A Journal of the Gesellschaft Deutscher Chemiker

Angewandte International Edition GDCh Chemie

www.angewandte.org

Accepted Article

Title: Autonomous discovery in the chemical sciences part II: Outlook

Authors: Connor W Coley, Natalie S Eyke, and Klavs F. Jensen

To be cited as: *Angew. Chem. Int. Ed.* 10.1002/anie.201909989
Angew. Chem. 10.1002/ange.201909989

Link to VoR: http://dx.doi.org/10.1002/anie.201909989
http://dx.doi.org/10.1002/ange.201909989

WILEY-VCH

10.1002/anie.201909989

Autonomous discovery in the chemical sciences part II: Outlook

Connor W. Coley*†, Natalie S. Eyke*, Klavs F. Jensen*‡

Keywords: automation, chemoinformatics, machine learning, drug discovery, materials science

Author bios:

Connor W. Coley completed his B.S. in chemical engineering at the California Institute of Technology and his M.S.C.E.P. and Ph.D. in chemical engineering at the Massachusetts Institute of Technology (MIT). His research focuses on how data science and laboratory automation can be used to streamline discovery in the chemical sciences.

Natalie S. Eyke completed her B.S. in chemical engineering at the University of Michigan in 2014. After graduating, she joined the Chemical Engineering Research & Development department at Merck & Co., Inc., where she worked on process development for small molecule pharmaceuticals. In 2017, she began a Ph.D. in chemical engineering at the Massachusetts Institute of Technology (MIT), where she works for Professors Klavs F. Jensen and William H. Green. Her research focuses on combining active machine learning and high-throughput experimentation to facilitate reaction screening.

Klavs F. Jensen is the Warren K. Lewis Professor in Chemical Engineering and Materials Science and Engineering at the Massachusetts Institute of Technology. He is a co-director of MIT's Pharma AI consortium that aims to bring machine learning technology into pharmaceutical discovery and development. He received his MSc in Chemical Engineering from the Technical University of Denmark (DTU) and his Ph.D. in chemical engineering from the University of Wisconsin-Madison. His research interests include on-demand multistep synthesis, methods for automated synthesis, and machine learning techniques for chemical synthesis and interpreting large chemical data sets. Catalysis, chemical kinetics and transport phenomena are also topics of interest along with development of methods for predicting performance of reactive chemical systems.

*Department of Chemical Engineering, Massachusetts Institute of Technology, Cambridge, MA 02139
†ccoley@mit.edu
‡kfjensen@mit.edu

Contents

1 Abstract **3**

2 Reflection on case studies **3**

3 Challenges and trends **9**
3.1 Working with complex data . . . 9
3.1.1 Creating and maintaining datasets . . . 9
C: Establish open access databases with standardized data representations . . . 12
C: Address the inconsistent quality of existing data . . . 13
3.1.2 Building empirical models . . . 14
C: Improve representations of molecules and materials . . . 14
C: Improve empirical modeling performance in low-data environments . . . 15
C: Incorporate physical invariance and equivariance properties . . . 15
C: Unify and utilize heterogeneous datasets . . . 16
C: Improve interpretability of machine learning models . . . 17
3.2 Automated validation and feedback . . . 17
3.2.1 Experimental validation . . . 18
C: Expand the scope of automatable experiments . . . 18
C: Facilitate integration through systems engineering . . . 19
C: Automate the planning of multistep chemical syntheses . . . 20
3.2.2 Computational validation . . . 21
C: Accelerate code/software used for computational validation . . . 21
C: Broaden capabilites / applicability of first-principles calculations . . . 22
3.2.3 Shared challenges . . . 22
C: Ensure that validation reflects the real application . . . 22
C: Lower the cost of automated validation . . . 23
C: Combine newly acquired data with prior literature data . . . 23
3.3 Selection of experiments for validation and feedback . . . 24
C: Quantify model uncertainty and domain of applicability . . . 24
C: Quantify the tradeoff between experimental difficulty and information gain . . . 25
C: Define discovery goals at a higher level . . . 26
3.3.1 Proposing molecules and materials . . . 27
C: Bias generative models towards synthetic accessibility . . . 27
C: Benchmark problems for molecular generative models . . . 28
3.4 Evaluation . . . 29
C: Demonstrate extrapolative power of predictive models . . . 29
C: Demonstrate design-make-test beyond proof-of-concept . . . 30
C: Develop benchmark problems for discovery . . . 30

4 Conclusion **31**
4.1 Attribution of discovery to automation . . . 31
4.2 Changing definitions of discovery . . . 31
4.3 Role of the human . . . 32
4.4 Outlook . . . 33

5 Acknowledgements **34**

1 Abstract

This two-part review examines how automation has contributed to different aspects of discovery in the chemical sciences. In this second part, we reflect on a selection of exemplary studies. It is increasingly important to articulate what the role of automation and computation has been in the scientific process and how that has or has not accelerated discovery. One can argue that even the best automated systems have yet to "discover" despite being incredibly useful as laboratory assistants. We must carefully consider how they have been and can be applied to future problems of chemical discovery in order to effectively design and interact with future autonomous platforms.

The majority of this article defines a large set of open research directions, including improving our ability to work with complex data, build empirical models, automate both physical and computational experiments for validation, select experiments, and evaluate whether we are making progress toward the ultimate goal of autonomous discovery. Addressing these practical and methodological challenges will greatly advance the extent to which autonomous systems can make meaningful discoveries.

2 Reflection on case studies

In 2009, King et al. proposed a hypothetical, independent robot scientist that "automatically originates hypotheses to explain observations, devises experiments to test these hypotheses, physically runs the experiments by using laboratory robotics, interprets the results, and then repeats the cycle" [1]. To what extent have we closed the gap toward each component of this workflow, and what challenges remain?

The case studies in Part 1 illustrate many examples of the progress that has been made toward achieving machine autonomy in discovery. Several studies in particular, summarized in Table 1, represent what we consider to be exemplars of different discovery paradigms [2–15]. These include the successful execution of experimental and computational workflows as well as the successful implementation of automated experimental selection and belief revision. There are a great number of studies (some of which are described in part one) that follow the paradigm of (a) train a surrogate QSAR/QSPR model on existing data, (b) computationally design a molecule or material to optimize predicted performance, and (c) manually validate a few compounds. The table intentionally underrepresents such studies, as we believe iterative validation to be a distinguishing feature of autonomous workflows compared to "merely" automated calculations.

We encourage the reader to reflect on these case studies through the lens of the questions we proposed for assessing autonomous discovery: (i) How broadly is the goal defined? (ii) How constrained is the search/design space? (iii) How are experiments for validation/feedback selected? (iv) How superior to

a brute force search is navigation of the design space? (v) How are experiments for validation/feedback performed? (vi) How are results organized and interpreted? (vii) Does the discovery outcome contribute to broader scientific knowledge?

The goals of discovery are defined narrowly in most studies. We are not able to request that a platform identify a good therapeutic, come up with an interesting material, uncover a new reaction, or propose an interesting model. Instead, in most studies described to date, an expert defines a specific scalar performance objective that an algorithm tries to optimize. Out of the examples in Table 1, Kangas et al. has the highest-level goals: in one of their active learning evaluations, the goal could be described as finding strong activity for *any* of the 20,000 compounds against *any* of the 177 assays. While Adam attempts to find relationships between genes and the enzymes they encode (discover a causal model), it does so from a very small pool of hypotheses for the sake of compatibility with existing deletion mutants for a single yeast strain.

The search spaces used in these studies vary widely in terms of the constraints that are imposed upon them. Some are restricted out of necessity to ensure validation is automatable (e.g., Eve, Adam, Desai et al., ARES) or convenient (e.g., Weber et al., Fang et al.). Others constrain the search space to a greater degree than automated validation requires. This includes reductions of dimensionality by holding process parameters constant (e.g., Reizman et al., Ada) or the size of discrete candidate spaces (e.g., Gómez-Bombarelli et al., Thornton et al., Janet et al.). Computational studies that minimize constraints on their search (e.g., RMG, Segler et al.) do so under the assumption that the results of validation (e.g., simulation results, predictions from surrogate models) will be accurate across the full design space.

In all cases, human operators have implicitly or explicitly assumed that a good solution can be found in these restricted spaces. The extent to which domain expertise or prior knowledge is needed to establish the design space also varies. Molecular or materials design in relatively unbounded search spaces (e.g., Segler et al.) requires the least human input. Fixed candidate libraries that are small (e.g., Weber et al., Desai et al.) or derived from an expert-defined focused enumeration (e.g., RMG, Gómez-Bombarelli et al., Janet et al.) require significant application-specific domain knowledge; larger fixed candidate libraries may be application-agnostic (e.g., diverse screening libraries in Eve, Fang et al., Thornton et al.). Limiting process parameters (e.g., Reizman et al., ARES, Ada) require more general knowledge about typical parameter ranges where optima may lie.

The third question regarding experiment selection is one where the field has excelled. There are many frameworks for quantifying the value of an experiment in model-guided experimental design, both when optimizing for performance and when optimizing for information [17]. However, active learning with formal consideration of uncertainty from either a frequentist perspective [18] (e.g., Eve, Reizman et al.) or a Bayesian perspective [19, 20] (e.g., Ada) is less common than with *ad hoc* definitions of experimental diversity meant

to encourage exploration (e.g., Desai et al., Kangas et al.). Both are less common than greedy selection criteria (e.g., ARES, RMG, Gómez-Bombarelli et al., Thornton et al.). Model-free experiment selection, including the use of genetic algorithms (e.g., Weber et al., Janet et al.), is also quite prevalent but requires some additional overhead from domain experts to determine allowable mutations within the design space. When validation is not automated (e.g., Fang et al. or Gómez-Bombarelli et al.'s experimental follow-up), the selection of experiments is best described as pseudo-greedy, where the top predicted candidates are manually evaluated for practical factors like synthesizability.

The benefit of computer-assisted experiment selection is a function of the size of the design space and, when applicable, the initialization required.. In many cases, a brute force exploration of the design space is not prohibitively expensive (e.g., Eve, Adam, Desai et al., Janet et al.), although this is harder to quantify when some of the design variables are continuous (e.g., Reizman et al., ARES, Ada). Other design spaces are discrete but virtually infinite (e.g., RMG, Segler et al.), which makes the notion of a brute force search ill-defined. Regardless of whether the full design space can be screened, we can still achieve a reduction in the number of experiments required to find high-performing candidates, perhaps by a modest factor of 2-10 (e.g., Eve, Desai et al., Gómez-Bombarelli et al., Janet et al.) or even by 2-3 orders of magnitude (e.g., Weber et al., Kangas et al., Thornton et al.). It's possible that the experimental efficiency of some of these studies could be improved by reducing the number of experiments needed to initialize the workflow (e.g., Eve, Gómez-Bombarelli et al.).

The manner in which experiments for validation are performed depends on the nature of the design space and, of course, whether experiments are physical or computational. Examples where validation is automated are intentionally overrepresented in this review; there are *many* more examples of partially-autonomous discovery in which models prioritize experiments that are manually performed (e.g., similar to Weber et al., Fang et al.). There are also cases where *almost* all aspects of experimentation are automated but a few manual operations remain (e.g., transferring well plates for Adam). In computational workflows, one can often construct pipelines to automate calculations (e.g., RMG, Gómez-Bombarelli et al., Segler et al., Thornton et al., Janet et al.). In experimental workflows, one can programmatically set process conditions with tailor-made platforms (e.g., Desai et al., ARES, Ada) or robotically perform assays using in-stock compounds (e.g., Eve, Adam) or ones synthesized on-demand (e.g., Desai et al.). Pool-based active learning strategies lend themselves to retrospective validation, where an "experiment" simply reveals a previous result not known to the algorithm (e.g., Kangas et al.); this is trivially automated and thus attractive for method development. Note that in the workflow schematic for Kangas et al. in Table 1, we illustrate the revelation of PubChem measurements as experiments.

In iterative workflows, automating the organization and interpretation of results is a practical step toward

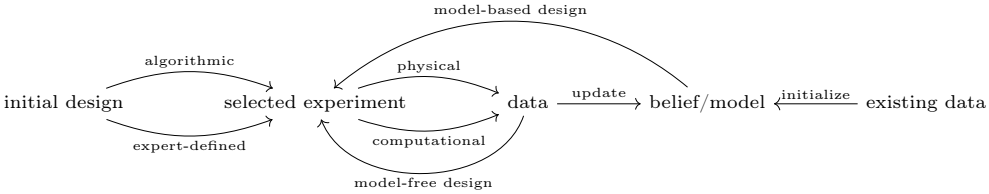
automating the subsequent selection of experiments. When workflows only proceed through a few iterations of batched experiments, humans may remain in the loop to simplify the logistics of organizing results and initializing each round (e.g., Gómez-Bombarelli et al., Thornton et al.), but nothing fundamentally prevents this step from being automated. When many iterations are required or expected, it behooves us to ensure that the results of experiments can be directly interpreted; otherwise, tens (e.g., Desai et al., Reizman et al., Segler et al., Janet et al., Ada) or hundreds (e.g., Eve, ARES, Kangas et al.) of interventions by human operators would be required. This can be the case when iterative experimental design is used with manual experimentation and analysis (e.g., the 20 iterations of a genetic algorithm conducted manually by Weber et al.). In non-iterative workflows with manual validation (e.g., Fang et al.), there is little benefit to automating the interpretation of new data. Relative to automating experiments and data acquisition, automating the interpretation thereof is rarely an obstacle to autonomy. Exceptions to this include cases where novel physical matter (a molecule or material) is synthesized and characterized (e.g., case studies in Part 1 related to experimental reaction discovery), where further advances in computer-aided structural elucidation (CASE) [21] are needed.

Our final question when assessing autonomy is whether the discovery outcome contributes to broader scientific knowledge. In Table 1, with the exception of Adam and RMG, we have focused on the discovery of physical matter or processes rather than models. The primary outcome of these discovery campaigns is the identification of a molecule, material, or set of process conditions that achieves or optimizes a human-defined performance objective. Workflows with model-based experimental designs (e.g., all but Weber et al. and Janet et al., who use genetic algorithms) have the secondary outcome of a surrogate model, which may or may not lend itself to interpretation. However, the point of this question is whether the contribution to broader scientific knowledge came *directly* from the autonomous platform, not through an *ex post facto* analysis by domain experts. These discoveries generally require manual interpretation, again excepting Adam, RMG, and similar platforms where what is discovered is part of a causal model.

