ACCOUNTS of chemical research

pubs.acs.org/accounts Article

Data-Driven Strategies for Accelerated Materials Design

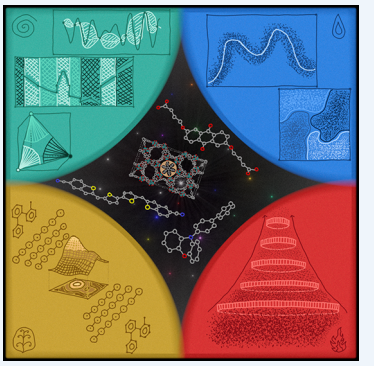
Published as part of the Accounts of Chemical Research special issue "Data Science Meets Chemistry".

Robert Pollice, Gabriel dos Passos Gomes, Matteo Aldeghi, Riley J. Hickman, Mario Krenn, Cyrille Lavigne, Michael Lindner-D'Addario, AkshatKumar Nigam, Cher Tian Ser, Zhenpeng Yao, and Alán Aspuru-Guzik*

Cite This: https://dx.doi.org/10.1021/acs.accounts.0c00785 Read Online

ACCESS | Metrics & More | Article Recommendations

CONSPECTUS: The ongoing revolution of the natural sciences by the advent of machine learning and artificial intelligence sparked significant interest in the material science community in recent years. The intrinsically high dimensionality of the space of realizable materials makes traditional approaches ineffective for large-scale explorations. Modern data science and machine learning tools developed for increasingly complicated problems are an attractive alternative. An imminent climate catastrophe calls for a clean energy transformation by overhauling current technologies within only several years of possible action available. Tackling this crisis requires the development of new materials at an unprecedented pace and scale. For example, organic photovoltaics have the potential to replace existing silicon-based materials to a large extent and open up new fields of application. In recent years, organic light-emitting diodes have emerged as state-of-the-art technology for digital screens and portable devices and are enabling new applications with flexible displays. Reticular frameworks allow the atom-precise synthesis of nanomaterials and promise to revolutionize the field by the potential to realize multifunctional nanoparticles with applications from gas storage, gas separation, and electrochemical energy storage to nanomedicine. In the recent decade, significant advances in all these fields have been facilitated by the comprehensive application of simulation and machine learning for property prediction, property optimization, and chemical space exploration enabled by considerable advances in computing power and algorithmic efficiency.



In this Account, we review the most recent contributions of our group in this thriving field of machine learning for material science. We start with a summary of the most important material classes our group has been involved in, focusing on small molecules as organic electronic materials and crystalline materials. Specifically, we highlight the data-driven approaches we employed to speed up discovery and derive material design strategies. Subsequently, our focus lies on the data-driven methodologies our group has developed and employed, elaborating on high-throughput virtual screening, inverse molecular design, Bayesian optimization, and supervised learning. We discuss the general ideas, their working principles, and their use cases with examples of successful implementations in data-driven material discovery and design efforts. Furthermore, we elaborate on potential pitfalls and remaining challenges of these methods. Finally, we provide a brief outlook for the field as we foresee increasing adaptation and implementation of large scale data-driven approaches in material discovery and design campaigns.

KEY REFERENCES

- Gómez-Bombarelli, R.; Aguilera-Iparraguirre, J.; Hirzel, T. D.; Duvenaud, D.; Maclaurin, D.; Blood-Forsythe, M. A.; Chae, H. S.; Einzinger, M.; Ha, D.-G.; Wu, T.; Markopoulos, G.; Jeon, S.; Kang, H.; Miyazaki, H.; Numata, M.; Kim, S.; Huang, W.; Hong, S. I.; Baldo, M.; Adams, R. P.; Aspuru-Guzik, A. Design of Efficient Molecular Organic Light-Emitting Diodes by a High-Throughput Virtual Screening and Experimental Approach. *Nat. Mater.* **2016**, *15* (10), 1120−1127.[1] *Realization of an integrated inverse design workflow from high-throughput virtual screening to device testing for organic light-emitting diode materials.*
- Yao, Z.; Sanchez-Lengeling, B.; Bobbitt, N. S.; Bucior, B. J.; Kumar, S. G. H.; Collins, S. P.; Burns, T.; Woo, T. K.; Farha, O.; Snurr, R. Q.; Aspuru-Guzik, A. Inverse Design of Nanoporous Crystalline Reticular Materials with Deep Generative Models. *Nat. Mach. Intell.* **2021**, *3*, 76.[2] *An automated nanoporous materials discovery platform powered by a supramolecular variational autoencoder was built and*

