[Connor W. Coley](http://cwcoley.com/) (B.S., Caltech; Ph.D., MIT; currently Assistant Professor of Chemical Engineering at MIT)

[Natalie S. Eyke](http://greengroup.mit.edu/natalie-eyke) (B.S., University of Michigan; working on Ph.D. at MIT)

[Klavs F. Jensen](https://jensenlab.mit.edu/) (Professor, Chemical Engineering, MIT)

The importance of automation in chemical science experiments lies in its ability to speed up the discovery process with few errors and human intervention. Automation provides inputs such as a new synthetic process, model, or molecule, and outputs which may include the outputs streamlining drug’s performance on chemical assays, or the model’s predictive performance. This research is important because human-guided drug discovery research and prone to errors, and though a computer model cannot identify a good therapeutic, automation can speed up certain aspects of drug discovery.

The authors discuss various types of discovery in the chemical sciences space, and recommend ways to unify them as searches in a high dimensional design space. They further define these discoveries into three fields: matter, processes, and models, and recommended a framework of questions to guide the evaluate the degree to which a discovery can be credited to automation or autonomy. The authors provide key steps to assess automation in the search discovery process, and propose that automatic discoveries are most valid if 1) the goal is broad or at its highest level (i.e. will meet a therapeutic need), 2) the researchers constrain the search and design space, 3) the validation and feedback selection process involves a small search space and cheap experiments in order to limit human intervention, 4) the design space includes intelligence in the search strategy (i.e. continuous design spaces with a quantifiable meaning not simply found through a brute force search), 5) the experiments have a system that can automatically collect information to support or refute a hypothesis —narrow experimental designs can mitigate an automatic validation/feedback process, 6) the results are structured in a way the contributes to previous scientific knowledge, and (optional) 7) the outcome contributes to overall scientific knowledge.

**The authors define the term discovery into three classifications:**

1. **Physical Matter**

Physical matter discoveries include that of new molecules, materials and devices, and tend to output new chemical matter that would become a part of a therapeutic application. For example, search spaces for physical matter are predetermined and . The discovery of new physical matter can be iterative, noniterative, model based, or include active learning and/or virtual screening.

1. **Processes**

Processes can be abstract such as pasteurization, or concrete like organic molecules or synthetic routes.

1. **Models**

Models are defined in the paper to include empirical, symbolic regressions, theories, natural laws, and structure-function relationships, and typically are a part of the discovery process and not the sole discovery. For examples, for natural laws that can explain observed phenomena that previous theories could not fully explain (i.e. Mendeleev’s periodic table).

**The authors argue that most discoveries are a form of a search problem:**

**1. Physical matter:** Molecular discovery is a search within a “chemical space” and has a combinatorial design space with theoretical molecules. The authors purport that materials discoveries are also combinatorial design spaces with structural compositions.

**2. Processes:** Discovering a chemical or physical is the result of a search for design spaces with process variables and/or operation sequences. The authors provide a chemical reaction optimization example where a reaction temperature or a solvent might need to be changed.

**3. Models:** The authors suggest that most models are empirical relationships that come from data fitting, and have defined search spaces. For example, symbolic regressions are derived in a combinatorial space with input variables and mathematical operations.

The automated discoveries product causal explanations from simplified terms and well-defined search spaces.

**The authors provide a framework for autonomous discovery by considering the following questions:**

**1. How broadly is the goal defined?**

In the autonomous discovery process the goal should have a broader application such as for a therapeutic purpose. If not, is the goal narrow in scope that it can optimize another process that has an impact on humans. Researchers can constrain a search space by limiting the experimental and computational capabilities of an autonomous platform.

**2. How constrained is the search/design space?**

Most researchers operate in an unconstrained search space which lends to errors—less human intervention can mitigate the automation process. Though some experiments may have too many limitations on the autonomous platform that can interfere with having the necessary automated validations.