Our first and last questions represent lofty goals in autonomous discovery: we specify high-level goals and receive human-interpretable, generalized conclusions *beyond* the identification of a molecule, material, device, process, or black box model. However, we have made tremendous progress in offloading both the manual and mental burden of navigating design spaces through computational experimental design and automated validation. We often impose constraints on design spaces to avoid unproductive exploration and focus the search on what we believe to be plausible candidates. To widen a design space requires that experiments remain automatable–less of a challenge for computational experiments than for physical ones–but may decrease the need for subject matter expertise and may increase the odds that the platform identifies an unexpected or superior result. Well-established frameworks of active learning and Bayesian optimization

10.1002/anie.201909989

have served us well for experimental selection, while new techniques in deep generative modeling have opened up opportunities for exploring virtually-infinite design spaces.



Reference	Discovery	Initialization	Design space	Data generation	Notes	Workflow
Eve [2]	bioactive, selective molecules	4,800 random compounds from design space	fixed library of 14,400 compounds	automated measurement of yeast growth curves	compound screening from a fixed library with an active search	
Adam [3]	gene-enzyme relationships	random experiment	15 open reading frame deletions	automated auxotrophy experiments	narrow hypothesis space, but nearly-closed-loop experimentation	
Weber et al. [4]	thrombin inhibitors	20 random compounds	virtual $10 \times 10 \times 40 \times 40$ compound library	manual synthesis and inhibition assay	iterative optimization using a genetic algorithm; design space defined by 4-component Ugi reaction to ensure synthesizability	
Desai et al. [5]	kinase inhibitors	random compound	27×10 candidates in make-on-demand library	automated microfluidic synthesis and biological testing	closed-loop synthesis and biological testing; narrow chemical design space	
Reizman et al. [6]	reaction conditions	algorithmic D-optimal design	concentration, temperature, time, 8 catalysts	automated synthesis and yield quantitation	closed-loop reaction optimization through a screening phase and an iterative phase	
ARES [7, 16]	carbon nanotube growth conditions	84 expert-defined experimental conditions	process conditions (temperature, pressures, and gas compositions)	automated nanotube growth and characterization	complex experimentation; uses RF model for regression	
Kangas et al. [8]	bioactive compounds against many assays	384 random measurements from design space	177 assay $\times$ 20,000 compound interactions	*simulated* experiments by revealing PubChem measurements	validation of pool-based active learning framework through retrospective analysis; iterative batches of 384 experiments	
RMG [9]	detailed gas-phase kinetic mechanisms	reaction conditions, optionally seeded by known mechanism	elementary reactions following expert-defined reaction templates	estimation of thermodynamic and kinetic parameters	iterative addition of hypothesized elementary reactions to a kinetic model based on simulations using intermediate models	
Fang et al. [10]	neuroprotective compounds	activity data from ChEMBL	in-house library of 28k candidates	none	literature-trained QSAR model applied to noniterative virtual screening with manual *in vitro* validation	
Gómez-Bombarelli et al. [11]	organic light-emitting diode molecules	40k random compounds from design space	virtual library of 1.6 M enumerated compounds	DFT calculations	iteratively selected batches of 40k calculations; manually validated a small number of compounds experimentally	
Segler et al. [12]	bioactive compounds	1.4M molecules from ChEMBL	all of chemical space	surrogate QSAR/QSPR models of activity	iteratively refined generative LSTM model (pretrained on ChEMBL) on active molecules identified via sampling an initial 100k + 8 rounds $\times$ 10k molecules	
Thornton et al. [13]	hydrogen storage materials	200 human-chosen subset and 200 random subset of search space	850k structures (Materials Genome)	grand canonical Monte Carlo simulations	few rounds of greedy optimization with batches of 1000 using surrogate QSPR model	
Janet et al. [14]	spin-state splitting inorganic complexes	random complexes from design space	708 ligand combinations $\times$ 8 transition metals	ANN surrogate model prediction	used genetic algorithm and computational evaluation for iterative optimization; relies on ANN pretrained on 2690 DFT calculations	
Ada [15]	organic hole transport materials	algorithmic	dopant ratio and annealing time	automated synthesis and analysis of thin film	complex experimentation successfully automated; simple design space	

Table 1: Selected examples of discovery accelerated by automation or computer assistance. The stages of the discovery workflow employed by each are shown as red arrows corresponding to the schematic above. Workflows may begin either with an initial set of experiments to run or by initializing a model with existing (external) data. Algorithmic initial designs include the selection of random experiments from the design space.

3 Challenges and trends

The capabilities required for autonomous discovery are coming together rapidly. This section emphasizes what we see as key remaining challenges associated with working with complex data, automating validation and feedback, selecting experiments, and evaluation.

3.1 Working with complex data

The discovery of complex phenomena requires a tight connection between knowledge and data [22]. A 1991 article laments the "growing gap between data generation and data understanding" and the great potential for knowledge discovery from databases [23]. While we continue to generate new data at an increasing rate, we have also dramatically improved our ability to make sense of complex datasets through new algorithms and advances in computing power.

We intentionally use "complex data" rather than "big data"–the latter generally refers only to the size or volume of data, and not its content. Here, we mean "complex data" when it would be difficult or impossible for a human to identify the same relationships or conclusions as an algorithm. This may be due to the size of the dataset (e.g., millions of bioactivity measurements), the lack of structure (e.g., journal articles), or the dimensionality (e.g., a regression of multidimensional process parameters).

Complex datasets come in many forms and have inspired an array of different algorithms for making sense of them (and leveraging them for discovery). Unstructured data can be mined and converted into structured data [24, 25] or directly analyzed as text, e.g. to develop hypotheses about new functional materials [26]. Empirical models can be generated and used to draw inferences about factors that influence complex chemical reactivity [27–30]. Virtually *any* dataset of (input, output) pairs describing a performance metric of a molecule, material, or process can serve as the basis for supervised learning of a surrogate model. Likewise, unsupervised techniques can be used to infer the structure of complex datasets [31–33] and form the basis of deep generative models that propose new physical matter [34, 35].

3.1.1 Creating and maintaining datasets

Many studies don't develop substantially novel methods, but instead take advantage of new data resources. This is facilitated by the increasing availability of public databases. The PubChem database, maintained by the NIH and currently the largest repository of open-access chemical information [36], has been leveraged by many studies for ligand-based drug discovery proofs and thus is a particularly noteworthy example of the value inherent in these curation efforts. Curation efforts spearheaded by government organizations as well as those led by individual research groups can both be enormously impactful, whether through amalgamation

of large existing datasets (a greater strength of the broad collaborations) or the accumulation of high-quality, well-curated data (which efforts by individual research groups tend may be better suited for).

Table 2 provides an incomplete list of some popular databases used for tasks related to chemical discovery. Additional databases related to materials science can be found in refs. 37 and 38. Some related to drug discovery are contained in refs. 39 and 40. Additional publications compare commercial screening libraries that can be useful in experimental or computational workflows [41, 42].

Table 2: Overview of some databases used to facilitate discovery in the chemical sciences. API: application programming interface

Chemical structures			
Name	Description	Size (approx.)	Availability
ZINC [43, 44]	commercially-available compounds	35 M	Open
ChemSpider [45]	structures and misc. data	67 M	API
SureChEMBL [46]	structures from ChEMBL	1.9 M	Open
Super Natural II [47]	natural product structures	325 k	Open
SAVI [48]	enumerated synthetically-accessible structures and their building blocks	283 M	Open
eMolecules [49]	commercially-available chemicals and prices	5.9 M	Commercial
MolPort [50]	in-stock chemicals and prices	7 M	On Request
REAL (Enamine) [51]	enumerated synthetically-accessible structures	11 B	Open
Chemspace [52]	in-stock chemicals and prices	1 M	On Request
GDB-11, GDB-13, GDB-17 [53–55]	exhaustively enumerated chemical structures	26.4 M; 970 M; 166 B	Open
SCUBIDOO [56]	enumerated synthetically-accessible structures	> 10 M	Open
CHIPMUNK [57]	enumerated synthetically-accessible structures	95 M	Open
Biological data			
Name	Description	Size	Availability
PubChem [36]	compounds and properties, emphasis on bioassay results	96 M	Open
ChEMBL [58, 59]	compounds and bioactivity measurements	1.9 M	Open
ChEBI [60]	compounds and biological relevance	56 k	Open
PDB [61]	biological macromolecular structures	150 k	Open
PDBBind [62]	protein binding affinity	20 k	Open
ProTherm [63]	thermodynamic data for proteins	10 k	Open
LINCS [64]	cellular interactions and perturbations	varies	Open
SKEMPI [65]	energetics of mutant protein interactions	7 k	Open
xMoDEL [66]	MD trajectories of proteins	1700	Open
GenBank [67]	species' nucleotide sequences	400 k	Open
DrugBank [68]	drug compounds, associated chemical properties, and pharmacological information	13 k	Open
BindingDB [69]	compounds and binding measurements	750 k; 1.7 M	Open
CDD [70]	collaborative drug discovery database for neglected tropical diseases	> 100 datasets	Registration
ToxCast [71, 72]	compounds and cellular responses	> 4500	Open
Tox21 [73, 74]	compounds and multiple bioassays	14k	Open
Chemical reactions			
Name	Description	Size	Availability
USPTO [75]	chemical reactions (patent literature)	3.3 M	Open
Pistachio [76]	chemical reactions (patent literature)	8.4 M	Commercial
Reaxys [77]	chemical reactions	>10 M	Commercial
CASREACT [78]	chemical reactions	>10 M	Commercial
SPRESI [79]	chemical reactions	4.3 M	Commercial
Organic Reactions [80]	chemical reactions	250 k	Commercial
EAWAG-BBD [81]	biocatalysis and biodegradation pathways	219	Open
NIST Chemical Kinetics [82]	gas-phase chemical reactions	38 k	Open
NMRShiftDB [83]	measured NMR spectra	52 k	Open

Molecular properties			
Name	Description	Size	Availability
QM7/QM7b [84, 85]	electronic properties (DFT)	7200	Open
QM9 [86]	electronic properties (DFT)	134 k	Open
QM8 [87]	spectra and excited state properties	22 k	Open
FreeSolv [88]	aqueous solvatio energies	642	Open
NIST Chemistry WebBook [89]	miscellaneous molecular properties	varies	Open
Materials			
Name	Description	Size	Availability
PoLyInfo [90]	polymer properties	400k	Open
COD [91–93]	crystal structures of organic, inorganic, metal-organics compounds and minerals	410k	Open
CoRE MOF [94]	properties of metal-organic frameworks	5k	Open
hMOF [95]	hypothetical metal-organic frameworks	140k	Open
CSD [96]	crystal structures	1 M	API
ICSD [97]	crystal structure data for inorganic compounds	180k	Commercial
NOMAD [98]	total energy calculations	50 M	Open
AFLOW [99, 100]	material compounds; calculated properties	2.1M; 282M	Open
OQMD [101]	total energy calculations	560 k	Open
Materials Project [102]	inorganic compounds and computed properties	87 k	API
Computational Materials Repository [103, 104]	inorganic compounds and computed properties	varies	Open
Pearson's [105]	crystal structures	319 k	Commercial
HOPV [106]	experimental photovoltaic data from literature, QM calculations	350 k	Open
Journal articles			
Name	Description	Size	Availability
Crossref [107]	journal article metadata	107 M	Open
PubMed [108]	biomedical citations	29 M	Open
arXiv [109]	arXiv articles (from many domains)	1.6 M	Open
Wiley [110]	full articles	millions	API
Elsevier [111]	full articles	millions	API

Several factors have contributed to the greater wealth and accessibility of chemical databases that can be used to facilitate discovery. First, hardware for automated experimentation has allowed us to generate data at a faster pace. Second, the successes of computational tools at leveraging large quantities of data has created a self-catalyzing phenomenon: as the capabilities of tools are more frequently and widely demonstrated, the incentive to collect and curate large datasets that can be used by these tools has grown.

Challenge: Establish open access databases with standardized data representations

Time invested in the creation of open databases of molecules, materials, and processes can have an outsized impact on discovery efforts that are able to make use of that data for supervised learning or screening.

Creating and maintaining these databases is not without its challenges. There's much to be done to capture and open-source the data generated by the scientific community. For cases where data must be protected by intellectual property agreements, we need software that can facilitate sharing between collaborators and guarantee privacy as needed [112, 113]. Even standardizing representations can be challenging, particularly for polymeric materials with stochastic structures and process-dependent attributes [114].

Government funding agencies in the EU and US are starting to prioritize accessibility of research results to the broader community [115]. Further evolution of the open data policies will accelerate discovery through broader analysis of data (crowdsourcing discovery [116–118]) and amalgamation of data for the purposes of machine learning. Best practices among experimentalists must begin to include the storage of experimental details and results in searchable, electronic repositories.

The data that exists in the literature that remains to be tapped by the curation efforts described above is vast. To access it, scientific researchers are gaining increasing interest in adapting information extraction techniques for use in chemistry [119–124]. Information extraction and natural language processing bring structure to unstructured data, e.g., published literature that presents information in text form. Methods have evolved from identifying co-occurrences of specifics words [125] to the formalization of domain-specific ontologies [126], learned word embeddings [26], knowledge graphs and databases [122], and causal models [25]. Learning from unstructured data presents additional challenges in terms of data set preparation and problem formulation, and is significantly less popular than working with pre-tabulated databases. Nevertheless, building knowledge graphs of chemical topics may eventually let us perform higher level reasoning [127] to identify and generalize from novel trends.

Challenge: Address the inconsistent quality of existing data

Existing datasets may not contain all of the information needed for a given prediction task (i.e., the input is underspecified or the schema is missing a metric of interest). Even when the right fields are present, there may be missing or misentered data from automated information extraction pipelines or manual entry.

As Williams et al. point out, data curation (which involves evaluating the accuracy of data stored in repositories) before the data are used to create a model or otherwise draw conclusions is very important: data submitted to the PDB is independently validated before it is added to the database, whereas data added to PubChem undergoes no prerequisite curation or validation [128]. Missing and/or misentered data curtails the accuracy of the resulting models. These issues plague databases including the PDB (misassigned electron density) and Reaxys (missing fields). As analytical technology continues to improve, one can further ask how much we should bother relying on old data in lieu of generating new data that we trust more.

Database curation policies must account for the potential for error propagation and incorporate standardization procedures that correct for errors when they arise [129, 130], for example by using ProsaII [131] to evaluate sequence-structure compatibility of PDB entries and identify errors [132]. While the type of crowdsourcing error correction exemplified by Venclovas et al. can be helpful, we argue that it shouldn't be relied upon [132]; curators should preemptively establish policies to help identify, control, and prevent errors.