Received: December 7, 2020

ACS Publications © XXXX American Chemical Society

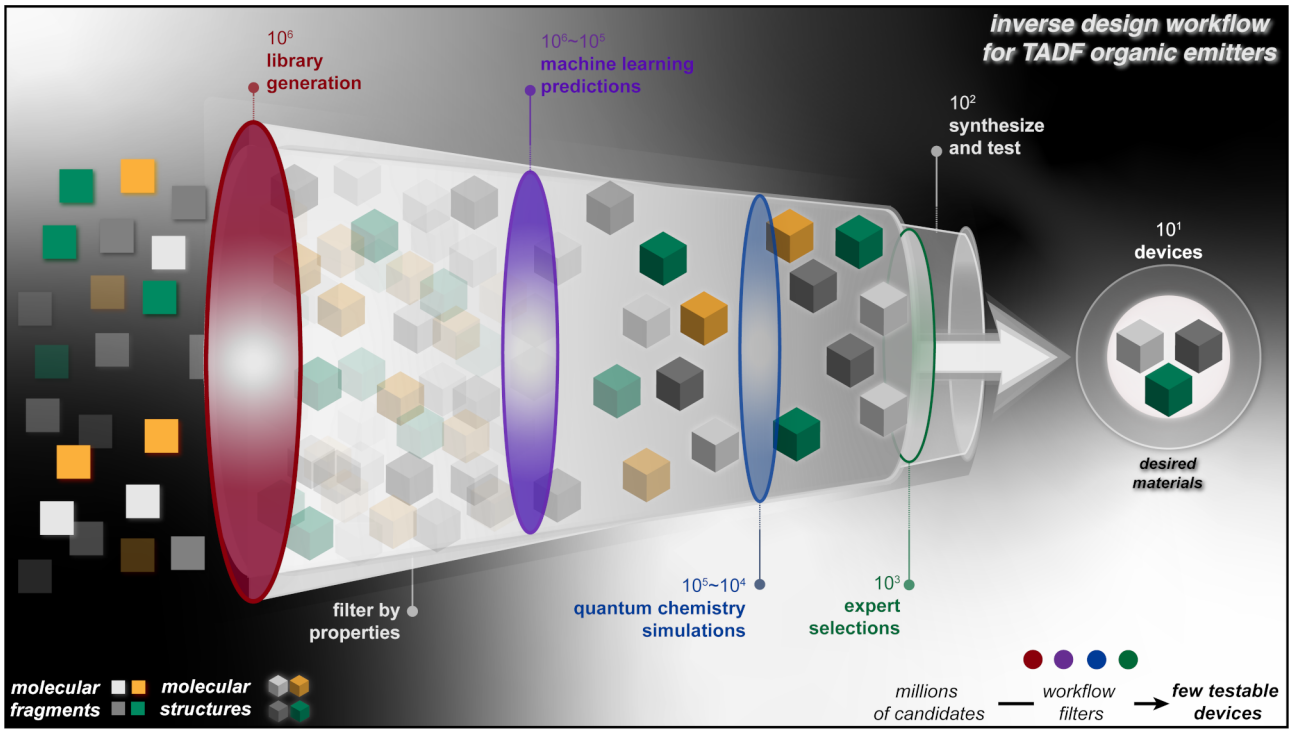


Figure 1. Inverse design workflow for thermally activated delayed fluorescence organic emitters from selecting fragments to device integration and testing.

demonstrated for the efficient exploration of the near infinite reticular chemical space and inverse design of reticular materials with desired functions like gas separation.

- Nigam, A.; Friederich, P.; Krenn, M.; Aspuru-Guzik, A. Augmenting Genetic Algorithms with Deep Neural Networks for Exploring the Chemical Space. In *International Conference on Learning Representations*; 2020..[3] *The proposal of a genetic algorithm enhanced by a neural network for inverse molecular design that can avoid convergence and bias molecule generation based on existing data sets.*
- Häse, F.; Roch, L. M.; Kreisbeck, C.; Aspuru-Guzik, A. Phoenics: A Bayesian Optimizer for Chemistry. *ACS Cent. Sci.* **20184** (9), 1134−1145.[4] *A probabilistic global optimization algorithm based on Bayesian kernel density estimation for the efficient parallel search of optimal experimental conditions.*

INTRODUCTION

The tremendous rise of data science and machine learning (ML) in the last decades led to the suggestion that it constitutes the fourth pillar of science.[5] While data has always been at the heart of research, current hardware enables its utilization at an unprecedented scale.[5] Accordingly, our group, the Matter Lab, has been using ML extensively to accelerate the discovery of new materials, especially for clean energy technologies to combat climate catastrophe and enable innovative technologies.

In this Account, we define discovery as observing a previously unknown natural phenomenon or object,[6,7] and design as rationally devising an object based on a particular plan.[8] Typically, discovery precedes and inspires materials design, as design requires at least minimal knowledge of the necessary features. Therefore, large scale discovery helps to speed up the establishment of material design principles, *i.e.*, heuristics to realize particular designs, because they enable identifying patterns in known matter with desired properties. In turn, successful design catalyzes the realization of new materials by restricting the search space to only the most promising regions in subsequent campaigns.

Herein, we review our work on organic electronic materials, crystalline materials, and data-driven methodologies for materials discovery and design, particularly high-throughput virtual screening, supervised learning, inverse molecular design, and Bayesian optimization. Moreover, we formulate general strategies for data-driven materials design our lab has adopted over the years and show how to implement them using ML. Finally, investigating these approaches critically, we propose typical use cases and highlight unsolved challenges.

APPLICATIONS

Organic Electronic Materials

One of our research foci has been organic electronic materials.[9] Compared to silicon-based electronics, they offer several advantages, including low cost, low density, high mechanical flexibility and toughness, low energy consumption, and easy processability. Further, chemical derivatization is well-established, making the accessible candidate space vast.

Accordingly, solar cells have experienced a remarkable surge because of the vast energy available from the sun and increasing efforts against a climate catastrophe. Organic photovoltaics[10] (OPVs) could replace commercial silicon-based devices if their power conversion efficiencies (PCEs) surpassed 10% and their lifetimes exceeded several thousands of hours. Notably, state-of-the-art OPVs reach 18% PCE in laboratory devices.[11] The Harvard Clean Energy Project (CEP) was initiated to find photoactive organic materials with high efficiencies.[12] Starting from 26 building blocks, selected based on expert knowledge to maximize performance and synthesizability,[13] 10^7 potential donors were generated. They were evaluated using high-

https://dx.doi.org/10.1021/acs.accounts.0c00785
Acc. Chem. Res. XXXX, XXX, XXX−XXX

throughput virtual screening (HTVS, *vide infra*) via increasingly expensive property predictions. First, the library was assessed using linear descriptor models constructed from experimental data. Subsequently, electronic structure calculations were performed, and PCEs were estimated using the Scharber model with a fullerene as acceptor.[14] That way, about 1000 candidates with estimated PCEs of 11% and higher were identified.

Additionally, statistical analysis of the top-performing molecules revealed design principles for photoactive donors identifying building blocks more likely to exhibit high performance. Notably, the screening efforts led to the experimental characterization of an organic crystal with one of the highest reported hole mobilities reported at the time.[15] Subsequently, extending the CEP to nonfullerene acceptors, over 51 000 candidates were generated based on 107 expertly chosen fragments.[16] More sophisticated property calibration with Gaussian processes and a modified Scharber model improved PCE predictions with a well-studied electron donor. Overall, 838 molecules with predicted PCEs of 8% or larger were found. Moreover, statistical analysis of the candidate structures was performed with respect to both Morgan fingerprints and the building blocks, establishing a general architecture for nonfullerene acceptors.

Similarly, organic light-emitting diodes[17] (OLEDs) have found wide adoption in small displays, are becoming prevalent in screens and lighting applications, and are entering the market in flexible displays. Thermally activated delayed fluorescence (TADF) emitters have become the main OLED class because of their high quantum efficiency, operational stability, and low cost. Their essential property is a small energy gap between the first excited singlet and triplet states so that energetically favored but nonemissive, triplet excitons can be upconverted to emissive singlet excitons. Based on knowledge about the TADF mechanism, our group carried out HTVS of emitters covering 10^6 candidates (Figure 1).[1] Key methodology included efficient quantum chemistry, calibrated against experiment via supervised learning (*vide infra*). Linear regression and neural networks were used for property predictions across the entire space.

