after 50 iterations of learning as a function of the scale of the Gaussian noise in the noisy measurements (left: 1D Dejong function and right: 1D Levy function). Each data point is the mean of 1,000 runs initialized with different random starting points. The vertical lines on each data point show the error bars. For some data points the error bar is smaller than the point. The black horizontal dotted line marks the optimal function value.

mance.[46,47] This behavior was seen to be particularly effective for complex optimization landscapes where adding noise can help overcome local minima.

These plots also show how simple models such as grid search (or HTE) can be quite powerful at low-dimensional problems (see right panel, Fig. 6). The benchmarks above were done with the 1D version of the functions, which gave very few grid-points to evaluate. The benchmark below shows the performance of the same algorithms with functions that have increasingly more dimensions.

Fig. 7 shows that for both synthetic functions – Dejong (left) and Levy (right) – the ML optimizers perform consistently better than grid and random-search as a function of increasing dimensionality. While the random search and HTE approaches might be suitable for simple problems, most real-world problems will require smarter strategies that leverage the learned information.

### 3.3 *Beyond Global Optimization*

To further speed up discovery, it is of utmost importance to develop accurate transfer learning methods that can leverage pre-existing knowledge and databases. Efficient transfer learning strategies require that only the relevant information is extracted from pre-existing experimental results. This will allow the generation of 'data-driven chemical intuition' that can be generalized and used in new applications. Such a transfer learning approach would allow researchers to construct informative priors for new applications based on results from other applications.

These transfer-learning methods need to be an integral part of the closed-loop optimization process. Importantly, these methods should (i) be method or surrogate agnostic, (ii) select automatically useful sources of prior information and (iii) head start optimization to explore only promising regions without biasing the search.

Several methods for transfer learning have been published in the ML literature. These include a compound acquisition function, multi-task Gaussian processes[48–52] from BO,[53] shape and shrink the initial search space *a priori* to replace initial sampling, or meta-learning and few-shot learning.[54] Recently, we have demonstrated that our transfer-learning algorithm – Atinary SeMOpt – can systematically improve the knowledge extracted from prior information by combining neural processes for meta-learning with a compound acquisition function.[55,56]

Last but not least, real-world optimization problems are very often limited by multiple constraints. For example, the optimization of a chemical process may exhibit known physical and/or manufacturing constraints. These constraints have a direct impact on the parameter space of the optimization problem by effectively reducing its volume to only the subset of points that fulfill all the constraints (*i.e.* the feasible regions). Therefore, optimization techniques that lack the ability to incorporate this knowledge or constraints are inefficient and may choose parameter points that are outside the feasible regions.

Recently, we reported a new ML algorithm, Atinary™ Emmental, that allows users to add a set of constraints in optimization problems. Emmental uses the constraints to define the feasible regions in the parameter space, and avoid the non-feasible regions when solving the optimization problem (Fig. 8). Emmental is compatible with Atinary's proprietary suite of ML algorithms available on SDLabs, and is available through the SDLabs graphical user interface (GUI) or application programming interface (API) or software development kit (SDK).

Compared to popular Bayesian optimization algorithms, such as GPyOpt or Botorch,[57] Emmental can handle a broader variety of non-linear optimization problems with multiple parameters and constraints. Specifically, Emmental supports constrained optimization problems involving continuous, discrete and/or categorical parameters in combination with the following constraint types: (i) Exclusion constraints, which specify ranges of intermediate values that continuous/discrete parameters must avoid; (ii) Conditional-exclusion constraints, which define relationships between multiple parameters that are to be avoided; (iii) Inequality and equality constraints, which indicate linear and non-linear relations between parameters.

Atinary Falcon and Atinary Emmental, together with a suite of open-source ML algorithms, are made available through Atinary's no-code ML platform SDLabs. The platform allows users to define their experiments and start planning experiments *via* either its GUI or its API/SDK. Fig. 9 illustrates these means of interactions.

## 4. The Importance of Data Interpretability

Not only is it important for the ML algorithm and optimization approach to find the local and global optima, but the ML algorithm also needs to maximize knowledge acquisition in order for data to be useful and interpretable. The ML optimizer must be able to interpret the data and draw conclusions and make predictions based on that data. The integration of data science with chemistry and materials science is of utmost importance to further accelerate discovery and innovation and to speed up the transition towards sustainable manufacturing.[58–60]

There are many factors that affect the interpretability of data. The most important factor is the quality of the data itself. If the data is of poor quality, it will be very difficult to interpret. The second factor is the amount of data available. The more data there is, the easier it will be to interpret. Finally, the methods used to analyze and visualize the data can also affect its interpretability.[58] Atinary SDLabs allows users to interact with their data through its 'Analytics' module, which provides a menu of eight plots to facilitate and streamline data interpretability.