3.1.2 Building empirical models

Various statistical methods have been used for model-building for many years. Some of the most dramatic improvements to statistical learning have been in the area of machine learning. Machine learning is now the go-to for developing empirical models that describe nonlinear structure-function relationships to estimate the properties of new physical matter and serve as surrogate models for expensive calculations or experiments. These models guide experimental selection for many discovery efforts, so improvements here significantly impact computer-aided discovery, even when the full workflow is not automated. Packages like scikit-learn, Tensorflow, and Pytorch have lowered the barrier for implementing empirical models, and chemistry-specific packages like DeepChem [133] and ChemML [134] represent further attempts to streamline model training and deployment (with mixed success in adoption).

Challenge: Improve representations of molecules and materials

There are many strategies for representing molecules and materials as inputs to empirical models, but certain aspects have yet to be adequately addressed by existing methods. Further, it is difficult to know which representation will perform best for a given objective *a priori*.

In the wake of the 2012 ImageNet competition, in which a convolutional neural network dominated rule-based systems for image classification [135], there has been a shift in modeling philosophy to avoid human feature engineering, and instead *learn* suitable representations [136]. This is in part enabled by new network architectures, such as message passing networks particularly suited to embedding molecular structures [137–139]. There is no consensus as to when the aforementioned deep learning techniques should be applied over "shallower" learning techniques like RFs or SVMs with fixed representations [140]; which method performs best is task-dependent and determined empirically [133, 141], although some heuristics, e.g. regarding the fingerprint granularity needed for a particular materials modeling task, do exist [142]. Use of molecular descriptors may make generalization to new inputs more predictable [28] but limits the space of relationships able to be described by presupposing that the descriptors contain all information needed for the prediction task. Further, selecting features for low-dimensional descriptor-based representations requires expert-level domain knowledge.

In addition to a better understanding of why different techniques perform well in different settings, there is a need for the techniques themselves to better capture relevant information about input structures. Some common representations in empirical QSAR/QSPR modeling are listed in Table 3. However, there are several types of inputs that current representations are unable to describe adequately. These include (a) polymers that are stochastically-generated ensembles of specific macromolecular structures, (b) heterogeneous materials with periodicity or order at multiple length scales, and (c) "2.5D" small molecules with

defined stereochemistry but flexible 3D conformations. Descriptor-based representations serve as a catch-all, as they rely on experts to encode input molecules, materials, or structures as numerical objects.

Representation	Description
Descriptors	Vector of calculated properties
Fingerprints	Vector of presence/absence or count of structural features (many types)
Coulomb matrices	Matrix of electrostatic interactions between nuclei
Images	2D line drawings of chemical structures
SMILES	String defining small molecule connectivity (can be tokenized or adapted in various ways, e.g., SELFIES [143], DeepSMILES [144])
FASTA	String for nucleotide or peptide sequences
Graphs	2D representation with connectivity information
Voxels	Discretized 3D or 4D representation of molecules
Spatial coordinates	3D representation with explicit coordinates for every atom

Table 3: Representations commonly used in empirical QSAR/QSPR modeling.

Challenge: Improve empirical modeling performance in low-data environments

Empirical modeling approaches must be validated on or extended to situations for which only tens of examples are available.

Defining a meaningful feature representation is especially important when data are limited [145, 146]. Challenging discovery problems may be those for which little performance data are available and validation is expensive. For empirical models to be useful in these settings, they must be able to make reasonably accurate predictions with only tens of data points. QSAR/QSPR performance in low-data environments is understudied, with few papers explicitly examining low-data problems (e.g., fewer than 100 examples) [147–149]. The amount of data "required" to train a model is dependent on the complexity of the task, the true (unknown) mathematical relationship between the input representation and output, the size of the domain over which predictions are made, and the coverage of the training set within that space.

Challenge: Incorporate physical invariance and equivariance properties

By ensuring that models are only sensitive to meaningful differences in input representations, one can more effectively learn an input-output relationship without requiring data augmentation to also learn input-input invariance or equivariance.

One potential way to improve low-data performance and generalization ability is to embed physical invariance or equivariance properties into models. Consider a model built to predict a physical property from a molecular structure: molecular embeddings from message passing neural networks are inherently invariant to atom ordering. In contrast, embeddings calculated from tensor operations on Coulomb matrices are not invariant. Sequence encoders using a SMILES string representation of a molecule have been shown to benefit from data augmentation strategies that teach the chemical equivalence of multiple SMILES strings

[150, 151]. There are strong parallels to image recognition tasks, where an object recognition model should be invariant to translation, rotation, and scale. When using 3D representations of molecules with explicit atomic coordinates, it is preferable to use embedding architectures that are inherently rotationally-invariant [152, 153] instead of relying on inefficient preprocessing steps of structure alignment [29] and/or rotational enumeration [154] for voxel representations, which still may lead to models that do not obey natural invariance or equivariance laws.

Challenge: Unify and utilize heterogeneous datasets

Vast quantities of unlabeled or labeled data can be used as baseline knowledge for pretraining empirical models or in a multitask setting when tasks are sufficiently related.

When human researchers approach a discovery task, they do so equipped with an intuition and knowledge base built by taking courses, reading papers, running experiments, etc. In computational discovery workflows with machine learning-based QSAR modeling, algorithms tend to focus only on the exact property or task and make little use of prior knowledge; only via the input representation, model architecture, and constraints on the search space is domain-specific knowledge embedded. Models are often trained from scratch on datasets that contain labeled (molecule, value) pairs.

Such isolated applications of machine learning to QSAR/QSPR modeling can be effective, but there is a potential benefit to multitask learning or transfer learning when predictions are sufficiently related [155–159]. Searls argues that drug discovery stands to benefit from integrating different datasets relating to various aspects of gene and protein functions [160]. As a simple example, one can consider that the prediction of phenotypes from suppressing specific protein sequences might benefit from knowledge of protein structure, given the connection between protein sequence → structure → function. For some therapeutic targets, there are dozens of databases known to be relevant that have not been meaningfully integrated [161]. Large-scale pretraining is a more general technique that can be used to learn an application-agnostic atom- or molecule-level representation prior to refinement on the actual QSAR task [162–165]. Performance on phenotypic assays has even been used directly as descriptors for molecules in other property prediction tasks [166], as has heterogeneous data on drug, protein, and drug-protein interactions [167].

Challenge: Improve interpretability of machine learning models

Machine learning models are typically applied as black box predictors with some minimal degree of *ex post facto* interpretation: analysis of descriptor importance, training example relevance, simplified decision trees, etc. Extracting explanations consistent with those used by human experts in the scientific literature requires the structure of the desired explanations to be considered and built into the modeling pipeline.

To the extent that existing autonomous discovery frameworks generate hypotheses that explain observations and interpret the results of experiments, they rarely do so in a way that is directly intelligible *to humans*, limiting the expansion of scientific knowledge that is derived from a campaign. In connection with this, many of the case studies from Part 1 focus on discoveries that are readily physically observable–identifying a new molecule that is active against a protein target, or a new material that can be used to improve energy capture–rather than something more abstract, such as answering a particular scientific question. We can probe model understanding by enumerating predictions for different inputs, but these are human-defined experiments to answer human-defined hypotheses (e.g., querying a reaction prediction model with substrates across a homologous series). Standard approaches to evaluating descriptor importance still require careful control experiments to ensure that the explanations we extract are not spurious, even if they align with human intuition [168]. We again refer readers to ref. 169 for a review of QSAR interpretability. Ref. 170 reviews additional aspects of explainable machine learning for scientific discovery.

Many challenges above can be approached by what Rueden et al. call *informed machine learning*: "the explicit incorporation of additional knowledge into machine learning models". The taxonomy they propose is reproduced in Figure 1. In particular, several points relate to (a) the integration of natural sciences (laws) and intuition into representations and model architectures and (b) the integration of world knowledge through pretraining or multitask/transfer learning.

3.2 Automated validation and feedback

Iterating between hypothesis generation and validation can be fundamental to the discovery process. One often needs to collect new data to refine or prove/disprove hypotheses. Sufficiently advanced automation can compensate for bad predictions by quickly falsifying hypotheses and identifying false positives [172] (i.e., being "fast to fail" [173]). The last several decades have brought significant advances in automation of small-scale screening, synthesis, and characterization, which facilitates validation via physical experiments, as well as advances in software for faster and more robust computational validation.

3.2.1 Experimental validation

Many of the case studies we present portray great strides in terms of the speed and scale of experimental validation. High-throughput and parallelized experimentation capabilities have been transformational in the biological space and increasingly are being imported into the chemistry space [174]. The adoption of HTE has simplified screening broad design spaces for new information [175–177]. Beyond brute-force experimentation, there are new *types* of experiments to accelerate the rate of data generation and hypothesis

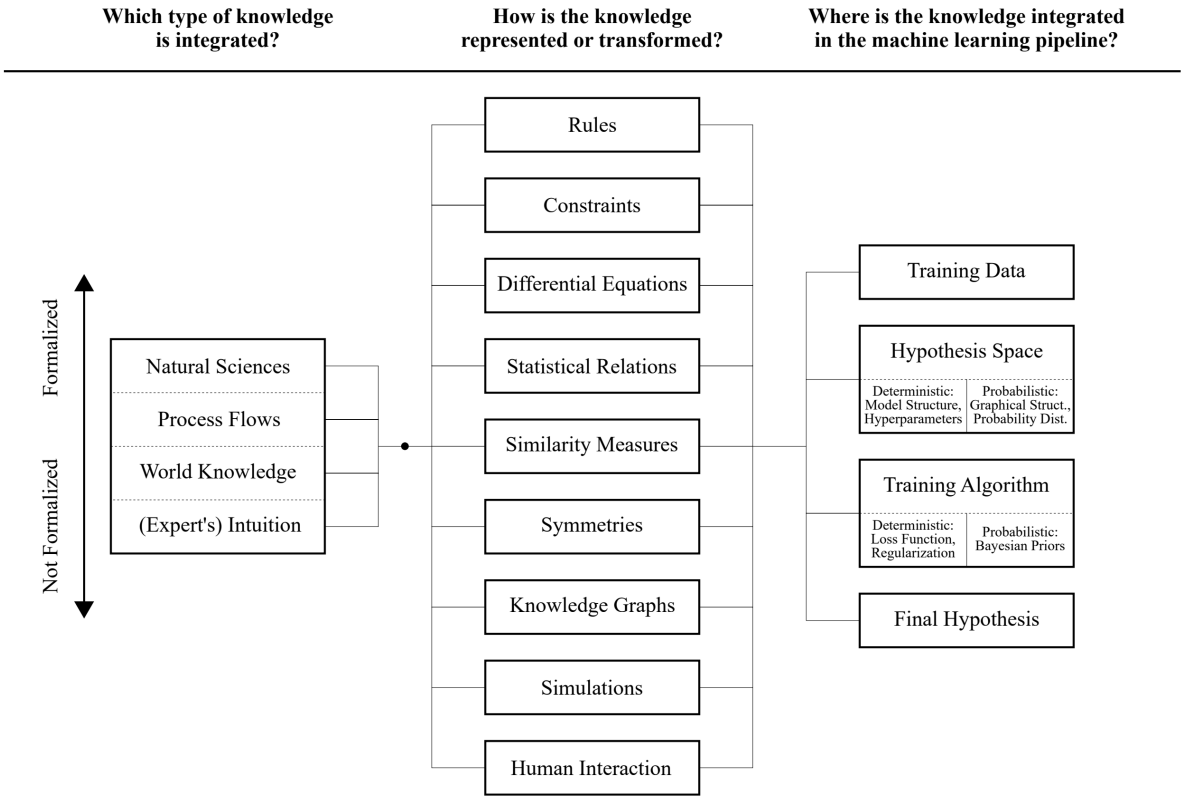


Figure 1: Taxonomy of *informed machine learning* proposed by Rueden et al. The incorporation of prior knowledge into machine learning modeling can take a number of forms. Figure reproduced from ref. 171.

validation. These include split-and-pool techniques and other combinatorial methods to study multiple candidates simultaneously [178–180].

Challenge: Expand the scope of automatable experiments

Whether an iterative discovery problem's hypotheses can be autonomously validated depends on whether the requisite experiments are amenable to automation.

If we are optimizing a complex design objective, such as in small molecule drug discovery, we benefit from having access to a large search space. Many syntheses and assays are compatible with a well-plate format and are routinely automated (e.g., Adam [3] and Eve [2]). Moving plates, aspirating/dispensing liquid, and heating/stirring are all routine tasks for automated platforms. Experiments requiring more complex operations may still be automatable, but require custom platforms, e.g., for the growth and characterization of nanotubes by ARES [7] or deposition and characterization of thin films by Ada [15]. Dispensing and metering of solids is important for many applications but is challenging at milligram scales, though new strategies are emerging that may decrease the precision required for dosing solid reagents [181]. Indeed, the set of automatable experiments is ever-increasing, but a universal chemical synthesizer [182] remains

elusive. The result of this gap is that design spaces may be not only constrained through prior knowledge (an intentional and useful narrowing of the space), but also *limited* by the capabilities of the automated hardware available. Characterizing the structure of physical matter is increasingly routine, but our ability to measure complex functions and connect them back to structure remains limited. Oliver et al. list several useful polymer characterization methods that have eluded full automation, such as differential scanning calorimetry and thermogravimetric analysis [177].

Challenge: Facilitate integration through systems engineering

Scheduling, performing, and analyzing experiments can involve coordinating tasks between several independent pieces of hardware/software that must be physically and programmatically linked.

Expanding the scope of experimental platforms may require the integration of independent pieces of equipment at both the hardware and software level. The wide variety of necessary tasks (scheduling, error-handling, etc.) means that designing control systems for such highly-integrated platforms is an enormously complex task [183]. As a result, developing software for integration of an experimental platform [184] (Figure 2) can be a large contributor to the cost. The lack of standard APIs and command sets between different hardware providers means that each requires its own driver and software wrapper; this is particularly troublesome for analytical equipment, which even lacks standardization in file formats for measured data. Programs like OVERLORD and Roch et al.'s ChemOS [185] are attempts to create higher-level controllers. Throughput-matching in sequential workflows is a challenge in general, requiring a plan for "parking" (and perhaps stabilizing) samples in the event of a bottleneck downstream. These practical issues must be resolved to benefit from increased integration.

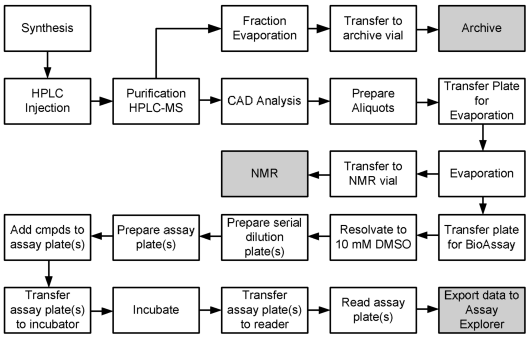


Figure 2: The workflow of automated synthesis, purification, and testing requires the scheduling of many independent operations handled by different pieces of hardware and software. Figure reproduced from Baranczak et al. [184].