Exploration was performed iteratively using a neural network to predict the most promising candidates, which were then simulated, minimizing evaluations. Not only were known emitters rediscovered, but new structures were also uncovered. Additionally, the systematic exploration exposed both established property trade-offs and unknown property limits. Moreover, the best leads were evaluated by human experts concerning synthesizability and novelty. Consequently, the most promising molecules after both computer and human-based evaluations were synthesized and incorporated into devices leading to high external quantum efficiencies of over 20%. This study serves as a prototype for the entire data-driven discovery pipeline from defining the candidate space to device integration.

Finally, renewable energy like wind and solar is intermittent, requiring large storage capacities to meet consumer demands. Redox-flow batteries (RFBs) resolve that by separating energy from power, enabling large grids to store immense amounts of energy scalable to varying demand loads.[18] Organic RFBs[19] (ORFBs) represent a sensible advancement, as redox-active organic electrolytes are tunable and cheaper than inorganic alternatives.[20] To identify ideal organic electrolytes, our group performed HTVS of quinones, which are well-known for their single-electron redox pairs.[21] The screening spanned 1710 single- and double-electron redox pairs to validate existing studies and find new redox couples.

The results indicated that quinone-exclusive electrolytes were promising aqueous ORFBs and revealed that functionalizations near the carbonyl groups largely affected redox potential and those away largely affected solubility. Subsequently, several experimental studies verified these predictions.[22,23] However, decomposition was found to deteriorate battery capacity irreversibly.[24] Hence, our group performed combined computational and experimental studies on the decomposition of quinones in aqueous environments.[18] HTVS was performed for over 140 000 redox pairs, including decomposition product analysis. The results identified a trade-off between redox potential, with a maximum near 0.95 V, close to experimental results at 0.85 V,[25] and stability. These results provide roadmaps for future studies, which are ongoing in our group, as the trade-off suggests that electrolyte stability must be considered.

Crystalline Materials

Crystalline energy storage materials with high energy density at low cost are cornerstones of renewable energy applications. For instance, multivalent calcium ion batteries[26] (CIBs) improve upon monovalent lithium-ion counterparts through increased capacities and higher material abundance while maintaining comparable operating voltages.[27] However, the development of CIBs is hindered by the failure of traditional graphite and calcium metal anodes due to the intercalation difficulty and the lack of efficient electrolytes. Recently, a high voltage (4.45 V) CIB cell using tin as the anode was reported to achieve a remarkable cyclability (over 300 cycles).[28]

Importantly, designing CIB anodes with improved performance requires a thorough exploration of the alloying space as calcium mixes with many elements. Hence, our group constructed a workflow to discover novel multivalent CIBs.[29] First, the tin electrochemical calciation reaction was investigated computationally and the reaction driving force as a function of calcium content was simulated. This exploration allowed the identification of threshold voltages governing the calciation limits. Consequently, a four-step screening strategy was adopted to look for high-performance CIB anodes. First, 357 metal−calcium binary and ternary compounds were identified from the Inorganic Crystal Structure Database (ICSD)[30] and further filtered to 115 candidates with existing decalciated metal/metalloid or binary intermetallic compounds. The calciation voltage profiles were calculated, and two threshold calciation voltages were defined, one stricter, based on the tin−calcium system, and the other more relaxed to account for potential differences in the driving force requirements. For each threshold, the maximum capacities, output voltages, volume expansions, and energy densities of the respective material were determined. Finally, metal−calcium systems with higher energy density than tin−calcium were identified, in which metalloids (Si, As, Sb, Ge), post-transition metals (Al, Pb, Cu, Cd, $CdCu_2$, Ga, Bi, In, Tl, Hg), and noble metals (Ag, Pt, Pd, Au) showed promise as alloying candidates for CIB anodes and calls for further experimental validations.

Additionally, reticular frameworks[31] (RFs), which include metal−organic frameworks (MOFs), are crystalline porous materials with high internal surface area and high stability and can be used for gas storage, gas separation, and electrochemical energy storage. They are constructed via self-assembly of molecular building blocks and exhibit a near-infinite combinatorial space, complicating their systematic exploration.

https://dx.doi.org/10.1021/acs.accounts.0c00785
Acc. Chem. Res. XXXX, XXX, XXX−XXX

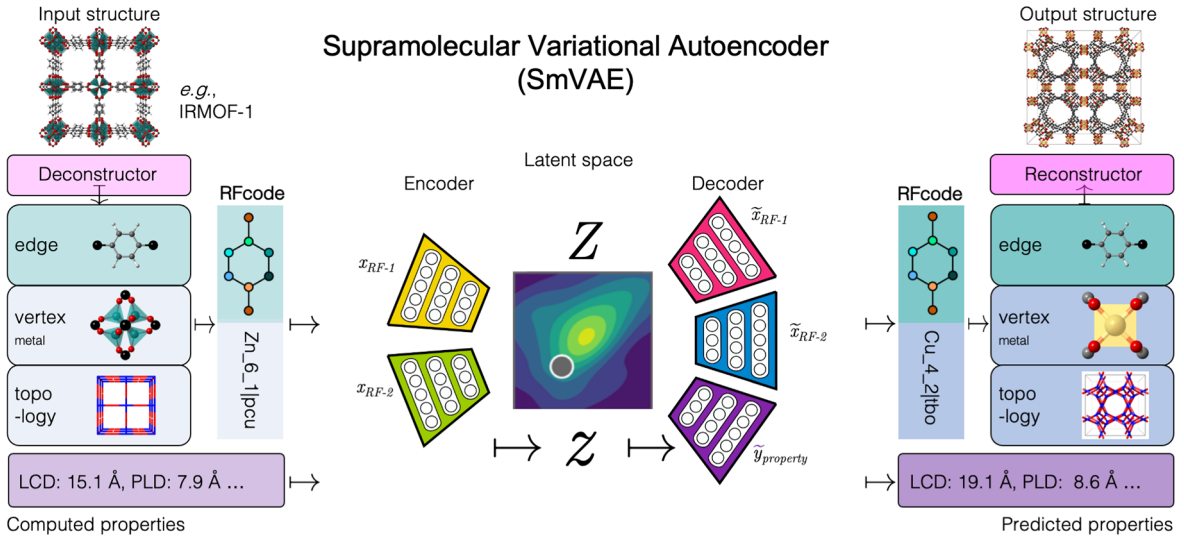


Figure 2. Automated reticular framework (RF) discovery platform using the supramolecular variational autoencoder (SmVAE). We construct the intermediate representation, RFcode, using unique, decomposed nets as a tuple of edges, vertices, and topologies. We consider the edges as SMILES, while vertices and topologies are categorical variables from known structures. SmVAE is a multicomponent variational autoencoder encoding and decoding each part of the RFcode separately ($\boldsymbol{x}_{\text{edge}} \rightarrow \tilde{\boldsymbol{x}}_{\text{edge}}$, $\boldsymbol{x}_{\text{RFcom}} \rightarrow \tilde{\boldsymbol{x}}_{\text{RFcom}}$). Structures are converted into/back from RFcode using the deconstructor/reconstructor, then transferred into continuous vectors ($\boldsymbol{z}$). To organize the latent space based on properties, we add a supervised model to predict properties ($\tilde{\boldsymbol{y}}_{\text{property}}$) based on labeled data ($\boldsymbol{y}$). Data from ref 2.