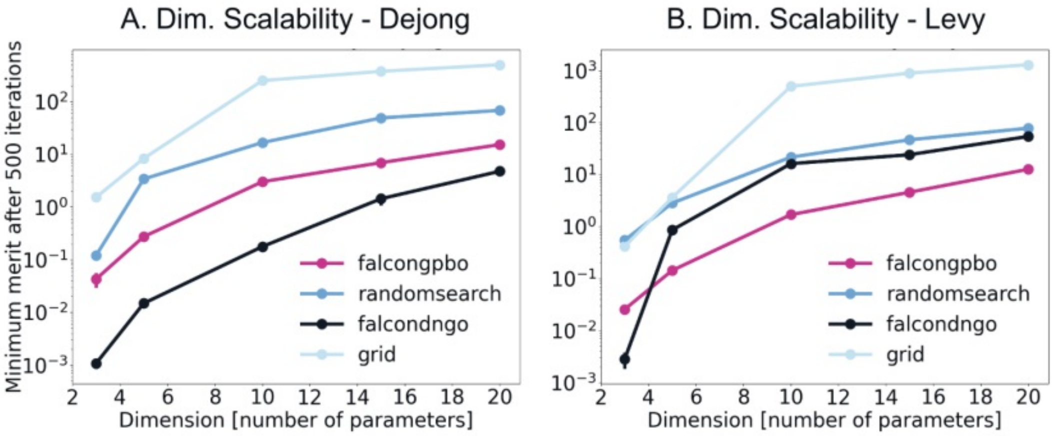


Fig. 7. Dimension scalability benchmark: Minimum merit reached after 500 iterations of learning as a function of the number of parameters to optimize. Each data point is the mean of 10 runs initialized with different random seeds. The vertical lines on each data point show the error bars. For some data points the error bar is smaller than the point.

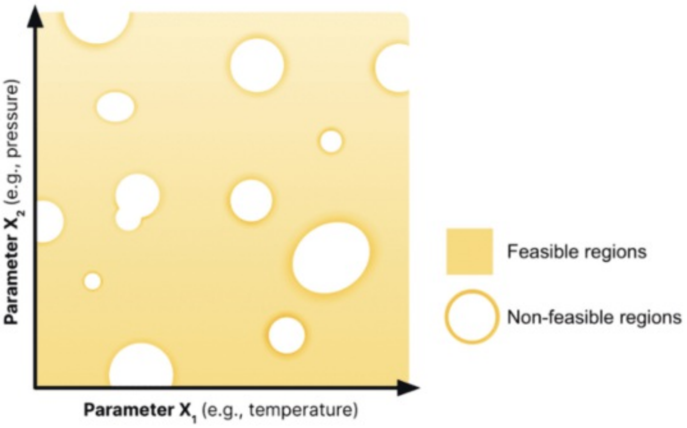

Parameter $X_2$ (e.g., pressure)
Parameter $X_1$ (e.g., temperature)
Feasible regions
Non-feasible regions

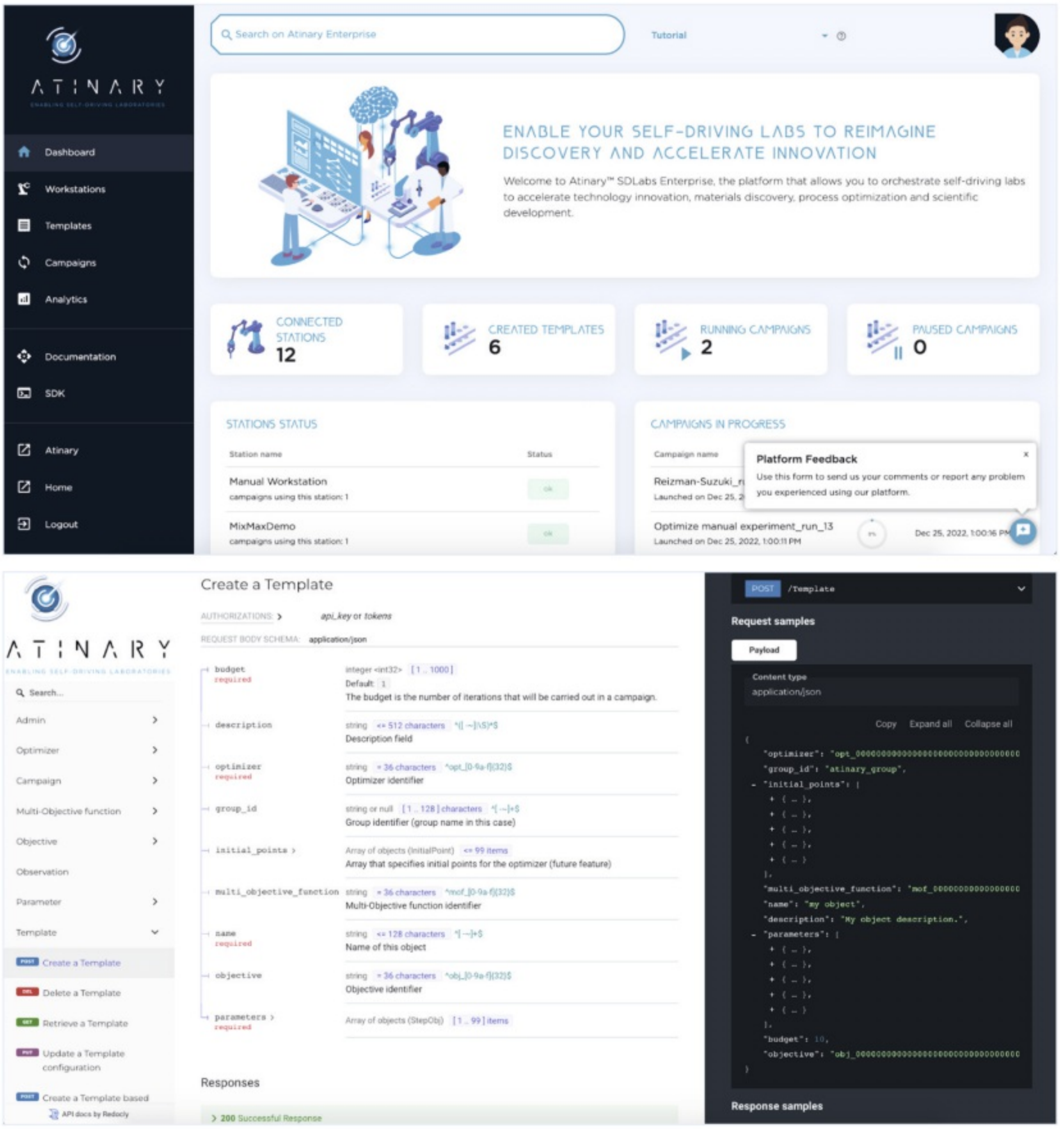
Fig. 8. Representation of Atinary™ Emmental for constrained optimization.