10.1002/anie.201909989

CHALLENGE: Automate the planning of multistep chemical syntheses

Many discovery tasks involve proposing new chemical matter; approaching these tasks with autonomous systems requires the ability to synthesize novel compounds on-demand.

A particularly challenging class of experiments is on-demand synthesis. The primary methodological challenge for general purpose automated synthesis is the *planning* of processes–designing multistep synthetic routes using available building blocks; selecting conditions for each reaction step including quantities, temperature, and time; and automating intermediate and final purifications. If reduced to stirring, heating, and fluid transfer operations, chemical syntheses are straightforward to automate [186–188], and existing robotic platforms (Figure 3) can execute a series of process steps if those steps are precisely planned [184, 189]. However, current CASP tools are unable to make directly implementable recommendations with this level of precision.

There are two diverging philosophies of how to approach automated synthesis: (a) the development of general-purpose machines able to carry out most chemical reactions, or (b) the development of specialized machines to perform a few general-purpose reactions that are still able to produce most molecules. The references in the preceding paragraph follow the former approach. Burke and co-workers have advocated for the latter and propose using advanced MIDA boronate building blocks and a single reaction/purification strategy to simplify process design [190]. Peptide and nucleic acid synthesizers exemplify this notion of automating a small number of chemical transformations to produce candidates within a vast design space.

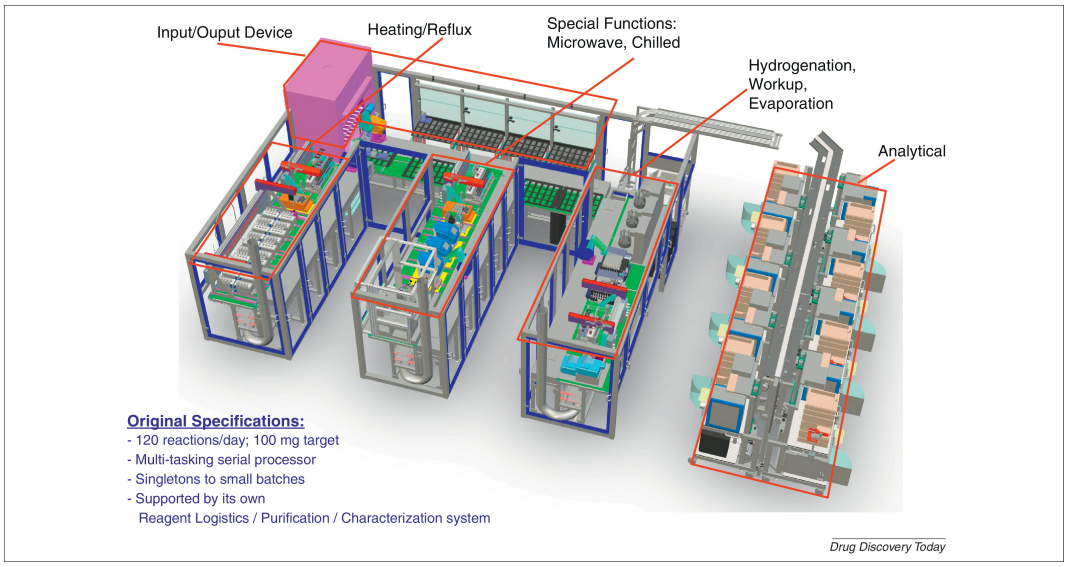


Figure 3: Rendering of Eli Lilly's first generation Automated Synthesis Laboratory (ASL) for automated synthesis and purification. Reproduced from ref. 189.

3.2.2 Computational validation

Many discoveries can be validated with high confidence through computational techniques alone. Where applicable, this can be extremely advantageous. This is because the logistics of the alternative (physical experiments) may be much more complex, e.g., relying on access to a large chemical inventory (of candidates or as precursors) to perform validation experiments within a large design space. An emblematic example of discoveries that can be validated through computation alone is that of physical matter whose desired function can be reliably estimated with first principles calculations.

Challenge: Accelerate code/software used for computational validation
Just as in physical experiments, there are practical challenges in computational experiments related to the throughput of high fidelity calculations.

A unique feature of computational validation is the well-recognized tradeoff between speed and accuracy. Consider Lyu et al.'s virtual screen of 170 million compounds to identify binders to two protein targets through rigid-body docking [191]. The computational pipeline was fast–requiring about one second per compound–only requiring ≈50,000 core-hours in total, and successfully yielded dozens of novel chemotypes and a few sub-nanomolar binders. While rigid-body docking is not as robust as, say, explicit molecular dynamics, its predictions still correlated with experimental binding and generated useful candidate molecules. An earlier study by Gómez-Bombarelli et al. used time-dependent DFT to evaluate a total of 400,000 candidate OLED molecules selected from a pool of 1.6 million enumerated candidates; the computational cost for this study was roughly 13 million core-hours [11]. There are billions upon billions of molecules in enumerated synthetically-accessible chemical space. Given our desire to work with broad design spaces, there is a need for faster workflows that can conduct large-scale computational validation.

One strategy to accelerate computational validation is to create surrogate models of first-principles calculations [11, 14, 192]. Predictions made using surrogate machine learning models regressed to measured or simulated data almost always carry greater uncertainty than the original experiment, and therefore confirming the final discovery is especially reliant on higher-fidelity, often physical validation. An orthogonal strategy is to accelerate the calculations themselves without sacrificing accuracy. Many computational chemistry and molecular modeling packages have been developed to take advantage of hardware acceleration on GPUs [193–195], FPGAs [196], and even ASICs [197].

Challenge: Broaden capabilites / applicability of first-principles calculations
Many properties of interest cannot be simulated accurately, forcing us to rely on experimental validation.

Expanding the scope of what can be accurately modeled would open up additional applications for purely

computational autonomous workflows. There are some tasks for which computational solutions exist but could be improved, including binding prediction through docking, reactive force field modeling, transition state searching, conformer generation, solvation energy prediction, and crystal structure prediction. There are other tasks with even fewer satisfactory approaches, including long timescale molecular dynamics and multiscale modeling in materials. Some grand challenges in computational chemistry are discussed in refs. 198 and 199.

3.2.3 Shared challenges

> **Challenge: Ensure that validation reflects the real application**
> Computational or experimental validation that lends itself to automation is often a proxy for a more expensive evaluation. If the proxy and the true metric are misaligned, an autonomous platform will not be able to generate any useful results.

Ideally, there would be perfect alignment between the approaches to validation compatible with an autonomous system and the real task at hand. This is impossible for tasks like drug discovery, where imperfect *in vitro* assays are virtually required before evaluating *in vivo* performance during preclinical development. For other tasks, assays are simplified for the sake of automation or cost, e.g., Ada's measurement of optoelectronic properties of a thin film as a proxy for hole mobility as a proxy of the efficiency of a multicomponent solar cell [15]. Assays used for validation in autonomous systems do not necessarily need to be high throughput, just high fidelity and automatable. Avoiding false results, especially false negatives in design spaces where positive hits are sparse (e.g., binders of an "undruggable" protein), is critical [200].

> **Challenge: Lower the cost of automated validation**
> Relatively few things can be automated cheaply. This is especially true for problems requiring complex experimental procedures, e.g., multi-step chemical synthesis.

While the equipment needed for a basic HTE setup is becoming increasingly accessible and compatible with the budget of many academic research groups [201, 202], we must increase the complexity of the automated platforms that are used for validation before increasing the complexity of problems they can address. Autonomous systems need not be high-throughput in nature, but, as we have mentioned several times throughout this review, one of the key goals of their development should be to facilitate exploration of ever-broader design spaces that we cannot explore manually. It is imperative that some attention is paid to affordability, lest cost inhibit adoption. Homegrown systems can be made inexpensively through integration of common hardware components and open-source languages for control [187, 203]. Miniaturization

reduces material consumption costs, but can complicate system fabrication and maintenance. The decision to automate a workflow will ultimately depend on a holistic evaluation of its return on investment [183].

The costs of computational assays are less of an impediment to autonomous discovery than experimental assays, given the accessibility of large-scale compute. Improving their accuracy is more of a priority. For example, the docking method used by Lyu et al. was sufficiently inexpensive to screen millions of compounds and obtain results that correlate with experimental binding affinity, but the majority of high scoring compounds are false positives and the differentiation of top candidates is poor [191, 204].

Challenge: Combine newly acquired data with prior literature data

Predictive models trained on existing data reflect beliefs about structure-property landscapes; when new data are acquired, that belief must be updated, preferably in a manner that reflects the relative confidence of the data sources.

A fundamental question yet to be addressed in studies combining data mining with automated validation is the following: how should new data acquired through experimental/computational validation be used to update models pretrained on literature data? The quintessential workflow for non-iterative data-driven discovery of physical matter includes (a) regressing a structure-property dataset, (b) proposing a new molecule, material, or device, and (c) validating the prediction for a small number of those predictions. Incorporating this new data into the model should account for the fact that the new data may be generated under more controlled conditions or may be higher fidelity than the literature data.

The nature of existing data can be different from what is newly acquired. For example, tabulated reaction data are available at the level of chemical species, temperature, time, intended major product, and yield. In the lab, we will know the conditions quantitatively (e.g., concentrations, order of addition), will have the opportunity to record additional factors (e.g., ambient temperature, humidity), and may be able to measure additional endpoints (e.g., identify side products). However, while we can more thoroughly evaluate different reaction conditions than what has been previously reported, the diversity of substrates reported in the literature exceeds what is practical to have in-stock in any one laboratory; we must figure out how to meaningfully integrate the two. For discovery tasks that aim to optimize physical matter with standardized assays, where databases contain exactly what we would calculate or measure, this notion of complementarity is less applicable.

3.3 Selection of experiments for validation and feedback

Excellent foundational work in statistics on (iterative) optimal experimental design strategies has been adapted to the domain of chemistry. Although iterative strategies often depend on manually-designed ini-

tializations and constrained search spaces, algorithms can be given the freedom to make decisions about which hypotheses to test. This flexibility makes iterative strategies inherently more relevant to autonomous discovery than noniterative ones.

A variety of algorithms exist for efficiently navigating design spaces and/or compound libraries (virtual or otherwise). Broadly speaking, these can be categorized as model-free–black box optimizations, including evolutionary algorithms (EAs)–or model-based–using surrogate models for predicting performance and/or model uncertainty. The latter category includes uncertainty-guided experimental selection where an acquisition function quantifies how useful a new experiment would be [17]; ref. 20 provides a tutorial on Bayesian optimization.

Challenge: Quantify model uncertainty and domain of applicability

Active learning strategies are crucially dependent on quantifying uncertainty; doing so reliably in QSAR/QSPR modeling remains elusive, and current strategies cannot anticipate structure-activity cliffs or other rough features.

Accurate uncertainty quantification drives discovery by drawing attention to underexplored areas of a design space and helping to triage experiments, e.g., in combination with Bayesian optimization [205]. Statistical and probabilistic frameworks can account for uncertainty when analyzing data and selecting new experiments [205–209], but we must be able to meaningfully estimate our uncertainty to use them. Common frequentist methods for estimating uncertainty include model ensembling [210] and Monte Carlo (MC) dropout [211]; various Bayesian approaches like the use of Gaussian process models have been used as well [209, 212]. Not only is it difficult to generate meaningful outcomes with these methods, but also they tend to be computationally expensive (MC dropout less so than the others). In QSAR/QSPR, one often tries to define a domain of applicability (DOA) as a coarser version of uncertainty, where the DOA can be thought of as the input space for which the prediction and uncertainty estimation is meaningful [213–215].

There is little to no agreement on the correct way to estimate epistemic uncertainty (as opposed to aleatoric uncertainty, which is that which arises from measurement noise). In drug discovery, activity cliffs [216]–sharp changes in binding affinity resulting from minor structural changes–are especially troublesome and call into question any attempt to directly connect structural similarity to functional similarity [217, 218]. Even functional descriptor-based representations are unlikely to capture all salient features. Implicit or explicit assumptions must be made when choosing a representational and modeling technique, for example choosing an appropriate kernel and a prior on (or a hyperparameter controlling) the smoothness of the landscape in a Gaussian processes model [219].

Challenge: Quantify the tradeoff between experimental difficulty and information gain

Experiment selection criteria should be able to account for the difficulty of an experiment, i.e., employ cost-sensitive active learning.

Experiment selection methods rarely account for the *cost* of an experiment in any quantitative way. Separately, experiment selection is occasionally biased based on factors that are irrelevant to the hypothesis. If proposed experiments require the synthesis of several molecules (e.g., a compound library designed during lead optimization), an expert chemist will generally select those they determine to be easily synthesized, rather than those that are most informative. One must ask if it is worth spending weeks making a single compound that maximizes the expected improvement or if there is a small analogue library that is easier to synthesize that, collectively, offers a similar probability of improvement. In this setting, there will almost always be a tradeoff between data that is fast and inexpensive to acquire and data that is most useful for the discovery. The term in experimental design for this is *cost-sensitive* active learning [220].

Understanding that tradeoff is essential for autonomous systems where experiments can have very different costs (e.g., selecting molecules to be synthesized) or likelihoods of success (e.g., electronic structure simulations prone to failure) in contrast to where experiments have similar costs (e.g., selecting virtual molecules for rigid-body docking). The situation becomes more complex for batched optimizations where, e.g., the cost of synthesizing 96 molecules in a parallel well-plate format is not merely the sum of their individual costs, but depends on overlap in the precursors and reaction conditions they employ.

Williams et al. provide one example of how to roughly quantify the value of active learning-based screening for Eve [2]. It is easy to imagine how one might augment this framework to account for cost as part of the experiment selection process. However, the utility calculation heuristics used by Williams et al. would need to be substantially improved in order to be usefully applied to cases where the cost of experiments vary, which is the interesting setting here. To date, the experiments able to be conducted by a given automated or autonomous workflow are of comparable costs and the decision about whether those costs are reasonable is made by the human designer of the platform.

Challenge: Define discovery goals at a higher level

The ability of an automated system to make surprising or significant discoveries relies upon its ability to extrapolate and explore beyond what is known. This could be encouraged by defining broader objectives than what are currently defined.

In current data analyses, the structure of hypotheses tend to be prescribed: a mathematical function relating an expert-selected input to an expert-selected output, a correlative measure between two chemical terms, a causal model that describes a sequence of events. Ideally, we would be able to generate hypotheses

from complex datasets in a more open-ended fashion where do not have to know exactly what we are looking for. Techniques in knowledge discovery [221], unsupervised learning [222], and novelty detection [223] are intended for just that purpose and may present a path toward more open-ended generation of scientific hypotheses (Figure 4).