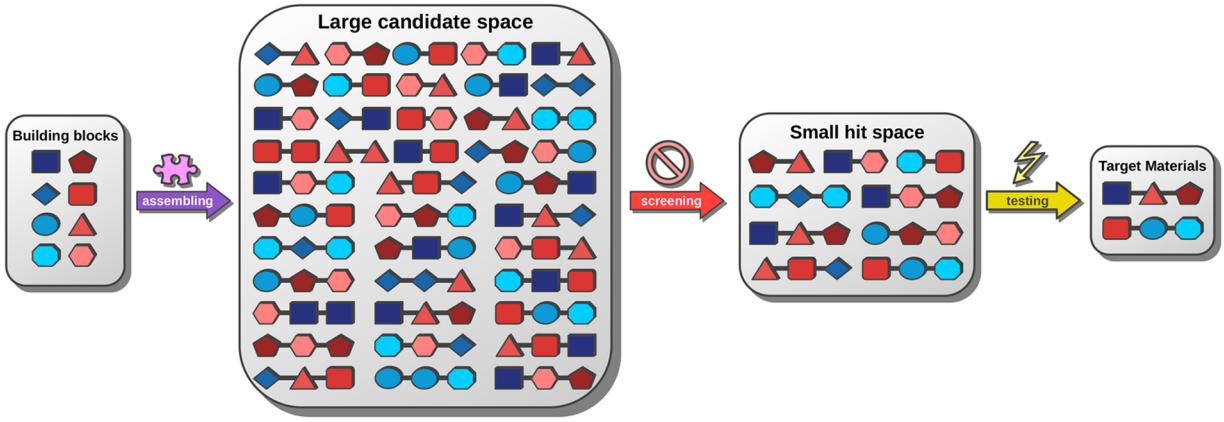


Figure 3. High-throughput virtual screening starts from a large space of candidates (*e.g.*, generated combinatorically, as illustrated). Using virtual screening, most candidates are eliminated, such that fewer (more expensive and time-consuming) experimental tests can be performed.

Recently, our group developed an invertible and efficient RF representation (Figure 2).[2,32] MOF fragments were extracted from the computation-ready, experimental (CoRE) MOF database[33] and augmented randomly with common functional groups. Furthermore, we added sets of multiconnected metal or organic nodes and sets of known MOF topologies generating a data set with around 2×10^6 MOF structures. Moreover, property simulations were performed for a random subset of about 40 000 MOF structures. The supramolecular variational autoencoder (SmVAE) with a MOF structure encoder-decoder, property prediction model, and framework generation algorithm was constructed with these structures (Figure 2), which can locate high performing MOFs through property optimization in the latent space. We demonstrated its capabilities for automatic design by proposing top candidates for gas separation adsorbent materials. We believe that the MOFs discovered are highly competitive against the best-performing MOFs/zeolites ever reported. Currently, their performance was validated using computational methods. Nevertheless, experimental verification is under way. Furthermore, the as-built platform can be applied to various supramolecular systems (*e.g.*, covalent-organic frameworks, coordination polymers, etc.) and applications (*e.g.*, batteries, catalysis, drug delivery).

METHODOLOGY

High-Throughput Virtual Screening (HTVS)

Virtual screening[34] denotes a selection process of candidate materials. Chemicals, either generated on-the-fly or from databases, are subject to simulations that estimate application-specific properties. Candidates failing computational tests are rejected, with the proviso that predicted performance is likely translatable to experimental performance. Thus, HTVS is a technique that reduces large candidate spaces to a manageable set of promising materials (Figure 3). In our search for new TADF emitters (*vide supra*),[1] the candidate space was narrowed down by 5 orders of magnitude via HTVS. Importantly, HTVS on large chemical spaces is inverse molecular design (*vide infra*) because, rather than designing structures directly, the computa-

https://dx.doi.org/10.1021/acs.accounts.0c00785
Acc. Chem. Res. XXXX, XXX, XXX−XXX