The importance of interpretability cannot be overstated. Data that cannot be interpreted is effectively useless. Interpretability is essential for researchers and businesses to be able to draw useful conclusions and make informed decisions. Interpretability helps scientists, engineers, and other professionals to make sense of data, allowing them to identify trends and anomalies in the data and take appropriate action.[59]

Sensitivity analysis is one such data interpretability tool. It consists of studying the effect that input variability has on the output variables. Through sensitivity analysis,[61,62] researchers can identify the most influential parameters in a system. It helps to identify which inputs are the most important in a model, and can be used to determine the optimal set of parameters and to identify areas that need further investigation. Sensitivity analysis can also be used to identify potential areas of model uncertainty, which can help to improve the accuracy and reliability of a model. This is a crucial element in the understanding of an application in order to refine the search space without affecting the level of quality, safety and efficiency in the process at hand.

Another critical aspect of enforcing data interpretability is the ability to construct 'maps' towards design control. Through advanced methods for dimensionality reduction, one can represent high dimensional space into a human readable format to assure that the application meets user needs, intended uses, and specified requirements. The figure below (Fig. 10) shows an example of a nine-dimensional parameter space reduced to a two-dimensional space. The color scheme displays the evaluation of a particular ML regression model as a function of the input parameters. The figure displays the nonlinear nature of what the ML model 'sees', with its local minima/maxima and valleys.

Apart from visualizing the regression models, it is also possible to use these models to simulate a given process. This simulator is known as a 'digital twin' which can serve as the platform



ENABLE YOUR SELF-DRIVING LABS TO REIMAGINE DISCOVERY AND ACCELERATE INNOVATION

Welcome to Atinary™ SDLabs Enterprise, the platform that allows you to orchestrate self-driving labs to accelerate technology innovation, materials discovery, process optimization and scientific development.

CONNECTED STATIONS 12

CREATED TEMPLATES 6

RUNNING CAMPAIGNS 2

PAUSED CAMPAIGNS 0

STATIONS STATUS

| Station name | Status |
|---|---|
| Manual Workstation<br>campaigns using this station: 1 | ok |
| MixMaxDemo<br>campaigns using this station: 1 | ok |

CAMPAIGNS IN PROGRESS

| Campaign name | | |
|---|---|---|
| Reizman-Suzuki_r<br>Launched on Dec 25, 2 | | |
| Optimize manual experiment_run_13<br>Launched on Dec 25, 2022, 1:00:11 PM | 0% | Dec 25, 2022, 1:00:16 PM |

Platform Feedback

Use this form to send us your comments or report any problem you experienced using our platform.

Create a Template

AUTHORIZATIONS: api_key or tokens

REQUEST BODY SCHEMA: application/json

| | |
|---|---|
| budget<br>required | integer <int32> [1 .. 1000]<br>Default: 1<br>The budget is the number of iterations that will be carried out in a campaign. |
| description | string <= 512 characters ^([ -~]\S)*$<br>Description field |
| optimizer<br>required | string = 36 characters ^opt_[0-9a-f]{32}$<br>Optimizer identifier |
| group_id | string or null [1 .. 128] characters ^[ -~]+$<br>Group identifier (group name in this case) |
| initial_points > | Array of objects (InitialPoint) <= 99 items<br>Array that specifies initial points for the optimizer (future feature) |
| multi_objective_function | string = 36 characters ^mof_[0-9a-f]{32}$<br>Multi-Objective function identifier |
| name<br>required | string <= 128 characters ^[ -~]+$<br>Name of this object |
| objective | string = 36 characters ^obj_[0-9a-f]{32}$<br>Objective identifier |
| parameters ><br>required | Array of objects (StepObj) [1 .. 99] items |