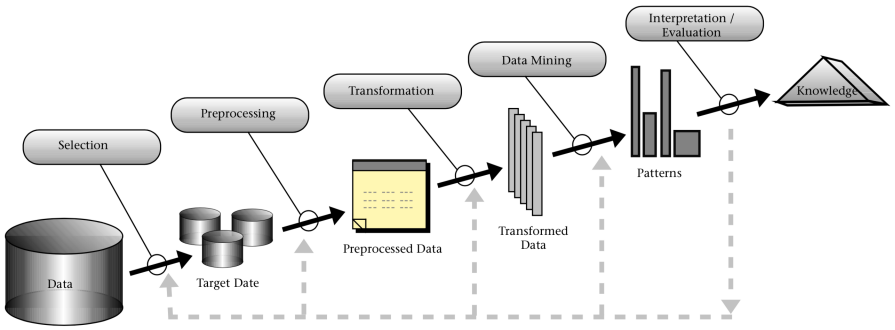


Figure 4: Overview of the process of knowledge discovery from databases. Figure reproduced from Fayyad et al. [221].

Experimental design can also be given greater flexibility by defining broad goals for discovery (performance, novelty, etc.) and using computational frameworks to learn tradeoffs in reaching those goals, e.g., through reinforcement learning. Consider the goal of compound selection during a drug discovery campaign: to identify a molecule that ends up being suitable for clinical trials. In the earlier information-gathering stages, we don't necessarily need to select the highest performing compounds, just the ones that provide information that lets us eventually identify them (i.e., a future reward). More generally, the experiments proposed for validation and feedback in a discovery campaign should be selected to achieve a higher-order goal (*eventually*, finding the best candidate) rather than a narrow objective (maximizing performance within a designed compound library).

Open-ended inference is a general challenge in deep learning [224], as is achieving what we would call creativity in hypothesis generation [225]. At some level, in order to apply optimization strategies for experimental design or analysis, the goal of a discovery search must be reducible to a scalar objective function. We should strive to develop techniques for guided extrapolation toward the challenging-to-quantify goals that the field has used when defining discovery: novelty, interestingness, intelligibility, and utility.

3.3.1 Proposing molecules and materials

Strategies for selecting molecules and materials for validation in discovery workflows are worth additional discussion (Figure 5). Iterative strategies of the sort described above apply here, with active learning being useful for selecting compounds from a fixed virtual library and evolutionary/generative algorithms being

10.1002/anie.201909989

useful for designing molecules on-the-fly. Generative models are a particularly attractive way to design molecules and materials with minimal human input, primarily biased by knowledge of the chemical space on which they are trained (Figure 6) [34, 35, 226].

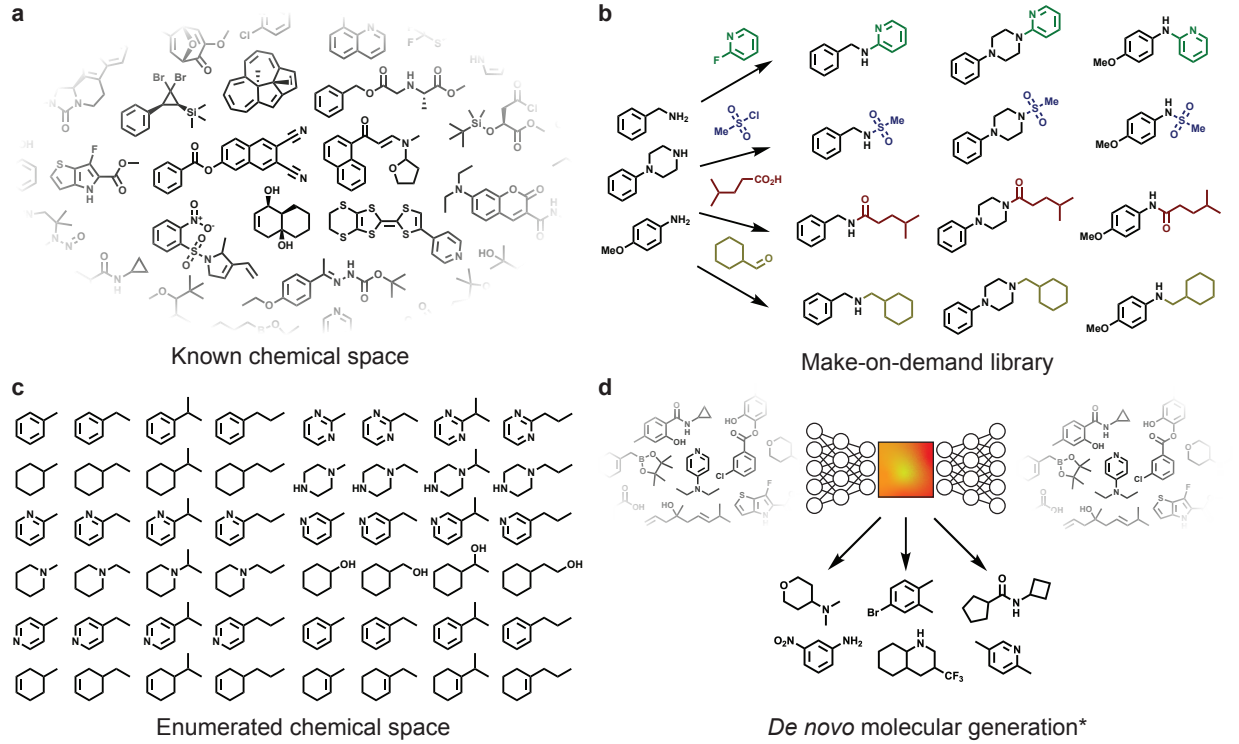


Figure 5: Common sources of molecules from which to select those that fulfill some design objective. Molecules can be selected from (a) a fixed, known chemical space, (b) a make-on-demand library of synthesizable compounds, (c) an enumerated library (via systematic enumeration or evolutionary methods), and (d) molecules proposed *de novo* from a generative model. *An autoencoder architecture is shown as a representative type of generative model.

Challenge: Bias generative models towards synthetic accessibility

Compared to fixed virtual libraries, a shortcoming of generative models is that the molecules or materials they propose may not be easily realizable.

Algorithms that can leverage existing data to suggest promising, as-yet-untested possibilities exist, but these do not yet function on the level of a human scientist in part because they do not understand what experiments are possible. Generative models can concoct new molecules in some abstract latent space, but simplistic measures of synthesizability [233, 234] are not enough to steer the models toward accessible chemical space. Make-on-demand virtual libraries provide a distinct advantage in that one is more confident proposed molecule can be made in short timeframe. Achieving that same confidence will be essential for the adoption of *de novo* methods, some of which are beginning to combine molecular generation and virtual

10.1002/anie.201909989

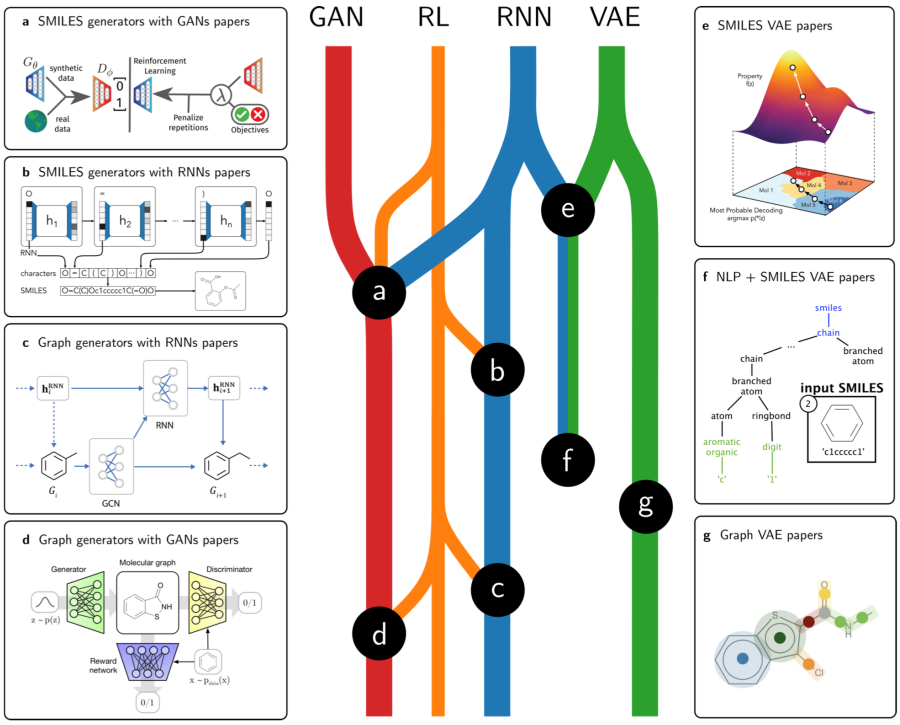


Figure 6: Schwalbe-Koda and Gómez-Bombarelli's timeline of generative model development for molecules (top to bottom). Figure reproduced from ref. 35. Figure subparts originally from refs. 227, 228, 229, 230, 231, and 232.

enumeration [235]. Some applications of generative models, like to peptide design, do not suffer from this limitation as, to a first approximation, most peptides are equally synthesizable.

Challenge: Benchmark problems for molecular generative models

The current evaluations for generative models do not reflect the complexity of real discovery problems.

The explosion of techniques for molecular generation has outpaced our ability to meaningfully assess their performance. A metric introduced early on as a proxy objective is the "penalized logP" metric for molecular optimization. While not used for any actual discovery efforts, a heuristic function of estimated logP, synthetic accessibility, and a penalty for rings larger than 6 atoms was introduced for (and continues to be used for) benchmarking. This metric bears little resemblance to any multiobjective function one would use in practice. Only recently have more systematic benchmarks been introduced to cover a wider range of learning objectives: either maximizing a scalar objective or learning to mimic a distribution of training molecules. Two frameworks for such model comparisons include GuacaMol [236] and MOSES [237]. However, these do not consider the number of function evaluations required by each method and still represent simplistic goals. Optimization goals that better reflect the complexity of real discovery tasks might include binding or selectivity as predicted by docking scores [238]

3.4 Evaluation

Challenge: Demonstrate extrapolative power of predictive models

If the ultimate goal of computer-aided discovery is to generate new scientific knowledge, extrapolation beyond what is known is a necessity.

The majority of approaches to automated discovery of physical matter rely on predictive models to guide the selection of experiments. The most effective models to facilitate this process will be able to at least partially extrapolate from our current knowledge to new chemical matter, eliminating the need for brute force experimentation. This extrapolative power–the ability of QSAR/QSPR models to generalize to design spaces they have not been trained on–should be prioritized as an evaluation metrics during model development. The potential for algorithms to guide us toward novel areas of chemical or reactivity space was emphasized in a recent review by Gromski et al. [239].

Challenge: Demonstrate design-make-test beyond proof-of-concept

All studies to-date that demonstrate closed-loop design, make, test cycles have been proof-of-concepts limited to narrow search spaces, severely limiting their practical utility.

A compelling demonstration of autonomous discovery in chemistry would be the closed-loop design, synthesis, and testing of new molecules to optimize a certain property of interest or build a structure-property model. There has not been much progress since early proof-of-concept studies that could access only a limited chemical space [5, 240] despite significant advances in the requisite areas of molecular design, CASP, and automated synthesis. These constraints on the design space to ensure compatibility with automated validation prevent us from addressing many interesting questions and optimization objectives. Chow et al. describe several case studies where certain steps in the drug discovery have been integrated with each other for increased efficiency, but acknowledge–as others have–that *all* stages must be automated and integrated for maximal efficiency [241–243].

Challenge: Develop benchmark problems for discovery

Developing methods for autonomous discovery would benefit from a "sandbox" that doesn't eliminate all of the complexity of real domain applications.

There is no unified strategy for the use of existing data and the acquisition of new data for discovering functional physical matter, processes, or models. The existence of benchmarks would encourage method development and makes it easier to evaluate when new techniques are an improvement over existing ones. We have such evaluations for purely computational problems like numerical optimization and transition state

searching, but there are no realistic benchmarks upon which to test algorithms for autonomous discovery (e.g., hypothesis generation, experimental selection, etc.).

Vempaty et al. describe one way to evaluate knowledge discovery algorithms through a simplified "coupon-collector model"; this model assumes that domain knowledge is a set of elements to be identified through a noisy observation process [244], which represents a limited problem formulation. Even for subtasks with seemingly better-defined goals like building empirical QSAR/QSPR models, there are no standard evaluations for assessing interpretability, uncertainty quantification, or generalizability. The field will need to collectively establish a set of problem formulations that describe many discovery tasks of interest to domain experts in order to benchmark components of autonomous discovery. Given the practical obstacles to validation through physical experiments, computational chemistry may be the right playground for advancing these techniques.

However, we do caution that an overemphasis on quantitative benchmarking can be detrimental. Language tasks have reached a point where the amount of compute required for competitive performance is inaccessible for all but the most well-resourced research groups [245]. Unless benchmarking controls what (open source) training data are permissible, a lack of access to compute and data may inadvertently discourage method development.

4 Conclusion

4.1 Attribution of discovery to automation

The case studies in this two-part article illustrate that computer assistance and automation have become ubiquitous parts of scientific discovery both by reducing the manual effort required to complete certain tasks and by enabling entirely new approaches to discovery at an unprecedented throughput. But to what extent can the discovery itself be considered a direct result of automation or autonomy?

As summarized in our reflection of the case studies in Part 1, very few studies can claim to have achieved a high level of autonomy. In particular, researchers frequently gloss over the fact that specifying the discovery objective, defining the search space, and narrowing that space to the "relevant" space that is ultimately explored requires *substantial* manual input. While there will always be a need for subject matter experts in constructing these platforms and associated workflows, we hope it will be possible to endow autonomous platforms with sufficiently broad background knowledge and validation capabilities that this initial narrowing of the search space is less critical to their success.

4.2 Changing definitions of discovery

The bar for what makes a noteworthy discovery is ever-increasing. Computer-aided structural elucidation, building structure-activity relationships, and automated reaction optimization are all discoveries under the definition we have presented here, but they are not perceived to be as significant as they were in the past. As computational algorithms become more flexible and adaptive in other contexts, and as the scope of automatable validation experiments expands, more and more workflows will appear routine.