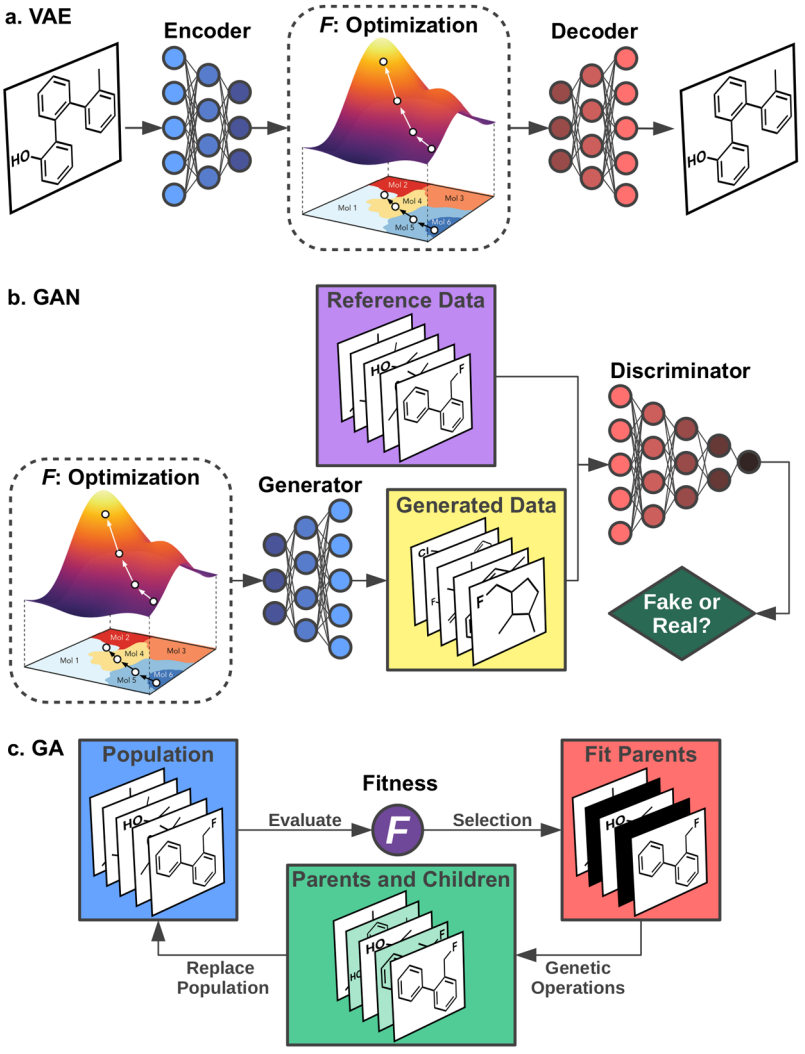


Figure 4. Inverse molecular design based on desired properties (F), with variational autoencoders (VAEs, a), generative adversarial networks (GANs, b), and genetic algorithms (GAs, c). Adapted with permission from ref 44. Copyright 2018 American Chemical Society.

tional tests and the candidate space are designed, which leads to the final hits based on the predicted properties.[35] Moreover, it can provide the basis for both generative and supervised models (*vide infra*), as they all rely on validated data.

Accordingly, HTVS is a powerful accelerator because computer simulation can be significantly less expensive than the respective experiments.[34] The continuing growth in computational power, which will soon reach the exascale, has made virtual screening highly scalable as it is embarrassingly parallel. Although HTVS is at least almost 20 years old,[36] it only recently started transforming materials science by advances in the accuracy and efficiency of density functional theory (DFT).[37] Besides computational cost, the main appeal of DFT was the possibility to tailor functional parameters to reproduce experiments, which increased its predictive power significantly.

For instance, linear response time-dependent DFT (TD-DFT) is accurate and computationally inexpensive for excited state properties. More importantly, it is robust, can be used in a black-box manner, and is readily deployed in simulations of tens of thousands of molecules with minimal failure rates.[14] However, one pernicious failure mode of TD-DFT is the description of excited states with significant double-excitation character, which is, *inter alia*, important in describing molecules with inverted singlet−triplet gaps,[38,39] such as the INVEST emitters recently described by our group.[40] Nevertheless, as computing power is increasing, more sophisticated *ab initio* approaches can be used in HTVS, allowing one to tackle ever more complicated problems and new material classes.

Yet, the impact of HTVS has been hampered by the difficulty in scaling the experimental confirmation of candidates,[1] as simulations feasible for high-throughput are still largely qualitative for condensed-phase properties.[41] A loose screen that accounts for computational inaccuracies minimizes false negatives, but the high cost of experimental validation means that almost all candidates must be rejected. The accuracy of computational screening can be maximized by implementing self-correcting filters such as checking whether simulations showed proper convergence catching false positives early on in the workflow. Nevertheless, ultimately, improvements in the experimental throughput are essential, calling for self-driving laboratories and closed-loop experimentation.[42,43]

AI-Powered Inverse Molecular Design

Inverse molecular design[35] starts at the desired properties and explores the chemical space to identify molecules optimizing them. Recently, various ML techniques have been employed to improve inverse molecular design, motivated by advances both

https://dx.doi.org/10.1021/acs.accounts.0c00785
Acc. Chem. Res. XXXX, XXX, XXX−XXX

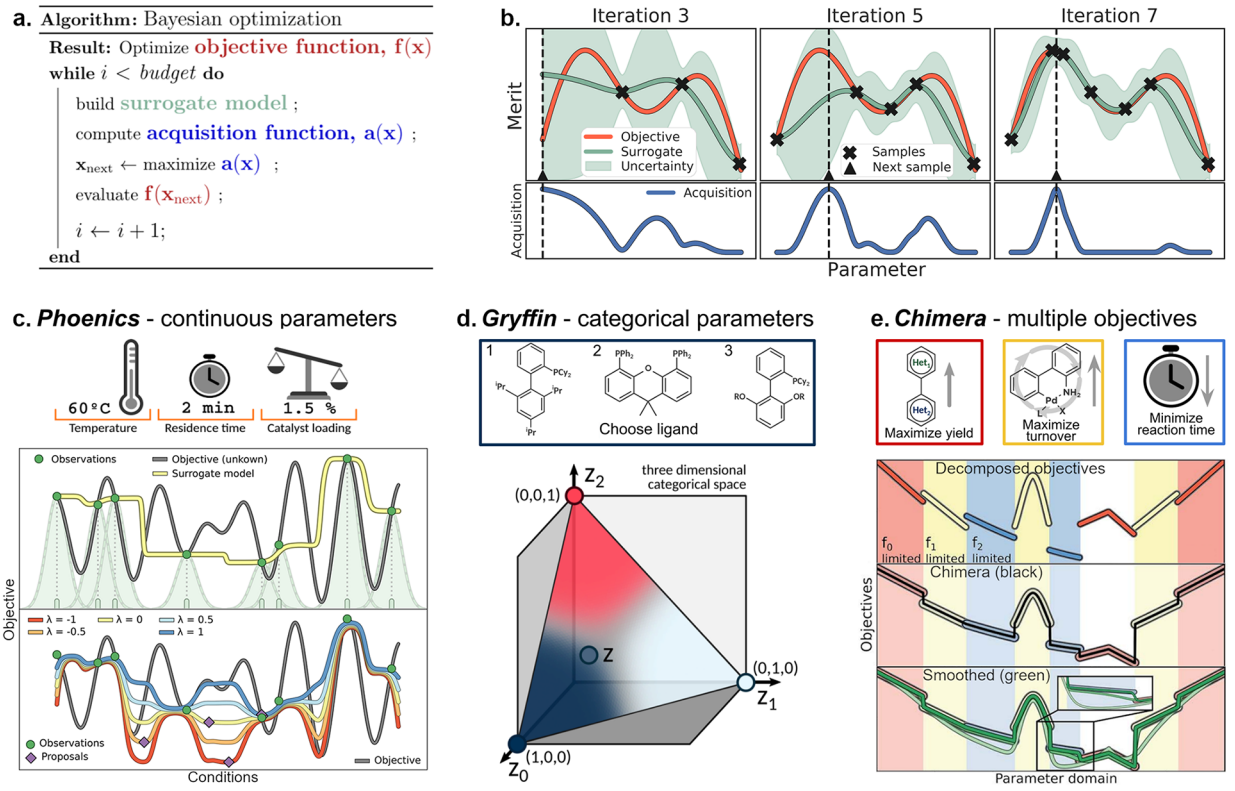


Figure 5. (a) General pseudocode for Bayesian optimization. (b) Visualization of Bayesian optimization of an objective function (red curve) using Gaussian processes. (c) Examples of continuous-valued parameters compatible with *Phoenics*, along with a sample surrogate model and acquisition functions generated by the algorithm. Adapted with permission from ref 4. Copyright 2018 American Chemical Society. (d) Depiction of the representation of a categorical variable in *Gryffin* with three options (*e.g.*, three ligands) on a simplex.[51] (e) Example of a multiobjective optimization problem for a chemical reaction, along with the construction of *Chimera* (bottom panel) from three 1-dimensional objective functions. Reproduced with permission from ref 52. Copyright 2018 Royal Society of Chemistry.