Responses

> 200 Successful Response

POST /Template

Request samples

Payload

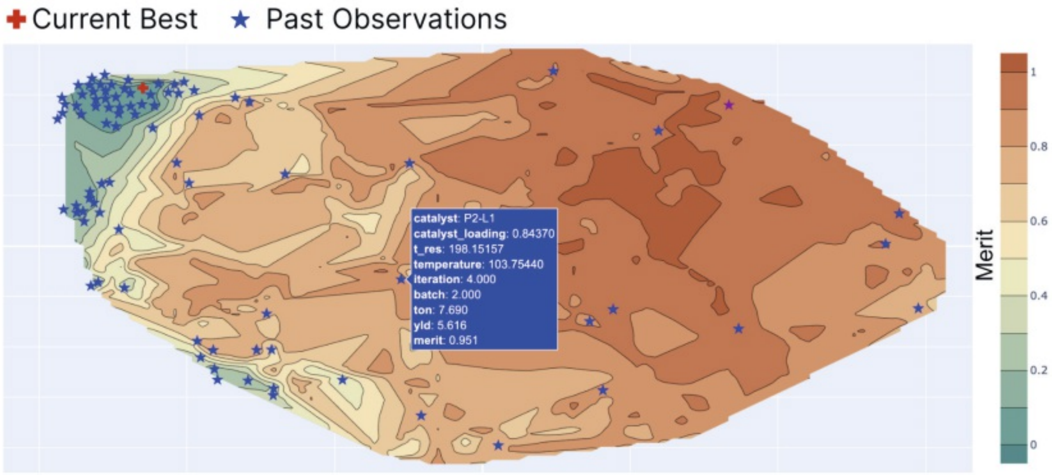
Content type
application/json

Copy Expand all Collapse all

```
{
   "optimizer": "opt_00000000000000000000000000000000
   "group_id": "atinary_group",
 - "initial_points": [
    + { … },
    + { … },
    + { … },
    + { … },
    + { … }
   ],
   "multi_objective_function": "mof_0000000000000000
   "name": "my object",
   "description": "My object description.",
 - "parameters": [
    + { … },
    + { … },
    + { … },
    + { … },
    + { … }
   ],
   "budget": 10,
   "objective": "obj_00000000000000000000000000000000
}
```

Response samples

Fig. 9. Screenshots of the Graphical User Interface (GUI, top) and application programming interface definition (API, bottom) of Atinary SDLabs.



catalyst: P2-L1
catalyst_loading: 0.84370
t_res: 198.15157
temperature: 103.75440
iteration: 4.000
batch: 2.000
ton: 7.690
yld: 5.616
merit: 0.951

Fig. 10. Interpretation of the application at hand from the ML algorithm. Dimensionality reduction from 6D to 2D towards design control for process scale-up.

to perform further digital experiments, benchmarks and calculations. Such digital twins can serve as proxies to real experiments, and can help users better leverage further R&D efforts.

The path to go beyond data interpretability and increase the adoption of data infrastructures to further maximize value extraction from the data is enumerated by the FAIR data principle, which is a set of guidelines for how data should be managed and shared.[63] 'FAIR' stands for Findable, Accessible, Interoperable, and Reusable, and it refers to the four guiding principles for effective data management and sharing:

- *Findable* requires that data should be discoverable, usually through a searchable repository, and that its metadata associated with the data are sufficient for other users to understand its context and potential uses.
- *Accessibility* means that data should be available in a format that is open and machine-readable, and should have appropriate licenses to ensure that the data can be used without any legal restrictions.
- *Interoperability* requires that data be structured and formatted in such a way that it can be accessed and used in combination with other data sources.
- *Reusability* means that data should be formatted in a way that makes it easy to be used for a variety of purposes.

The FAIR data principle is becoming increasingly important. It serves as the foundation of effective data management and sharing, ultimately allowing data to reach its fullest potential. Such a common communication stream is crucial at all levels to streamline sharing massive amounts of data. Although the FAIR principle establishes the foundation towards data management and sharing, one of the challenges lies in the importance for the scientific community to develop standards in order to maximize participation and adoption of new data infrastructure.[58]

## 5. The Self-driving Labs: Integrated Platform where Cutting-edge Technologies Meet

Self-driving labs™ are next-generation R&D labs. They leverage automated equipment along with the power of AI and ML to efficiently identify target candidate materials/molecules, in a closed-loop fashion. They have the potential to significantly improve the speed and accuracy of scientific experimentation and data collection, and revolutionize the discovery of new materials.[64–67]

By integrating the latest technological innovations in automation, robotics and computer science with current approaches in chemistry, materials synthesis and characterization, self-driving labs will act as a catalyst for revolutionizing the way research and development is conducted in both industry and academia.[33,68–75] The use of self-driving labs can also lead to faster and more efficient solutions to global challenges, helping to achieve the United Nations SDGs. However, these labs are still in their infancy and need to be more widely adopted in order to fully realize their potential.

However, on the hardware front, standardization – which is essential to embrace new technologies – is still a challenge as each laboratory may have different requirements for its operations. Various frameworks, such as SiLA, provide a standard framework for the integration of laboratory instruments and data, enabling the sharing and interoperability of laboratory data and processes. Such frameworks would catalyze the development, deployment, and maintenance of automated laboratory systems.

One of the key advantages of self-driving labs is that they can significantly improve the speed and accuracy of scientific experimentation.[76,77] By automating many of the tedious and repetitive tasks involved in research, these labs can help researchers focus on more complex and innovative work. Additionally, the use of AI and ML technology can help identify patterns and trends that might not be immediately apparent to human researchers, leading to new insights and discoveries.[76]

Self-driving labs have the potential to accelerate and revolutionize the discovery of new materials, which is essential for advancing many fields of science and technology.[78] By automating the process of materials synthesis and characterization, these labs can explore a wider range of potential materials more quickly and efficiently. This could lead to the development of new and improved materials for use in a wide range of applications, from renewable energy to medical treatments.