We have intentionally avoided a precise definition of the degree of confidence required for a discovery without direct experimental observation of a desired physical property. This is because this varies widely by domain and is rapidly evolving as computational validation techniques and proxy assays become more accurate. A computational prediction of a new chemical reaction would likely not be considered a discovery under any circumstances without experimental validation. A computational prediction of a set of bioactive compounds might, but with a subjective threshold for the precision of its recommendations. Whether the computational workflow has directly made the discovery of a new compound might depend if all of the top n compounds were found to be active, or if at least m of the top n were, etc.

4.3 Role of the human

The current role of humans in computer-assisted discovery is clear. Langley writes of the "developer's role" in terms of high-level tasks: formulating the discovery problem, settling on an effective representation, and transforming the output into results meaningful to the community [246, 247]. Honavar includes mapping the current state of knowledge and generating/prioritizing research questions [248].

Alan Turing's *imitation game* ("the Turing test") asks whether a computer program can be made indistinguishable from a human conversationalist [249]. It is interesting to wonder if we can reach a point where autonomous platforms *are* able to report insights and interpretations that are indistinguishable (both in presentation and scientific merit) from what a human researcher might publish in a journal article. Among other things, this would require substantial advances in hypothesis generation, explainability, and scientific text generation. Machine-generated review articles and textbooks may be the first to pass this test [250]. Kitano's more ambitious grand challenge in his call-to-arms is to make a discovery in the biomedical sciences worthy of a Nobel Prize [172].

We do not want to overstress a direct analogy of the Turing test to autonomous discoveries, because the type of discoveries typically enabled by automation and computational techniques are often distinct from those made by hand. For the field to have the broadest shared capabilities, the best discovery platforms will excel at tasks that humans can't easily or safely do. The scale of data generation, the size of a design

space that can be searched, and the ability to define new experiments that account for enormous quantities of existing information makes autonomous systems equipped to make discoveries in ways entirely distinct from humans.

Turing makes the point that the goals of machines and programs are distinct; that a human would lose in a race with an airplane does not mean we should slow down airplanes so their speeds are indistinguishable. Rephrased more recently by Steve Ley, "while people are always more important than machines, increasingly we think that it is foolish to do thing machines can do better than ourselves" [251]. Particularly when faced with the grunt work of some manual experimentation, "leaving such things to machines frees us for still better tasks" (Derek Lowe) [182]. We should *embrace* the divergence of human versus machine tasks.

4.4 Outlook

We join many others in touting the promise of autonomous or accelerated discovery [241, 243, 252–260]. Automation has brought increased productivity to the chemical sciences through efficiency, reproducibility, reduction in error, and the ability to cope with complex problems at scale; likewise, machine learning and data science through the identification of highly nonlinear relationships, trends, and patterns in complex data.

The previous section identified a number of directions in which additional effort is required to capture the full value of that promise: creating and maintaining high-quality open access datasets; building interpretable, data-efficient, and generalizable empirical models; expanding the scope of automated experimental platforms, particularly for multistep chemical synthesis; improving the applicability and speed of automated computational validation; aligning automated validation with prior knowledge and what is needed for different discovery applications, ideally not at significant cost; improving uncertainty quantification and cost-sensitive active learning; enabling open-ended hypothesis generation for experimental selection; and explicitly incorporating synthesizability considerations into generative models and benchmarking on realistic tasks. Evaluation will require the creation of benchmark problems that we argue should focus on whether algorithms facilitate extrapolation to underexplored, large design spaces that are currently expensive or intractable to explore.

Numerous research initiatives are supporting work in these directions. For example, the United States Department of Defense recently funded a multidisciplinary initiative to develop a Scientific Autonomous Reasoning Agent; the Defense Advanced Research Projects Agency (DARPA) has funded several programs relevant to autonomous discovery, including the Data-Driven Discovery of Models, Big Mechanism Project, Make-It, Accelerated Molecular Discovery, and Synergistic Discovery and Design; the Engineering and Physical Sciences Research Council (EPSRC) has an ongoing Dial-a-Molecule challenge that strives to debottleneck

synthesis, and recently launched a Centre of Doctoral Training in Automated Chemical Synthesis enabled by Digital Molecular Technologies; the Materials Genome Initiative, Materials Project, and Mission Innovation's Materials Acceleration Platform continue to bring sweeping changes to how data in materials science is collected, curated, and applied to discovery. Many more commercial efforts are underway as well, with significant investment from the pharmaceutical industry into the integration and digitization of their drug discovery workflows.

A 2004 perspective article by Glymour stated that we were in the midst of a revolution to automate scientific discovery [261]. Regardless of whether we were then, we certainly seem to be now.

5 Acknowledgements

We thank Thomas Struble for providing comments on the manuscript and our other colleagues and collaborators for useful conversations around this topic. This work was supported by the Machine Learning for Pharmaceutical Discovery and Synthesis Consortium and the DARPA Make-It program under contract ARO W911NF-16-2-0023.

References

[1] R. D. King et al., *Science* **2009**, *324*, 85–89.

[2] K. Williams et al., *J. R. Soc. Interface* **2015**, *12*, 20141289.

[3] R. D. King, K. E. Whelan, F. M. Jones, P. G. K. Reiser, C. H. Bryant, S. H. Muggleton, D. B. Kell, S. G. Oliver, *Nature* **2004**, *427*, 247–252.

[4] L. Weber, S. Wallbaum, C. Broger, K. Gubernator, *Angew. Chem. Int. Ed. in English* **1995**, *34*, 2280–2282.

[5] B. Desai et al., *J. Med. Chem.* **2013**, *56*, 3033–3047.

[6] B. J. Reizman, Y.-M. Wang, S. L. Buchwald, K. F. Jensen, *React. Chem. Eng.* **2016**, *1*, 658–666.

[7] P. Nikolaev, D. Hooper, N. Perea-López, M. Terrones, B. Maruyama, *ACS Nano* **2014**, *8*, 10214–10222.

[8] J. D. Kangas, A. W. Naik, R. F. Murphy, *BMC Bioinf.* **2014**, *15*, 143.

[9] C. W. Gao, J. W. Allen, W. H. Green, R. H. West, *Comput. Phys. Commun.* **2016**, *203*, 212–225.

[10] J. Fang, X. Pang, R. Yan, W. Lian, C. Li, Q. Wang, A.-L. Liu, G.-H. Du, *RSC Adv.* **2016**, *6*, 9857–9871.

[11] R. Gómez-Bombarelli et al., *Nature Mater.* **2016**, *15*, 1120–1127.

[12] M. H. S. Segler, T. Kogej, C. Tyrchan, M. P. Waller, *ACS Cent. Sci.* **2017**, *4*, 120–131.

[13] A. W. Thornton et al., *Chem. Mater.* **2017**, *29*, 2844–2854.

[14] J. P. Janet, L. Chan, H. J. Kulik, *J. Phys. Chem. Lett.* **2018**, *9*, 1064–1071.

[15] B. P. MacLeod et al., *arXiv:1906.05398 [cond-mat physics:physics]* **2019**.

[16] P. Nikolaev, D. Hooper, F. Webber, R. Rao, K. Decker, M. Krein, J. Poleski, R. Barto, B. Maruyama, *Npj Comput. Mater.* **2016**, *2*, 16031.

[17] B. Settles, *Synthesis Lectures on Artificial Intelligence and Machine Learning* **2012**, *6*, 1–114.

[18] C. M. Anderson-Cook, C. M. Borror, D. C. Montgomery, *J. Stat. Plan. Inference* **2009**, *139*, 629–641.

[19] J. Mockus, V. Tiesis, A. Zilinskas, *Towards global optimisation* **1978**, *2*, 117–129.

[20] P. I. Frazier, *arXiv preprint arXiv:1807.02811* **2018**.

[21] J. Aires de Sousa in *Applied Chemoinformatics*, (Eds.: T. Engel, J. Gasteiger), Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim, Germany, **2018**, pp. 133–163.

[22] Y. Gil, H. Hirsh in 2012 AAAI Fall Symposium Series, **2012**.

[23] A. Sharafi, *Knowledge Discovery in Databases*, Springer Fachmedien Wiesbaden, Cambridge, MA, USA, **2013**.

[24] E. Kim, K. Huang, A. Tomala, S. Matthews, E. Strubell, A. Saunders, A. McCallum, E. Olivetti, *Sci. Data* **2017**, *4*, 170127.

[25] B. M. Gyori, J. A. Bachman, K. Subramanian, J. L. Muhlich, L. Galescu, P. K. Sorger, *Mol. Syst. Biol.* **2017**, *13*, 954.

[26] V. Tshitoyan, J. Dagdelen, L. Weston, A. Dunn, Z. Rong, O. Kononova, K. A. Persson, G. Ceder, A. Jain, *Nature* **2019**, *571*, 95–98.

[27] P. Raccuglia, K. C. Elbert, P. D. F. Adler, C. Falk, M. B. Wenny, A. Mollo, M. Zeller, S. A. Friedler, J. Schrier, A. J. Norquist, *Nature* **2016**, *533*, 73–76.

[28] D. T. Ahneman, J. G. Estrada, S. Lin, S. D. Dreher, A. G. Doyle, *Science* **2018**, *360*, 186–190.

[29] A. F. Zahrt, J. J. Henle, B. T. Rose, Y. Wang, W. T. Darrow, S. E. Denmark, *Science* **2019**, *363*, eaau5631.

[30] J. P. Reid, M. S. Sigman, *Nature* **2019**, *571*, 343–348.

[31] W. F. Reinhart, A. W. Long, M. P. Howard, A. L. Ferguson, A. Z. Panagiotopoulos, *Soft Matter* **2017**, *13*, 4733–4745.

[32] A. Mardt, L. Pasquali, H. Wu, F. Noé, *Nat. Commun.* **2018**, *9*, 5.

[33] A. Rives, S. Goyal, J. Meier, D. Guo, M. Ott, C. L. Zitnick, J. Ma, R. Fergus, *bioRxiv* **2019**, 622803.

[34] D. C. Elton, Z. Boukouvalas, M. D. Fuge, P. W. Chung, *arXiv:1903.04388 [physics stat]* **2019**.

[35] D. Schwalbe-Koda, R. Gómez-Bombarelli, *arXiv:1907.01632 [physics stat]* **2019**.

[36] S. Kim et al., *Nucleic Acids Res.* **2019**, *47*, D1102–D1109.

[37] J. Hill, G. Mulholland, K. Persson, R. Seshadri, C. Wolverton, B. Meredig, *MRS Bull.* **2016**, *41*, 399–409.

[38] L. Himanen, A. Geurts, A. S. Foster, P. Rinke, *arXiv:1907.05644 [cond-mat physics:physics]* **2019**.

[39] D. J. Rigden, X. M. Fernández, *Nucleic Acids Res.* **2018**, *46*, D1–D7.

[40] leejunhyun, The Databases for Drug Discovery (DDD), **2019**, `https://github.com/LeeJunHyun/The-Databases-for-Drug-Discovery` (visited on 07/26/2019).

[41] M. Krier, G. Bret, D. Rognan, *J. Chem. Inf. Model.* **2006**, *46*, 512–524.

[42] S. R. Langdon, N. Brown, J. Blagg, *J. Chem. Inf. Model.* **2011**, *51*, 2174–2185.

[43] J. J. Irwin, B. K. Shoichet, *J. Chem. Inf. Model.* **2005**, *45*, 177–182.

[44] T. Sterling, J. J. Irwin, *J. Chem. Inf. Model.* **2015**, *55*, 2324–2337.

[45] ChemSpider | Search and share chemistry, `http://www.chemspider.com/` (visited on 02/12/2019).

[46] G. Papadatos et al., *Nucleic Acids Res.* **2016**, *44*, D1220–D1228.

[47] P. Banerjee, J. Erehman, B.-O. Gohlke, T. Wilhelm, R. Preissner, M. Dunkel, *Nucleic Acids Res.* **2015**, *43*, D935–D939.

[48] Synthetically Accessible Virtual Inventory (SAVI) Database Download Page, `https://cactus.nci.nih.gov/download/savi%5C_download/` (visited on 02/12/2019).

[49] eMolecules Database Download - eMolecules, https://www.emolecules.com/info/plus/download-database (visited on 07/31/2019).

[50] MolPort: Download Compound Database | Available Compounds, https://www.molport.com/shop/database-download (visited on 07/31/2019).

[51] REAL Compounds - Enamine, https://enamine.net/library-synthesis/real-compounds (visited on 07/25/2019).

[52] Chemspace | Compound Libraries, https://chem-space.com/compounds (visited on 07/31/2019).

[53] T. Fink, J.-L. Reymond, *J. Chem. Inf. Model.* **2007**, *47*, 342–353.

[54] L. C. Blum, J.-L. Reymond, *J. Am. Chem. Soc.* **2009**, *131*, 8732–8733.

[55] L. Ruddigkeit, R. van Deursen, L. C. Blum, J.-L. Reymond, *J. Chem. Inf. Model.* **2012**, *52*, 2864–2875.

[56] F. Chevillard, P. Kolb, *J. Chem. Inf. Model.* **2015**, *55*, 1824–1835.

[57] L. Humbeck, S. Weigang, T. Schäfer, P. Mutzel, O. Koch, *ChemMedChem* **2018**, *13*, 532–539.

[58] ChEMBL, https://www.ebi.ac.uk/chembl/ (visited on 02/12/2019).

[59] A. Gaulton et al., *Nucleic Acids Res.* **2012**, *40*, D1100–D1107.

[60] J. Hastings, G. Owen, A. Dekker, M. Ennis, N. Kale, V. Muthukrishnan, S. Turner, N. Swainston, P. Mendes, C. Steinbeck, *Nucleic Acids Res.* **2016**, *44*, D1214–D1219.

[61] RCSB PDB: Homepage, https://www.rcsb.org/ (visited on 02/12/2019).

[62] Welcome to PDBbind-CN Database, http://www.pdbbind.org.cn/ (visited on 02/12/2019).

[63] M. M. Gromiha, H. Uedaira, J. An, S. Selvaraj, P. Prabakaran, A. Sarai, *Nucleic Acids Res.* **2002**, *30*, 301–302.

[64] A. B. Keenan et al., *Cell Syst.* **2018**, *6*, 13–24.

[65] J. Jankauskaitė, B. Jiménez-García, J. Dapkūnas, J. Fernández-Recio, I. H. Moal, *Bioinformatics* **2019**, *35*, 462–469.

[66] xMoDEL: Molecular Dynamics Libraries | Molecular Modeling and Bioinformatics Group, http://mmb.pcb.ub.es/www/node/356 (visited on 02/12/2019).

[67] D. A. Benson, M. Cavanaugh, K. Clark, I. Karsch-Mizrachi, J. Ostell, K. D. Pruitt, E. W. Sayers, *Nucleic Acids Res.* **2018**, *46*, D41–D47.

[68] D. S. Wishart, *Nucleic Acids Res.* **2006**, *34*, D668–D672.