on the algorithmic (powerful ML libraries) and the hardware sides (GPU improvements for large neural networks). Importantly, inverse molecular design approaches can be separated roughly into two classes: model-based ML algorithms and evolutionary techniques.

Model-based ML algorithms for inverse design models use neural networks to learn patterns in molecular structures from existing data. After training, these models suggest new molecules covering important chemical features from the data set. Several methodologies exist. Herein we will discuss variational autoencoders (VAEs) and generative adversarial networks (GANs) because our group, to the best of our knowledge, was the first to apply these tools in chemistry. VAEs (Figure 4a) are capable of forming continuous (latent) spaces from discrete representations. They are trained to minimize the combined losses of latent space smoothness and input reconstruction enabling gradient-based optimization in the latent space. For inverse design, the latent space of VAEs is coupled with a property estimation model using supervised learning (*vide infra*).[44] Consequently, the latent space is arranged based on the property values allowing for a direct search of desired materials. GANs (Figure 4b) are generative models with joint training of two competing networks, a generator, and a discriminator. The generator produces examples from a high dimensional (often Gaussian) space, attempting to fool the discriminator, which tries to distinguish generated samples from reference structures. For molecules, our group proposed a sequential GAN (ORGAN), where the model is trained using reinforcement learning.[45] Desired molecular properties are used as a reward for generating good structures.

Notably, both VAEs and GANs are trained in a supervised way. Hence, they rely on existing data and mimic their distribution. Thus, they are limited in the exploration of the chemical space as compared to evolutionary techniques such as genetic algorithms (GAs, cf. Figure 4c). As its name implies, GAs are inspired by natural evolution. An initial population seeds the algorithm, each member being evaluated. The top-performing members proceed to the next iteration, the worst members are removed or replaced by better offspring. For inverse molecular design, the fitness function corresponds to the determination of desired molecular properties.

In contrast to deep learning-based models, GAs are not biased by user-defined data sets. Therefore, they are superior in unbiased explorations.[3] Recently, we have shown that GAs augmented with neural networks to estimate the similarity of a molecule with a given data set can explore specific structural classes without the large data requirements of GANs and VAEs. Additionally, neural network-based learning was used to detect and avoid local minima trapping the GA to amplify exploration by avoiding convergence.[3] Notably, this shows that ML-based inverse design techniques can be effectively combined with evolutionary algorithms.

Importantly, in all these approaches, molecular representation plays a crucial role. Molecular graphs are used for computational efficiency, as they avoid conformations. Simplified Molecular Input Line Entry System (SMILES)[46] strings are commonly used as a flat encoding of molecular graphs. However, they have

https://dx.doi.org/10.1021/acs.accounts.0c00785
Acc. Chem. Res. XXXX, XXX, XXX−XXX

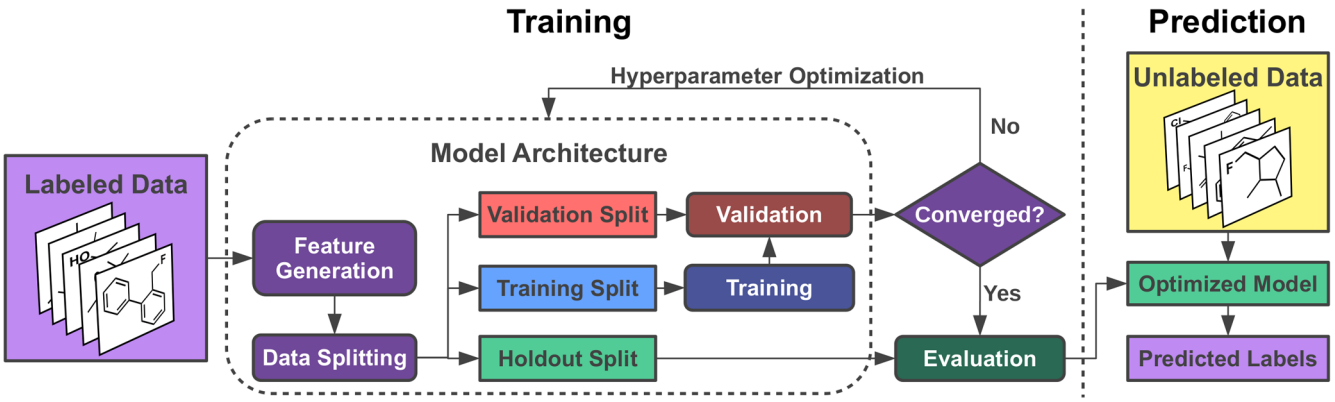


Figure 6. Workflow for supervised learning of molecular properties. A known (labeled) data set is used to optimize a model, which is subsequently used to estimate molecular properties for an unknown (unlabeled) data set.

a complex structure making a large fraction of molecules decoded from arbitrary SMILES invalid. This problem was solved recently by our group in a fundamental way by replacing SMILES with SELFIES (Self-Referencing Embedded Strings),[47] which is available on GitHub.[48] SELFIES is a 100% valid molecular string representation suitable as input for any inverse-design algorithm that outperformed alternative approaches in many benchmarks, such as validity and diversity of generated molecules, molecular density in the latent space of VAEs, or molecular optimization tasks with GAs.[3]

Bayesian Optimization

Several tasks across chemistry can be framed as optimization problems, where controllable parameters optimizing a desired objective are sought. For materials, such optimizations are challenging, as they are typically high-dimensional, nonconvex, and subject to noise and the objectives are expensive to evaluate. Suitable optimization strategies ought to be sample-efficient, global, and noise-tolerant. That is, they need to identify optimal parameter choices with as few measurements as possible, be able to escape local minima, and mitigate the detrimental effect of noise. A plethora of experiment planning strategies for optimization are currently available,[49] from traditional design of experiment to evolutionary and heuristic approaches. Among these, Bayesian optimization[50] (BO) has emerged as the strategy that best meets these requirements.