It is up to the scientific community to take the lead in advancing the development and deployment of self-driving labs in order to tackle the world's most pressing problems and work towards achieving the SDGs. This will require collaboration among all players in academia, industry, governments and society as a whole. Global challenges require global actions and global solutions.

## 6. Conclusions and Outlook

The convergence of key technologies, such as artificial intelligence, machine learning, robotics and data science enables a drastic reduction in the time and cost necessary to identify new molecules, materials and process parameters compared to traditional trial-and-error approaches.

The use of advanced tools such as Bayesian optimization, transfer learning, constrained optimization, and data science can help to optimize experiment planning and accelerate R&D and discovery of advanced materials and molecules in a number of ways. Machine learning algorithms can help identify patterns and trends in data that might not be immediately apparent to human researchers. This can enable researchers to make better and more informed decisions about how to design and conduct experiments, leading to better results and more accurate data. Additionally, the use of machine learning algorithms can help explore a larger piece of the vast and complex materials and molecular space that has not been explored yet, leading to new and unexpected insights.

However, to successfully deploy AI/ML in R&D labs, ML experts, lab scientists and domain experts must jointly define achievable use cases where the applications of ML can be used in real experiments. Such ML technology needs to be tested or benchmarked alongside the existing technology it aims to replace or complement. This, in order to demonstrate that ML can improve the discovery and development process in the industry, while also lowering failure rates and costs.

Self-driving labs, also referred to as materials acceleration platforms or autonomous labs, are a new type of R&D laboratories that fully integrate these tools and augment researchers to execute data-driven experimentation. Augmenting researchers with the power of these technologies and automating many of the tedious and repetitive tasks involved in experiment planning and execution, allows researchers to focus on more complex and innovative work.

Overall, the integration of state-of-the-art technologies into the R&D process can significantly accelerate the pace of scientific research and drive the development of new materials and molecules, as well as new technologies that can address global challenges, such as climate change, energy poverty, and the circular economy. By leveraging the power of these technologies, researchers can more quickly and efficiently identify solutions to some of the world's most pressing problems, helping to create a more sustainable and equitable future for all.

***Acknowledgements***

The authors would like to thank Alberto García-Durán and Riley Hickman for their contribution to fruitful discussions. All experiments, benchmarks and computations were carried out on the Exoscale data centers in Switzerland.