[69] M. K. Gilson, T. Liu, M. Baitaluk, G. Nicola, L. Hwang, J. Chong, *Nucleic Acids Res.* **2016**, *44*, D1045–D1053.

[70] S. Ekins, B. A. Bunin in *In Silico Models for Drug Discovery*, Springer, **2013**, pp. 139–154.

[71] D. J. Dix, K. A. Houck, M. T. Martin, A. M. Richard, R. W. Setzer, R. J. Kavlock, *Toxicol. Sci.* **2006**, *95*, 5–12.

[72] A. M. Richard et al., *Chem. Res. Toxicol.* **2016**, *29*, 1225–1251.

[73] R. R. Tice, C. P. Austin, R. J. Kavlock, J. R. Bucher, *Environ. Health Perspect.* **2013**, *121*, 756–765.

[74] Tox21 Data Browser, https://tripod.nih.gov/tox21/index (visited on 08/06/2019).

[75] D. Lowe, Chemical reactions from US patents (1976-Sep2016), **2017**.

[76] Pistachio, https://doi.org/10.1036/1097-8542.519800 (visited on 04/04/2019).

[77] Reaxys, https://www.reaxys.com (visited on 02/12/2019).

[78] Reactions - CASREACT - Answers to your chemical reaction questions | CAS, /support/documentation/reactions (visited on 02/12/2019).

[79] InfoChem - SPRESI - Storage and retrieval of chemical structure and reaction information - infochem, http://www.infochem.de/products/databases/spresi.shtml (visited on 02/12/2019).

[80] Databases - Librarians - Wiley Online Library, https://onlinelibrary.wiley.com/library-info/products/databases (visited on 02/12/2019).

[81] J. Gao, L. B. M. Ellis, L. P. Wackett, *Nucleic Acids Res.* **2010**, *38*, D488–D491.

[82] NIST Chemical Kinetics Database, https://kinetics.nist.gov/kinetics/index.jsp (visited on 07/31/2019).

[83] C. Steinbeck, S. Kuhn, *Phytochemistry* **2004**, *65*, 2711–2717.

[84] M. Rupp, A. Tkatchenko, K.-R. Müller, O. A. von Lilienfeld, *Phys. Rev. Lett.* **2012**, *108*, 058301.

[85] G. Montavon, M. Rupp, V. Gobre, A. Vazquez-Mayagoitia, K. Hansen, A. Tkatchenko, K.-R. Müller, O. Anatole von Lilienfeld, *New J. Phys.* **2013**, *15*, 095003.

[86] R. Ramakrishnan, P. O. Dral, M. Rupp, O. A. von Lilienfeld, *Scientific Data* **2014**, *1*, 140022.

[87] R. Ramakrishnan, M. Hartmann, E. Tapavicza, O. A. von Lilienfeld, *J. Chem. Phys.* **2015**, *143*, 084111.

[88] D. L. Mobley, J. P. Guthrie, *J. Comput.-Aided Mol. Des.* **2014**, *28*, 711–720.

[89] P. J. Linstrom, W. G. Mallard, *J. Chem. Eng. Data* **2001**, *46*, 1059–1063.

[90] S. Otsuka, I. Kuwajima, J. Hosoya, Y. Xu, M. Yamazaki in 2011 International Conference on Emerging Intelligent Data and Web Technologies, IEEE, **2011**, pp. 22–29.

[91] A. Merkys, A. Vaitkus, J. Butkus, M. Okulič-Kazarinas, V. Kairys, S. Gražulis, *J Appl Crystallogr* **2016**, *49*, 292–301.

[92] S. Gražulis, A. Merkys, A. Vaitkus, M. Okulič-Kazarinas, *J Appl Crystallogr* **2015**, *48*, 85–91.

[93] S. Gražulis, A. Daškevič, A. Merkys, D. Chateigner, L. Lutterotti, M. Quirós, N. R. Serebryanaya, P. Moeck, R. T. Downs, A. Le Bail, *Nucleic Acids Res* **2012**, *40*, D420–D427.

[94] Y. G. Chung, J. Camp, M. Haranczyk, B. J. Sikora, W. Bury, V. Krungleviciute, T. Yildirim, O. K. Farha, D. S. Sholl, R. Q. Snurr, *Chem. Mater.* **2014**, *26*, 6185–6192.

[95] C. E. Wilmer, M. Leaf, C. Y. Lee, O. K. Farha, B. G. Hauser, J. T. Hupp, R. Q. Snurr, *Nature Chem.* **2011**, *4*, 83–89.

[96] C. R. Groom, I. J. Bruno, M. P. Lightfoot, S. C. Ward, *Acta Cryst B* **2016**, *72*, 171–179.

[97] G. Bergerhoff, R. Hundt, R. Sievers, I. D. Brown, *J. Chem. Inf. Comput. Sci.* **1983**, *23*, 66–69.

[98] NOMAD Repository, https://repository.nomad-coe.eu/ (visited on 02/12/2019).

[99] S. Curtarolo et al., *Comput. Mater. Sci.* **2012**, *58*, 218–226.

[100] Aflow - Automatic - FLOW for Materials Discovery, http://aflowlib.org/ (visited on 02/12/2019).

[101] S. Kirklin, J. E. Saal, B. Meredig, A. Thompson, J. W. Doak, M. Aykol, S. Rühl, C. Wolverton, *Npj Comput. Mater.* **2015**, *1*, 15010.

[102] A. Jain et al., *APL Materials* **2013**, *1*, 011002.

[103] D. D. Landis, J. S. Hummelshoj, S. Nestorov, J. Greeley, M. Dulak, T. Bligaard, J. K. Norskov, K. W. Jacobsen, *Comput. Sci. Eng.* **2012**, *14*, 51–57.

[104] Projects — COMPUTATIONAL MATERIALS REPOSITORY, https://cmr.fysik.dtu.dk/ (visited on 02/12/2019).

[105] Pearson's Crystal Data, http://www.crystalimpact.com/pcd/ (visited on 02/12/2019).

[106] S. A. Lopez, E. O. Pyzer-Knapp, G. N. Simm, T. Lutzow, K. Li, L. R. Seress, J. Hachmann, A. Aspuru-Guzik, *Sci Data* **2016**, *3*, 160086.

[107] R. Lammey, *Sci. Ed.* **2015**, *2*, 22–27.

[108] pubmeddev, Home - PubMed - NCBI, https://www.ncbi.nlm.nih.gov/pubmed/ (visited on 02/12/2019).

[109] arXiv Bulk Data Access | arXiv e-print repository, https://arxiv.org/help/bulk_data (visited on 08/02/2019).

[110] Text and Data Mining Agreement - Wiley Online Library, `http://olabout.wiley.com/WileyCDA/Section/id-826542.html` (visited on 08/02/2019).

[111] Text and data mining policy - Elsevier, `https://www.elsevier.com/about/policies/text-and-data-mining` (visited on 08/02/2019).

[112] S. Ekins, A. M. Clark, S. J. Swamidass, N. Litterman, A. J. Williams, *J. Comput.-Aided Mol. Des.* **2014**, *28*, 997–1008.

[113] B. Hie, H. Cho, B. Berger, *Science* **2018**, *362*, 347–350.

[114] D. J. Audus, J. J. de Pablo, *ACS Macro Lett.* **2017**, *6*, 1078–1082.

[115] I. V. Tetko, O. Engkvist, U. Koch, J.-L. Reymond, H. Chen, *Mol. Inf.* **2016**, *35*, 615–621.

[116] S. Ekins, A. J. Williams, *Pharm. Res.* **2010**, *27*, 393–395.

[117] T. C. Norman, C. Bountra, A. M. Edwards, K. R. Yamamoto, S. H. Friend, *Sci. Transl. Med.* **2011**, *3*, 88mr1–88mr1.

[118] S. Ekins, A. M. Clark, A. J. Williams, *Mol. Inf.* **2012**, *31*, 585–597.

[119] M. Krallinger, F. Leitner, O. Rabal, M. Vazquez, J. Oyarzabal, A. Valencia, *J. Cheminform.* **2015**, *7*, S1.

[120] M. C. Swain, J. M. Cole, *J. Chem. Inf. Model.* **2016**, *56*, 1894–1904.

[121] M. Krallinger, O. Rabal, A. Lourenço, J. Oyarzabal, A. Valencia, *Chem. Rev.* **2017**, *117*, 7673–7761.

[122] E. Kim, K. Huang, A. Saunders, A. McCallum, G. Ceder, E. Olivetti, *Chem. Mater.* **2017**, *29*, 9436–9444.

[123] Z. Zhai, D. Q. Nguyen, S. A. Akhondi, C. Thorne, C. Druckenbrodt, T. Cohn, M. Gregory, K. Verspoor, *arXiv:1907.02679 [cs]* **2019**.

[124] S. Zheng, S. Dharssi, M. Wu, J. Li, Z. Lu in *Methods in Molecular Biology*, (Eds.: R. S. Larson, T. I. Oprea), Methods in Molecular Biology, Springer New York, New York, NY, **2019**, pp. 231–252.

[125] D. R. Swanson, N. R. Smalheiser, *Artif. Intell.*, Scientific Discovery **1997**, *91*, 183–203.

[126] A. Gomez-Perez, M. Martinez-Romero, A. Rodriguez-Gonzalez, G. Vazquez, J. M. Vazquez-Naya, Ontologies in Medicinal Chemistry: Current Status and Future Challenges, en, Text, **2013**.

[127] P. W. Battaglia et al., *arXiv:1806.01261 [cs stat]* **2018**.

[128] A. J. Williams, S. Elkins, V. Tkachenko, C. Lipinski, A. Tropsha, *Drug Discovery World* **2009**, *10*, 33–39.

[129] M. Jaskolski, *Acta Crystallogr D Biol Cryst* **2013**, *69*, 1865–1866.

[130] H. Berman, G. J. Kleywegt, H. Nakamura, J. L. Markley, *Acta Crystallogr D Biol Cryst* **2013**, *69*, 2297–2297.

[131] M. J. Sippl, *Proteins* **1993**, *17*, 355–362.

[132] Č. Venclovas, K. Ginalski, C. Kang, *Protein Sci.* **2004**, *13*, 1594–1602.

[133] Z. Wu, B. Ramsundar, E. N. Feinberg, J. Gomes, C. Geniesse, A. S. Pappu, K. Leswing, V. Pande, *Chem. Sci.* **2018**, *9*, 513–530.

[134] M. Haghighatlari, G. Vishwakarma, D. Altarawy, R. Subramanian, B. U. Kota, A. Sonpal, S. Setlur, J. Hachmann, **2019**, DOI `10.26434/chemrxiv.8323271.v1`.

[135] A. Krizhevsky, I. Sutskever, G. E. Hinton, *Commun. ACM* **2017**, *60*, 84–90.

[136] P. Hop, B. Allgood, J. Yu, *Mol. Pharmaceutics* **2018**, *15*, 4371–4377.

[137] D. K. Duvenaud, D. Maclaurin, J. Iparraguirre, R. Bombarell, T. Hirzel, A. Aspuru-Guzik, R. P. Adams in *Advances in Neural Information Processing Systems 28*, (Eds.: C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, R. Garnett), Curran Associates, Inc., **2015**, pp. 2224–2232.

[138] S. Kearnes, K. McCloskey, M. Berndl, V. Pande, P. Riley, *J. Comput.-Aided Mol. Des.* **2016**, *30*, 595–608.

[139] F. A. Faber, L. Hutchison, B. Huang, J. Gilmer, S. S. Schoenholz, G. E. Dahl, O. Vinyals, S. Kearnes, P. F. Riley, O. A. von Lilienfeld, *arXiv preprint arXiv:1704.01212* **2017**, *13*, 5255–5264.

[140] V. Korolev, A. Mitrofanov, A. Korotcov, V. Tkachenko, *arXiv:1906.06256 [physics]* **2019**.

[141] K. Yang et al., *J. Chem. Inf. Model.* **2019**, DOI 10.1021/acs.jcim.9b00237.

[142] R. Ramprasad, R. Batra, G. Pilania, A. Mannodi-Kanakkithodi, C. Kim, *Npj Comput. Mater.* **2017**, DOI 10.1038/s41524-017-0056-5.

[143] M. Krenn, F. Häse, A. Nigam, P. Friederich, A. Aspuru-Guzik, *arXiv:1905.13741 [physics physics:quant-ph stat]* **2019**.

[144] N. O'Boyle, A. Dalke, **2018**, DOI 10.26434/chemrxiv.7097960.v1.

[145] L. M. Ghiringhelli, J. Vybiral, S. V. Levchenko, C. Draxl, M. Scheffler, *Phys. Rev. Lett.* **2015**, *114*, 105503.

[146] J. P. Janet, H. J. Kulik, *J. Phys. Chem. A* **2017**, *121*, 8939–8954.

[147] H. Altae-Tran, B. Ramsundar, A. S. Pappu, V. Pande, *ACS Cent. Sci.* **2017**, *3*, 283–293.

[148] J. Li, T. Chen, K. Lim, L. Chen, S. A. Khan, J. Xie, X. Wang, *Advanced Intelligent Systems* **2019**, *1*, arXiv: 1811.02771, 1900029.

[149] Y. Zhang, C. Ling, *Npj Comput. Mater.* **2018**, *4*, DOI 10.1038/s41524-018-0081-z.

[150] E. J. Bjerrum, *arXiv preprint arXiv:1703.07076* **2017**.

[151] E. J. Bjerrum, B. Sattarov, *Biomolecules* **2018**, *8*, 131.

[152] A. P. Bartók, R. Kondor, G. Csányi, *Phys. Rev. B* **2013**, *87*, 184115.

[153] K. T. Schütt, P.-J. Kindermans, H. E. Sauceda, S. Chmiela, A. Tkatchenko, K.-R. Müller, *arXiv:1706.08566 [physics stat]* **2017**.

[154] M. Ragoza, J. Hochuli, E. Idrobo, J. Sunseri, D. R. Koes, *J. Chem. Inf. Model.* **2017**, *57*, 942–957.

[155] G. E. Dahl, N. Jaitly, R. Salakhutdinov, *arXiv preprint arXiv:1406.1231* **2014**.

[156] B. Ramsundar, S. Kearnes, P. Riley, D. Webster, D. Konerding, V. Pande, *arXiv:1502.02072 [cs stat]* **2015**.

[157] C. Fare, L. Turcani, E. O. Pyzer-Knapp, *arXiv:1809.06334 [physics stat]* **2018**.

[158] A. Gupta, A. T. Müller, B. J. H. Huisman, J. A. Fuchs, P. Schneider, G. Schneider, *Mol. Inf.* **2017**, *37*, 1700111.