**3. How are experiments for validation/feedback selected?**

Experiments that require few high level human decisions and have a restrained search space are cheap can facilitate the autonomous process. Cummings and Bruni regarding the levels of automation e.g. collaboration with humans and the computer. One level includes humans making all the local decisions and the computer narrows down any selections and executes the functions, and another level includes the computational algorithms suggesting assessments for a certain compound that a human needs to manually synthesize.

**4. How superior to a brute force search is navigation of the design space?**

The design space should have some level of intelligence that is beyond that of a brute-force search. For example, when evaluating a library of compounds, a brute force search requires that use of the entire design space, while an active learning screening may only require a small percentage of the experiment to derive an optimal selection.

**5. How are experiments for validation/feedback performed?**

Gathering data to support or refute a hypothesis is important can be done entirely by a human, computer, or a combination of the two. Though, The more narrow an experiment design space is, the likelier that the validation/feedback process can be automated.

**6. How are results organized and interpreted?**

In an iterative workflow, automation can organize and interpret experiment results to improve upon and contribute to current literature.

**7. Does the discovery outcome contribute to broader scientific knowledge?**

The authors discuss how intelligibility and impact are criteria for autonomous discovery to advance the existing catalogue of scientific knowledge. The studies listed in Table 1 focus on discovery campaigns with a primary outcome of identifying a molecule, material, or set of process conditions that achieve or optimize a human-defined performance objective. Autonomous discovery can be used to specify a high-level goal (e.g., find compounds that react strongly in a given set of assays) to receive human-interpretable, generalized conclusions that can advance the drug discovery process. This reduces the mental burden for humans and speeds up the process.

**The authors discuss the enabling factors that contribute to autonomous discovery**

1. **Data availability:** Digitization of the varied and expansive knowledge available in academic journals, public databases etc. have facilitated the ability to search for information and derive new analyses. In order to utilize the existing literature, machine learning can assist in compiling large amounts of data and literature to
2. **Computational tools:** Machine learning algorithmsare necessary to make computational recommendations and can discover patterns in the available literature and create parameters in a process or model.
3. **Hardware automation:** Experimental capabilities can automate workflows in the lab

**The authors outline a few examples of discoveries that have partial automation including:**

**1. Discovery for Pharmaceutical Applications**

The authors also discuss the contribution of machine learning in the pharmaceutical field, particularly with Eve an active learning strategy platform which has the capacity to screen specific compound libraries against different assays at >10,000 compounds a day. The platform is highly selective in the screening process rather than expanding the search design.

**2. Discovery of New Synthetic Pathways**

The discovery process of new synthetic pathways (precursor to producing a molecule) involves using machine learning such as learned value functions and action policies to respectively estimate the synthetics of a reactant molecule and find shorter pathways, and predict the transformation of the molecule using prior literature.

**Conclusion**

Autonomous scientific discovery has yet to fully manifest particularly in the chemical sciences. The limitations of experiments, it can by improve the levels of outputs and limit the burden of human intervention in manual tasks. The authors’ guidelines for autonomous scientific discovery can help researchers accelerate workflows in the laboratory and eventually produce therapeutics drugs and other human-based applications. Since few discoveries have full an autonomous process, human intervention is still necessary. While few studies identify the search space definition for autonomous discovery requires significant human input. The human developer’s role is still significant -- to formulate the discovery problem, an effective representation, and transform the output into meaningful scientific results.

Overall, this paper can inform Emergent.Bio on what may constitute an autonomous discovery through provided guidelines and examples of previous scientific work that uses automation in biomedical/biochemical models. This work relates to Emergent.Bio as the company seeks to predict markers by querying markers rather than behaviors. An empirical computer engine could run experiments and simulations (based on public information available from drug discovery databases, for example) and crunch the numbers to yield novel insights and advance the field. This work could occur in conjunction with research labs to speed up drug discovery and help accelerate drug discovery.