Received: December 13, 2022

[1] W. R. Stahel, *Nature* **2016**, *531*, 435, https://doi.org/10.1038/531435a.
[2] A. Stefanakis, I. Nikolaou, 'Circular Economy and Sustainability', Vol. 2: 'Environmental Engineering', Elsevier, **2021**.
[3] 21st Century Challenges: https://21stcenturychallenges.org/
[4] F. Sariatli, *Visegrad J. Bioecon. Sust. Dev*. **2017**, *6*, 31, https://doi.org/10.1515/vjbsd-2017-0005
[5] The SDGs are a call for action by all countries to promote prosperity while protecting the planet: https://www.un.org/sustainabledevelopment/
[6] E. Maine, E. Garnsey, *Res. Policy* **2006**, *35*, 375, https://doi.org/10.1016/j.respol.2005.12.006.
[7] T. P. Hughes 'American Genesis: A Century of Invention and Technological Enthusiasm, 1870-1970', Viking Press; **1989**.
[8] D. Bishop, E. Gill, *J. R. Soc. Med.* **2020**, *113*, 79, https://doi.org/10.1177/0141076820902625.
[9] J. W. Scannell, A. Blanckley, H. Boldon, B. Warrington, *Nat. Rev. Drug Discov.* **2012**, *11*, 191, https://doi.org/10.1038/nrd3681.
[10] Eroom's law is Moore's law spelled backwards. Moore's Law describes the exponential increase in the number of transistors that can be placed onto an integrated circuit. This number doubled every 2 years from the 1970s. The term is used more generally for technologies that improve exponentially over time.
[11] J. F. Box, *Am. Stat*. **1980**, *34*, 1, https://doi.org/10.2307/2682986.
[12] F. Yates, *Biometrics* **1964**, *20*, 307, https://doi.org/10.2307/2528399.
[13] J. C. Stanley, *Am. Ed. Res. J.* **1966**. *3*, 223, https://doi.org/10.3102/00028312003003223.
[14] R. F. Gunst, R. H. Myers, D. C. Montgomery, *Technometrics* **1996**, *38*, 285, https://doi.org/10.2307/1270613.
[15] G. Taguchi, 'Introduction to Quality Engineering: Designing Quality Into Products and Processes', Quality Resources, **1986**.
[16] C. Daniel, *J. Am. Stat. Assoc*. **1973**, *68*, 353, https://doi.org/10.1080/01621459.1973.10482433.
[17] R. S. Sutton, A. G. Barto, *IEEE Trans Neural Networks* **1998**, *9*, 1054, https://doi.org/10.1109/TNN.1998.712192.
[18] M. Ghavamzadeh, S. Mannor, J. Pineau, A. Tamar, *Mach. Learn*. **2015**, *8*, 359, https://doi.org/10.1561/2200000049.
[19] R. Eberhart, J. Kennedy, 'A new optimizer using particle swarm theory', in 'MHS'95 Proceedings of the Sixth International Symposium on Micro Machine and Human Science', **1995**, pp. 39-43.
[20] Y. Shi, R. Eberhart, 'A modified particle swarm optimizer', in 'IEEE International Conference on Evolutionary Computation Proceedings IEEE World Congress on Computational Intelligence (Cat No98TH8360)', IEEE; **2002**, https://doi.org/10.1109/ICEC.1998.699146.
[21] S. Kirkpatrick, C. D. Gelatt Jr, M. P. Vecchi, *Science* **1983**, *220*, 671, https://doi.org/10.1126/science.220.4598.671
[22] V. Černý, *J. Optim. Theory Appl.* **1985**, *45*, 41, https://doi.org/10.1007/BF00940812.
[23] N. Hansen, A. Ostermeier, *Evol. Comput*. **2001**, *9*, 159, https://doi.org/10.1162/106365601750190398.
[24] N. Hansen, S. D. Müller, P. Koumoutsakos, *Evol. Comput.* **2003**, *11*, 1, https://doi.org/10.1162/106365603321828970.
[25] J. Mockus, 'The Bayesian approach to global optimization. System Modeling and Optimization', Berlin/Heidelberg: Springer-Verlag, **2005,** pp. 473-481.
[26] H. J. Kushner, *J. Basic Eng*. **1964**, *86*, 97, https://doi.org/10.1115/1.3653121.
[27] T. Agrawal, 'Bayesian Optimization. Hyperparameter Optimization in Machine Learning', Springer, **2021**, pp. 81-108, https://doi.org/10.1007/978-1-4842-6579-6_4
[28] E. O. Pyzer-Knapp, G. N. Simm, A. Aspuru Guzik, *Mater. Horiz.* **2016**, *3*, 226, https://doi.org/10.1039/C5MH00282F.
[29] F. Häse, M. Aldeghi, R. J. Hickman, L. M. Roch, A. Aspuru-Guzik, *Appl. Phys. Rev*. **2021**, *8*, 031406, https://doi.org/10.1063/5.0048164.
[30] D. Reker, 'Active learning for drug discovery and automated data curation', in 'Artificial Intelligence in Drug Discovery', Chap. 4, Cambridge: Royal Society of Chemistry, **2020**, pp. 301-326, https://doi.org/10.1039/9781788016841-00301
[31] 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.
[32] B. Burger, P. M. Maffettone, V. V. Gusev, C. M. Aitchison, Y. Bai, X. Wang, X. Li, B. M. Alston, B. Li, R. Clowes, N. Rankin, B. Harris, R. S. Sprick, A. I. Cooper, *Nature* **2020**, *583*, 237, https://doi.org/10.1038/s41586-020-2442-2
[33] A. Dave, J. Mitchell, K. Kandasamy, H. Wang, S. Burke, B. Paria, B. Poczos, J. Whitacre, V. Viswanathan, *Cell Rep. Phys. Sci*. **2020**, *1*, 100264, https://doi.org/10.1016/j.xcrp.2020.100264.
[34] Atinary Technologies is a machine learning startup offering a no-code ML platform to accelerate R&D: www.atinary.com
[35] J. Snoek, H. Larochelle, R. P. Adams, *Adv Neural Inf. Proc. Syst*. **2012**, 25, https://proceedings.neurips.cc/paper/4522-practical-bayesian-optimization
[36] C. M. Bishop, 'Pattern Recognition and Machine Learning', Springer, **2006**.
[37] J. Snoek, O. Rippel, K. Swersky, R. Kiros, N. Satish, N. Sundaram, Md. M. A. Patwary, R. P. Adams, 'Scalable Bayesian optimization using deep neural networks', **2015**, https://doi.org/10.48550/arXiv.1502.05700.
[38] J. González, Z. Dai, 'GPyOpt: a Bayesian optimization framework in Python', Accessed, **2016**.
[39] K. Kandasamy, K. R. Vysyaraju, W. Neiswanger, B. Paria, C. R. Collins, J. Schneider, **2019**, https://doi.org/10.48550/ARXIV.1903.06694
[40] K. C. Felton, J. G. Rittig, A. A. Lapkin, *Chem. Meth.* **2021**, *1*, 116, https://doi.org/10.1002/cmtd.202000051.
[41] Press release available on LinkedIn: https://www.linkedin.com/feed/update/urn:li:activity:6963449379466723328.
[42] R. J. Hickman, M. Aldeghi, F. Häse, A. Aspuru-Guzik, *Digital Discov*. **2022**, *1*, 732, https://doi.org/10.1039/D2DD00028H.
[43] M. Aldeghi, F. Häse, R. J. Hickman, I. Tamblyn, A. Aspuru-Guzik, *Chem. Sci.* **2021**, *12*, 14792, https://doi.org/10.1039/D1SC01545A.
[44] Available at: https://www.sfu.ca/~ssurjano/dejong5.html.
[45] Available at: https://www.sfu.ca/~ssurjano/levy.html.
[46] A. Neelakantan, L. Vilnis, Q. V. Le, I. Sutskever, L. Kaiser, K. Kurach, J. Martens, 'Adding gradient noise improves learning for very deep networks', **2015**, https://doi.org/10.48550/ARXIV.1511.06807.
[47] C. Wang, J. C. Principe, *IEEE Trans. Neural Netw.* **1999**, *10*, 1511, https://doi.org/10.1109/72.809097.
[48] K. Yu, V. Tresp, A. Schwaighofer, 'Learning Gaussian processes from multiple tasks', in 'Proceedings of the 22nd international conference on Machine learning - ICML '05', New York, ACM Press, **2005**, https://doi.org/10.1145/1102351.1102479.
[49] E. V. Bonilla, K. Chai, C. Williams, 'Multi-task Gaussian Process Prediction', in 'Advances in Neural Information Processing Systems', Eds. J. Platt, D. Koller, Y. Singer, S. Roweis, Curran Associates, Inc.; **2007**, https://proceedings.neurips.cc/paper/2007/file/66368270ffd51418ec58bd793f2d9b1b-Paper.pdf