[159] D. Merk, L. Friedrich, F. Grisoni, G. Schneider, *Mol. Inf.* **2018**, *37*, 1700153.

[160] D. B. Searls, *Nat. Rev. Drug Discov.* **2005**, *4*, 45–58.

[161] J. C. Sundaramurthi, S. Brindha, T. Reddy, L. E. Hanna, *Tuberculosis* **2012**, *92*, 133–138.

[162] G. B. Goh, N. O. Hodas, C. Siegel, A. Vishnu, *arXiv:1712.02034* **2017**.

[163] G. B. Goh, C. Siegel, A. Vishnu, N. O. Hodas, N. Baker, *arXiv:1706.06689 [cs stat]* **2017**.

[164] Q. Zhou, P. Tang, S. Liu, J. Pan, Q. Yan, S.-C. Zhang, *PNAS* **2018**, *115*, E6411–E6417.

[165] W. Hu, B. Liu, J. Gomes, M. Zitnik, P. Liang, V. Pande, J. Leskovec, *arXiv:1905.12265 [cs stat]* **2019**.

[166] R. Sawada, H. Iwata, S. Mizutani, Y. Yamanishi, *J. Chem. Inf. Model.* **2015**, *55*, 2717–2730.

[167] Y. Luo, X. Zhao, J. Zhou, J. Yang, Y. Zhang, W. Kuang, J. Peng, L. Chen, J. Zeng, *Nat. Commun.* **2017**, *8*, 573.

[168] K. V. Chuang, M. J. Keiser, *Science* **2018**, *362*, eaat8603.

[169] P. Polishchuk, *J. Chem. Inf. Model.* **2017**, *57*, 2618–2639.

[170] R. Roscher, B. Bohn, M. F. Duarte, J. Garcke, *arXiv preprint arXiv:1905.08883* **2019**.

[171] L. von Rueden, S. Mayer, J. Garcke, C. Bauckhage, J. Schuecker, *arXiv:1903.12394 [cs stat]* **2019**.

[172] H. Kitano, *AIMag* **2016**, *37*, 39.

[173] S. M. Paul, D. S. Mytelka, C. T. Dunwiddie, C. C. Persinger, B. H. Munos, S. R. Lindborg, A. L. Schacht, *Nat. Rev. Drug Discov.* **2010**, *9*, 203–214.

[174] M. Shevlin, *ACS Med. Chem. Lett.* **2017**, *8*, 601–607.

[175] A. Buitrago Santanilla et al., *Science* **2014**, *347*, 49–53.

[176] N. J. Gesmundo, B. Sauvagnat, P. J. Curran, M. P. Richards, C. L. Andrews, P. J. Dandliker, T. Cernak, *Nature* **2018**, *557*, 228–232.

[177] S. Oliver, L. Zhao, A. J. Gormley, R. Chapman, C. Boyer, *Macromolecules* **2018**, *52*, 3–23.

[178] M. L. Green, I. Takeuchi, J. R. Hattrick-Simpers, *J. Appl. Phys.* **2013**, *113*, 231101.

[179] R. K. O'Reilly, A. J. Turberfield, T. R. Wilks, *Acc. Chem. Res.* **2017**, *50*, 2496–2509.

[180] K. Troshin, J. F. Hartwig, *Science* **2017**, *357*, 175–181.

[181] N. P. Tu, A. W. Dombrowski, G. M. Goshu, A. Vasudevan, S. W. Djuric, Y. Wang, *Angew. Chem. Int. Ed.* **2019**, *58*, 7987–7991.

[182] Lowe, Derek, Automated Chemistry: A Vision, en-US, **2018**.

[183] J. Y. Pan, *ACS Med. Chem. Lett.* **2019**, *10*, 703–707.

[184] A. Baranczak, N. P. Tu, J. Marjanovic, P. A. Searle, A. Vasudevan, S. W. Djuric, *ACS Med. Chem. Lett.* **2017**, *8*, 461–465.

[185] L. M. Roch, F. Häse, C. Kreisbeck, T. Tamayo-Mendoza, L. P. E. Yunker, J. E. Hein, A. Aspuru-Guzik, *Sci. Robot.* **2018**, *3*, eaat5559.

[186] K. Machida, Y. Hirose, S. Fuse, T. Sugawara, T. Takahashi, *CHEMICAL & PHARMACEUTICAL BULLETIN* **2010**, *58*, 87–93.

[187] S. Steiner et al., *Science* **2018**, *363*, eaav2211.

[188] T. Jiang, S. Bordi, A. E. McMillan, K.-Y. Chen, F. Saito, P. Nichols, B. Wanner, J. Bode, **2019**, DOI 10.26434/chemrxiv.7882799.v1.

[189] A. G. Godfrey, T. Masquelin, H. Hemmerle, *Drug Discov. Today* **2013**, *18*, 795–802.

[190] J. Li, S. G. Ballmer, E. P. Gillis, S. Fujii, M. J. Schmidt, A. M. E. Palazzolo, J. W. Lehmann, G. F. Morehouse, M. D. Burke, *Science* **2015**, *347*, 1221–1226.

[191] J. Lyu et al., *Nature* **2019**, *566*, 224–229.

[192] J. S. Smith, O. Isayev, A. E. Roitberg, *Chem. Sci.* **2017**, *8*, 3192–3203.

[193] J. E. Stone, J. C. Phillips, P. L. Freddolino, D. J. Hardy, L. G. Trabuco, K. Schulten, *J Comput Chem* **2007**, *28*, 2618–2640.

[194] I. S. Ufimtsev, T. J. Martínez, *Comput. Sci. Eng.* **2008**, *10*, 26–34.

[195] J. E. Stone, D. J. Hardy, I. S. Ufimtsev, K. Schulten, *Journal of Molecular Graphics and Modelling* **2010**, *29*, 116–125.

[196] C. Yang et al., *arXiv:1905.05359 [cs]* **2019**.

[197] D. E. Shaw et al., *Commun. ACM* **2008**, *51*, 91–97.

[198] T. S. Hofer, *Front Chem* **2013**, *1*, DOI 10.3389/fchem.2013.00006.

[199] S. Grimme, P. R. Schreiner, *Angewandte Chemie International Edition* **2018**, *57*, 4170–4176.

[200] N. Malo, J. A. Hanley, S. Cerquozzi, J. Pelletier, R. Nadon, *Nat. Biotechnol.* **2006**, *24*, 167–175.

[201] M. Baker, *Nat. Methods* **2010**, *7*, 787–792.

[202] C. L. Allen, D. C. Leitch, M. S. Anson, M. A. Zajac, *Nature Catalysis* **2019**, *2*, 2–4.

[203] M. O'Brien, L. Konings, M. Martin, J. Heap, *Tetrahedron Lett.* **2017**, *58*, 2409–2413.

[204] M. Su, Q. Yang, Y. Du, G. Feng, Z. Liu, Y. Li, R. Wang, *J. Chem. Inf. Model.* **2019**, *59*, 895–913.

[205] J. P. Janet, F. Liu, A. Nandy, C. Duan, T. Yang, S. Lin, H. J. Kulik, *Inorg. Chem.* **2019**, DOI 10.1021/acs.inorgchem.9b00109.

[206] Z. Ghahramani, *Nature* **2015**, *521*, 452–459.

[207] T. Ueno, T. D. Rhone, Z. Hou, T. Mizoguchi, K. Tsuda, *Materials Discovery* **2016**, *4*, 18–21.

[208] A. A. Peterson, R. Christensen, A. Khorshidi, *Phys. Chem. Chem. Phys.* **2017**, *19*, 10978–10985.

[209] F. Häse, L. M. Roch, C. Kreisbeck, A. Aspuru-Guzik, *ACS Cent. Sci.* **2018**, *4*, 1134–1145.

[210] I. Cortés-Ciriano, A. Bender, *J. Chem. Inf. Model.* **2019**, *59*, 1269–1281.

[211] Y. Gal, Z. Ghahramani, 10.

[212] G. N. Simm, M. Reiher, *J. Chem. Theory Comput.* **2018**, *14*, 5238–5248.

[213] A. Tropsha, A. Golbraikh, *Current pharmaceutical design* **2007**, *13*, 3494–3504.

[214] A. Tropsha, *Mol. Inf.* **2010**, *29*, 476–488.

[215] M. Toplak, R. Močnik, M. Polajnar, Z. Bosnić, L. Carlsson, C. Hasselgren, J. Demšar, S. Boyer, B. Zupan, J. Stålring, *J. Chem. Inf. Model.* **2014**, *54*, 431–441.

[216] D. Stumpfe, J. Bajorath, *J. Med. Chem.* **2012**, *55*, 2932–2942.

[217] J. Bajorath, *Expert Opin. Drug Discovery* **2017**, *12*, 879–883.

[218] R. Liu, A. Wallqvist, *J. Chem. Inf. Model.* **2019**, *59*, 181–189.

[219] O. Obrezanova, G. Csányi, J. M. R. Gola, M. D. Segall, *J. Chem. Inf. Model.* **2007**, *47*, 1847–1857.

[220] P. Donmez, J. G. Carbonell in Proceeding of the 17th ACM conference on Information and knowledge mining - CIKM '08, Proceeding of the 17th ACM conference, ACM Press, Napa Valley, California, USA, **2008**, p. 619.

[221] U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, *AI magazine* **1996**, *17*, 37–37.

[222] Y. Bengio, A. C. Courville, P. Vincent, *CoRR abs/1206.5538* **2012**, *1*, 2012.

[223] M. A. F. Pimentel, D. A. Clifton, L. Clifton, L. Tarassenko, *Signal Processing* **2014**, *99*, 215–249.

[224] G. Marcus, *arXiv:1801.00631 [cs stat]* **2018**.

[225] M. A. Boden, *Artif. Intell.* **1998**, *103*, 347–356.

[226] B. Sanchez-Lengeling, A. Aspuru-Guzik, *Science* **2018**, *361*, 360–365.

[227] R. Gómez-Bombarelli, J. N. Wei, D. Duvenaud, J. M. Hernández-Lobato, B. Sánchez-Lengeling, D. Sheberla, J. Aguilera-Iparraguirre, T. D. Hirzel, R. P. Adams, A. Aspuru-Guzik, *ACS Cent. Sci.* **2018**, *4*, 268–276.

[228] G. L. Guimaraes, B. Sanchez-Lengeling, C. Outeiral, P. L. C. Farias, A. Aspuru-Guzik, **2017**.

[229] M. J. Kusner, B. Paige, J. M. Hernández-Lobato, *arXiv:1703.01925 [stat]* **2017**.

[230] W. Jin, R. Barzilay, T. Jaakkola, *arXiv:1802.04364* **2018**.

[231] N. De Cao, T. Kipf, *arXiv:1805.11973 [cs stat]* **2018**.

[232] Y. Li, O. Vinyals, C. Dyer, R. Pascanu, P. Battaglia, *arXiv:1803.03324 [cs stat]* **2018**.

[233] P. Ertl, A. Schuffenhauer, *J. Cheminform.* **2009**, *1*, 8.

[234] C. W. Coley, L. Rogers, W. H. Green, K. F. Jensen, *J. Chem. Inf. Model.* **2018**, *58*, 252–261.

[235] J. Bradshaw, B. Paige, M. J. Kusner, M. H. S. Segler, J. M. Hernández-Lobato, *arXiv:1906.05221 [physics stat]* **2019**.

[236] N. Brown, M. Fiscato, M. H. Segler, A. C. Vaucher, *arXiv:1811.09621 [physics q-bio]* **2018**, *59*, 1096–1108.

[237] D. Polykovskiy et al., **2018**.

[238] T. Aumentado-Armstrong, *arXiv:1809.02032 [cs q-bio]* **2018**.

[239] P. S. Gromski, A. B. Henson, J. M. Granda, L. Cronin, *Nat. Rev. Chem.* **2019**, *3*, 119–128.

[240] W. Czechtizky et al., *ACS Med. Chem. Lett.* **2013**, *4*, 768–772.

[241] S. Chow, S. Liver, A. Nelson, *Nat. Rev. Chem.* **2018**, *2*, 174–183.

[242] C. A. Nicolaou et al., *ACS Med. Chem. Lett.* **2019**, *10*, 278–286.

[243] S. K. Saikin, C. Kreisbeck, D. Sheberla, J. S. Becker, A.-G. A., *Expert Opin. Drug Discovery* **2019**, *14*, 1–4.

[244] A. Vempaty, L. R. Varshney, P. K. Varshney, *arXiv:1708.03833 [stat]* **2017**, arXiv: 1708.03833.

[245] A. Rogers, How the Transformers broke NLP leaderboards, **2019**.

[246] P. Langley in Discovey Science, (Eds.: S. Arikawa, H. Motoda), Springer Berlin Heidelberg, **1998**, pp. 25–39.

[247] P. Langley, *Int. J. Hum. Comput. Stud.* **2000**, *53*, 393–410.

[248] V. G. Honavar, *Review of Policy Research* **2014**, *31*, 326–330.

[249] A. M. Turing, *Mind* **1950**, *LIX*, 433–460.

[250] B. Writer, *Lithium-Ion Batteries*, Springer, **2019**.

[251] S. V. Ley, D. E. Fitzpatrick, R. J. Ingham, R. M. Myers, *Angew. Chem. Int. Ed.* **2015**, *54*, 3449–3464.

[252] D. Waltz, B. G. Buchanan, *Science* **2009**, *324*, 43–44.

[253] Y. Gil, M. Greaves, J. Hendler, H. Hirsh, *Science* **2014**, *346*, 171–172.

[254] K. Alberi et al., *J. Phys. D: Appl. Phys.* **2018**, *52*, 013001.

[255] G. Schneider, *Nat. Rev. Drug Discov.* **2017**, *17*, 97–113.

[256] T. Dimitrov, C. Kreisbeck, J. S. Becker, A. Aspuru-Guzik, S. K. Saikin, *ACS Appl. Mater. Interfaces* **2019**, *11*, 24825–24836.

[257] P. Friederich, A. Fediai, S. Kaiser, M. Konrad, N. Jung, W. Wenzel, *Advanced Materials* **2019**, *0*, 1808256.

[258] J. Vamathevan et al., *Nat. Rev. Drug Discov.* **2019**, *18*, 463–477.

[259] J.-P. Correa-Baena, K. Hippalgaonkar, J. van Duren, S. Jaffer, V. R. Chandrasekhar, V. Stevanovic, C. Wadia, S. Guha, T. Buonassisi, *Joule* **2018**, *2*, 1410–1420.

[260] D. P. Tabor et al., *Nat. Rev. Mater.* **2018**, *3*, 5–20.

[261] C. Glymour, *Daedalus* **2004**, *133*, 69–77.