BO is an experiment planning algorithm that, in contrast to most other approaches, uses an ML model to learn from previous observations before suggesting the next iteration (Figure 5a).[50] In its most widely adopted form, BO employs techniques such as Gaussian processes to build a *surrogate* model that captures the features of the underlying objective function. Based on this surrogate, an *acquisition* function is defined, which determines the strategy used to propose new experiments (Figure 5b). Just like BO formulations using different ML models exist, various acquisition functions have been developed. Due to the use of an ML model, BO is sample-efficient. It is also noise-tolerant, as these models explicitly account for it. Finally, BO is a global approach that balances the *exploitation* of the best local optima identified with the *exploration* of unprobed areas of parameter space.

Typical BO approaches are inherently sequential and require heavy computations for each iteration. Therefore, BO can be unduly expensive when used in conjunction with high-throughput evaluations. Thus, our group has developed *Phoenics* (Figure 5c), a linear-scaling BO approach that supports parallel experiments.[4] *Phoenics* employs Bayesian neural networks (BNNs) to build a kernel density estimate of the objective function, and its acquisition function allows to select batches of evaluations to be run in parallel. Importantly, *Phoenics* is suitable for the optimization of continuous parameters, such as temperature and concentration. To also optimize categorical parameters, such as the choice of solvent, we developed *Gryffin* (Figure 5d), which uses categorical kernel densities that can be relaxed to continuous ones.[51] In addition, *Gryffin* allows for expert knowledge, in the form of descriptors for each categorical choice, to be provided to improve the optimization efficiency. Often, multiple competing objectives are present in materials science. *Chimera* (Figure 5e) is a general-purpose approach to multiobjective optimization.[52] It allows defining a hierarchy of objective preferences, which are combined into a single function to be optimized with any algorithm of choice.

Importantly, all the aforementioned algorithms can be combined with automated laboratories to enable autonomous experimentation.[42] These self-driving platforms are able to execute closed-loop workflows for the self-optimization of materials and processes. However, this requires robust software connections between automated hardware and experiment planning methods. *ChemOS* is a flexible, modular, open source and portable *Python* package that provides this interface between experiment planning and automated experiments.[53,54] Accordingly, in our laboratory, we have deployed *ChemOS*, together with *Phoenics*, *Gryffin*, and *Chimera*, for the autonomous optimization of manufacturing processes of thin-film materials,[55] multicomponent polymer OPV blends,[56] and reaction conditions of stereoselective Suzuki coupling.[57]

Supervised Learning

The costs associated with property measurement, from both experiments and simulations, are a major obstacle to the widespread expansion of HTVS, optimization, and inverse design. All of these techniques require some form of data acquisition, *i.e.*, simulations, measurements, or data mining. However, adapting experimental design to suit the needs of automated protocols is challenging, despite self-driving approaches likely being overall cost-effective. The promise of accurate and practically free inference of new results from existing data via supervised learning is a major driver of the ongoing ML revolution in the physical sciences.[58]

Supervised learning requires a data set of features and labels.[59] For molecular property prediction, this data set contains molecules in a specific representation (features) and their

https://dx.doi.org/10.1021/acs.accounts.0c00785
Acc. Chem. Res. XXXX, XXX, XXX−XXX

corresponding properties (labels). First, the data set is split into three, training, validation and holdout sets. The model is trained stepwise on the training set, usually by gradient descent or related algorithms. In general, hyperparameters, *i.e.*, choice of features, training set, and model architecture, influence predictive performance. These hyperparameters are optimized by maximizing prediction accuracy on the validation set. Eventually, model performance is evaluated via prediction accuracy for the holdout set, and the final model can be used to predict properties for unlabeled molecules. The entire workflow is illustrated in Figure 6. Our group developed several model architectures for supervised learning of molecular properties, most notably graph convolutional neural networks.[60,61]

Importantly, supervised learning has been used successfully for materials discovery. For example, our group used the CEP data set for property prediction.[62] After training on more than 200 000 molecules, a neural network predicted the result of DFT calculations consistently at a fraction of the computational expense. Additionally, our group applied this approach to reduce the number of simulations in HTVS significantly, with training on a set of similar size.[1] Moreover, our group also used Gaussian process regression to calibrate for systematic errors in DFT.[16] Crucially, in these studies, ML algorithms, representations, acquisition of training data, and validation procedures for models were tightly integrated with an understanding of the problem space, as opposed to sole reliance on existing data from various sources. We believe these considerations are key when it comes to the practical application of ML in chemistry.

Moreover, fruitful applications of supervised learning in materials science start from well-defined scientific goals. In contrast, the excitement brought upon by ML has generated many studies that focus on learning performance rather than scientific objectives. Generally, this is based on the (debatable and often unsupported) idea that performance metrics on one data set are transferable to other data sets or related problems. However, ML algorithms are highly parametrized and thus can readily overfit.[63] Indeed, the model choice can itself become a form of overfitting, especially when done on performance considerations alone.[64] Moreover, training data bias can contaminate predictions[65] but accounting for these biases appropriately is problem-specific. Furthermore, many studies are focused on error estimates obtained from statistical measures such as cross-validation. Although validation error can be a useful guide to the true prediction error on new data, it is not a replacement for it[66] and is often too optimistic.[67] In many ways, these issues arise when focus on the scientific goals is lost, as ultimately the best test of supervised learning is whether it solves problems.

CONCLUSION AND OUTLOOK

In this Account, we have reviewed data-driven approaches our group has employed for the design of materials, especially for clean energy applications, in the past decade. One of the first large scale campaigns our group embarked on was the CEP, where we implemented supervised learning together with HTVS using quantum chemistry simulations to investigate 10^7 potential donor molecules for organic solar cells and devised design principles by statistical analysis of structure−function relationships.[12] In the subsequent years, we refined these ML strategies and expanded our efforts toward other important materials such as OLEDs, OFRBs, multivalent CIBs, and RFs. In all these projects, data-driven workflows were key to speed up both the discovery and the design of new materials.