[50] K. Hayashi, T. Takenouchi, R. Tomioka, H. Kashima, *Trans. Jpn. Soc. Artif. Intell.* **2012**, *27*, 103, https://doi.org/10.1527/tjsai.27.103.

[51] B. Rakitsch, C. Lippert, K. Borgwardt, O. Stegle, *Adv. Neural Inf. Process Syst*. **2013**, *26*, https://proceedings.neurips.cc/paper/2013/hash/59c33016884a62116be975a9bb8257e3-Ab stract.html.

[52] J. Zhu, S. Sun, in 'Chinese Conference on Pattern Recognition', Communications in Computer and Information Science Book Series, Springer Berlin Heidelberg; **2014**, pp. 54-62, https://doi.org/10.1007/978-3-662-45646-0_6

[53] K. Swersky, J. Snoek, R. P. Adams, *Adv. Neural Inf. Process Syst*. **2013**, 26, https://proceedings.neurips.cc/paper/2013/hash/f33ba15effa5c10e873bf3842afb46a6-Abstr act.html

[54] M. Wistuba, J. Grabocka, 'Few-shot Bayesian optimization with deep kernel surrogates', **2021**, https://doi.org/10.48550/ARXIV.2101.07667.

[55] R. Hickman, J. Ruza, L. Roch, H. Tribukait, A. García-Durán, 'Equipping data-driven experiment planning for Self-driving Laboratories with semantic memory: case studies of transfer learning in chemical reaction optimization' *ChemRxiv* **2022**, https://doi.org/10.26434/chemrxiv-2022-jt4sm.

[56] V. Shekar, V. Yu, B. J. Garcia, D. B. Gordon, G. E. Moran, D. M. Blei, L. M. Roch, A. Garcia-Duran, M. A. Najeeb, M. Zeile, P. W. Nega, Z. Li, M. A. Kim, E. M. Chan, A. J. Norquist, S. Friedler, J. Schrier, 'Serendipity based recommender system for perovskites material discovery: balancing exploration and exploitation across multiple models' **2022**, https://doi.org/10.26434/chemrxiv-2022-l1wpf

[57] M. Balandat, B. Karrer, D. R. Jiang, S. Daulton, B. Letham, A. G. Wilson, E. Bakshy, 'BoTorch: A framework for efficient Monte-Carlo Bayesian optimization', **2019**, https://doi.org/10.48550/ARXIV.1910.06403.

[58] L. Himanen, A. Geurts, A. S. Foster, P. Rinke, *Adv. Sci.* **2019**, *6*, 1900808, https://doi.org/10.1002/advs.201900808.

[59] J. Yano, K. J. Gaffney, J. Gregoire, L. Hung, A. Ourmazd, J. Schrier, J. A. Sethian, F. M. Toma, *Nat. Rev. Chem*. **2022**, *6*, 357, https://doi.org/10.1038/s41570-022-00382-w.

[60] F. Delgado-Licona, M. Abolhasani, *Adv. Intell. Syst*. **2022**, 2200331, https://doi.org/10.1002/aisy.202200331.

[61] A. Saltelli, *Risk Anal.* **2002**, *22*, 579, https://doi.org/10.1111/0272-4332.00040.

[62] A. A. Saltelli, M. Ratto, T. Andres, F. Campolongo, J. Cariboni, D. Gatelli, M. Saisana, S. Tarantola, 'Global Sensitivity Analysis: The Primer', Hoboken, NJ: Wiley-Blackwell, **2008**, https://doi.org/10.1002/9780470725184

[63] M. D. Wilkinson, M. Dumontier, I. J. J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, J.-W. Boiten, L. Bonino da Silva Santos, P. E. Bourne, J. Bouwman, A. J. Brookes, T. Clark, M. Crosas, I. Dillo, O. Dumon, S. Edmunds, C. T. Evelo, R. Finkers, A. Gonzalez-Beltran, A. J. G. Gray, P. Groth, C. Goble, J. S. Grethe, J. Heringa, P. A. C.'t Joen, R. Hooft, T. Kuhn, R. Kok, J. Kok, S. J. Lusher, M. E. Martone, A. Mons, A. L. Packer, B. Persson, P. Rocca-Serra, M. Roos, R. van Schaik, S.-A. Sansone, E. Schultes, T. Sengstag, T. Slater, G. Strawn, M. A. Swertz, M. Thompson, J. van der Lei, E. van Mulligen, J. Velterop, A. Waagmeester, P. Wittenburg, K. Wolstencroft, J. Zhao, B. Mons, *Sci. Data* **2016**, *3*, 160018, https://doi.org/10.1038/sdata.2016.18.

[64] F. Häse, L. M. Roch, A. Aspuru-Guzik, *Trends Chem.* **2019**, *1*, 282, https://doi.org/10.1016/j.trechm.2019.02.007.

[65] D. P. Tabor, L. M. Roch, S. K. Saikin, C. Kreisbeck, D. Sheberla, J. H. Montoya, S. Dwaraknath, M. Aykol, C. Ortiz, H. Tribukait, C. Amador-Bella, C. J. Brabec, B. Maruyama, K. A. Persson, A. Aspuru-Guzik, *Nat. Rev. Mater.* **2018**, *3*, 5, https://doi.org/10.1038/s41578-018-0005-z.

[66] T. C. Wu, A. Aguilar-Granda, K. Hotta, S. A. Yazdani, R. Pollice, J. Vestfrid, H. Hao, C. Lavigne, M. Seifrid, N. Angello, F. Bencheikh, J. E. Hein, M. Burke, C. Adachi, A. Aspuru-Guzik, *Adv. Mater*. **2022**, e2207070, https://doi.org/10.1002/adma.202207070.

[67] H. G. Martin, T. Radivojevic, J. Zucker, K. Bouchard, J. Sustarich, S. Peisert, D. Arnold, N. Hillson, G. Babnigg, J. M. Marti, C. J. Mungall, G. T. Beckham, L. Waldburger, J. Carothers, S. Sundaram, D. Agarwal, B. A. Simmons, T. Backman, D. Banerjee, D. Tanjore, A. Singh, *Curr. Opin. Biotechnol.* **2022**, *79*, 102881, https://doi.org/10.1016/j.copbio.2022.102881.

[68] 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*, https://doi.org/10.1038/s42004-021-00550-x

[69] B. P. MacLeod, F. G. L. Parlane, T. D. Morrissey, F. Häse, L. M. Roch, K. E. Dettelbach, R. Moreira, L. P. E. Yunker, M. B. Rooney, J. R. Deeth, V. Lai, G. J. Ng, H. Situ, R. H. Zhang, M. S. Elliot, T. H. Haley, D. J. Dvorak, A. Aspuru-Guzik, J. E. Hein, C. P. Berlinguette, *Sci. Adv*. **2020**, *6*, eaaz8867, https://doi.org/10.1126/sciadv.aaz8867.

[70] D. Becker, C. Schmitt, L. Bovetto, C. Rauh, C. McHardy, C. Hartmann, *Innov. Food Sci. Emerg. Technol*. **2023**, *83*, 103232, https://doi.org/10.1016/j.ifset.2022.103232.

[71] A. Dave, J. Mitchell, S. Burke, H. Lin, J. Whitacre, V. Viswanathan, *Nat. Commun.* **2022**, *13*, 5454, https://doi.org/10.1038/s41467-022-32938-1.

[72] P. Nikolaev, D. Hooper, F. Webber, R. Rao, K. Decker, M. Krein, J. Poleski, R. Barto, B. Maruyama, *Npj Comput. Mater*. **2016**, *2*, 16031, https://doi.org/10.1038/npjcompumats.2016.31

[73] Y. Jiang, D. Salley, A. Sharma, G. Keenan, M. Mullin, L. Cronin, *Sci. Adv*. **2022**, *8*, eabo2626, https://doi.org/10.1126/sciadv.abo2626

[74] M. B. Rooney, B. P. MacLeod, R. Oldford, Z. J. Thompson, K. L. White, J. Tungjunyatham, B. J. Stankiewicz, C. P. Berlinguette, *Digital Discov*. **2022**, *1*, 382, https://doi.org/10.1039/D2DD00029F.

[75] B. P. MacLeod, F. G. L. Parlane, C. C. Rupnow, K. E. Dettelbach, M. S. Elliott, T. D. Morrissey, T. H. Haley, O. Proskurin, M. B. Rooney, N. Taherimakhsousi, D. J. Dvorak, H. N. Chiu, C. E. B. Waizenegger, K. Ocean, M. Mokhtari, C. P. Berlinguette, *Nat. Commun.* **2022**, *13*, 995, https://doi.org/10.1038/s41467-022-28580-6.

[76] A. Aspuru-Guzik, K. Persson, 'Materials Acceleration Platform: Accelerating Advanced Energy Materials Discovery by Integrating High-Throughput Methods and Artificial Intelligence', **2018**, https://dash.harvard.edu/bitstream/handle/1/35164974/233.pdf?sequence=1.

[77] L. M. Roch, F. Häse, C. Kreisbeck, T. Tamayo-Mendoza, L. P. E. Yunker, J. E. Hein, A. Aspuru-Guzik, *Sci. Robot.* **2018**, *3*, https://doi.org/10.1126/scirobotics.aat5559.

[78] S. G. Baird, T. D. Sparks, *Matter* **2022**, *5*, 4170, https://doi.org/10.1016/j.matt.2022.11.007.

[79] T. Cowen, B. Southwood, 'Is the Rate of Scientific Progress Slowing Down?', GMU Working Paper in Economics No. 21-31, Aug. 5, **2019**, http://dx.doi.org/10.2139/ssm.3822691.

### *License and Terms*

The definitive version of this article is the electronic one that can be found at https://doi.org/10.2533/chimia.2023.7