However, we believe that the full potential of data-driven strategies is yet to be unleashed. For instance, many properties are currently not investigated in HTVS because of their prohibitive computational cost. One such property is molecular stability with respect to common decomposition pathways. The associated problem is the huge dimensionality of potential reactions molecules can undergo, which greatly exceeds the chemical compound space in complexity. Recently, our group developed a method for the automatic discovery of chemical reactions based on the selection of reactive internal coordinates such as weak chemical bonds.[68] We believe this approach, together with empirical rules or heuristics for selecting reactive internal coordinates, could be used for HTVS of reactivity and stability of materials, and research in that direction is ongoing. Other properties too prohibitive for HTVS include the influence of explicit solvation on spectroscopic properties and the direct simulation of amorphous solid-state structures and properties. The main challenge therein is the large number of particles and degrees of freedom in the model systems and the associated multitude of interactions.

Furthermore, some of the methodologies we developed have only been tested on benchmark problems but are yet to be employed in real applications. Particularly, the genetic algorithm augmented with neural networks using SELFIES as molecular representation[47] our group proposed recently has outperformed most alternative generative models in benchmarks. However, it has yet to be implemented for designing functional materials, and we are actively working on that.[3] Finally, one of the most critical challenges of ML is model interpretability. Typically, supervised learning approaches are employed in a black box fashion without gaining insight into what the model actually learned. However, our group has shown recently that regression methods such as gradient boosting, when trained on molecular graph features, can be used to reveal important chemical moieties influencing the properties.[69,70] The trained model can be interpreted by human experts and rationalizing the feature importance can lead to new scientific understanding. We believe that similar approaches have the potential to change the way science is carried out in the near future.

However, the bottleneck of materials design campaigns is experimental synthesis and characterization, usually by a large margin.[71] Any material, no matter how good its (predicted) performance, needs to be synthesized for it to be used in real life. In particular for clean energy applications, material syntheses need to be performed on a huge scale requiring reliable, safe and green chemical processes. Accordingly, the continuing speed-up in computer power providing unprecedented prediction capabilities needs to be paralleled by increased experimental throughput. Accelerating materials design ultimately requires close integration of computer simulation, ML and experimentation in self-driving platforms, which our group termed Materials Acceleration Platforms (MAPs).[43]

One essential feature of MAPs is a closed-loop materials discovery workflow incorporating experimentation, computation, and human intuition. Online characterization techniques in conjunction with automated robotic synthesis[72−74] are central enabling technologies in these platforms. Making and measuring molecules on-demand in a feedback loop with self-correcting computational screening and ML is key to finding true "needle-in-a-haystack" materials. Currently, our group is implementing such a MAP for the realization of innovative materials making use of robust cross coupling chemistry, parallel robotic synthesis, and in-line characterization of spectroscopic properties coupled

https://dx.doi.org/10.1021/acs.accounts.0c00785
Acc. Chem. Res. XXXX, XXX, XXX−XXX

with computer simulation and ML. Details of this implementation will be described in an upcoming Account our group is working on in due course. Accordingly, the data-driven methods described above are a stepping stone to accelerate materials design. However, to realize their true potential, they need to percolate into experimental systems, and we are looking forward to witnessing applications of these methods in closed-loop experimental material design campaigns in the near future.

AUTHOR INFORMATION

Corresponding Author

Alán Aspuru-Guzik − *Chemical Physics Theory Group, Department of Chemistry and Department of Computer Science, University of Toronto, Toronto, Ontario M5S 3H6, Canada; Vector Institute for Artificial Intelligence, Toronto, Ontario M5G 1M1, Canada; Lebovic Fellow, Canadian Institute for Advanced Research (CIFAR), Toronto, Ontario M5G, Canada;* orcid.org/0000-0002-8277-4434; Email: aspuru@utoronto.ca

Authors

Robert Pollice − *Chemical Physics Theory Group, Department of Chemistry and Department of Computer Science, University of Toronto, Toronto, Ontario M5S 3H6, Canada;* orcid.org/0000-0001-8836-6266

Gabriel dos Passos Gomes − *Chemical Physics Theory Group, Department of Chemistry and Department of Computer Science, University of Toronto, Toronto, Ontario M5S 3H6, Canada;* orcid.org/0000-0002-8235-5969

Matteo Aldeghi − *Chemical Physics Theory Group, Department of Chemistry and Department of Computer Science, University of Toronto, Toronto, Ontario M5S 3H6, Canada; Vector Institute for Artificial Intelligence, Toronto, Ontario M5G 1M1, Canada;* orcid.org/0000-0003-0019-8806

Riley J. Hickman − *Chemical Physics Theory Group, Department of Chemistry and Department of Computer Science, University of Toronto, Toronto, Ontario M5S 3H6, Canada*

Mario Krenn − *Chemical Physics Theory Group, Department of Chemistry and Department of Computer Science, University of Toronto, Toronto, Ontario M5S 3H6, Canada; Vector Institute for Artificial Intelligence, Toronto, Ontario M5G 1M1, Canada*

Cyrille Lavigne − *Chemical Physics Theory Group, Department of Chemistry and Department of Computer Science, University of Toronto, Toronto, Ontario M5S 3H6, Canada;* orcid.org/0000-0003-2778-1866

Michael Lindner-D'Addario − *Chemical Physics Theory Group, Department of Chemistry and Department of Computer Science, University of Toronto, Toronto, Ontario M5S 3H6, Canada*

AkshatKumar Nigam − *Chemical Physics Theory Group, Department of Chemistry and Department of Computer Science, University of Toronto, Toronto, Ontario M5S 3H6, Canada;* orcid.org/0000-0002-5152-2082

Cher Tian Ser − *Chemical Physics Theory Group, Department of Chemistry and Department of Computer Science, University of Toronto, Toronto, Ontario M5S 3H6, Canada*

Zhenpeng Yao − *Chemical Physics Theory Group, Department of Chemistry and Department of Computer Science, University of Toronto, Toronto, Ontario M5S 3H6, Canada;* orcid.org/0000-0001-8286-8257

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.accounts.0c00785

Author Contributions

M.A., R.J.H., M.K., C.L., M.L.-D., A.K.N., C.T.S., and Z.Y. contributed equally to this work. R.P. and G.P.G. conceived the general outline and structure of this manuscript, and all authors contributed toward refining the structure. The manuscript was written through contributions of all authors. All authors have approved the final version of the manuscript.

Notes

The authors declare the following competing financial interest(s): A.A.-G. is co-founder and Chief Visionary Officer of Kebotix, Inc.

Biographies

Robert Pollice is an SNSF postdoctoral fellow at the University of Toronto.

Gabriel dos Passos Gomes is an NSERC Banting postdoctoral fellow at the University of Toronto.

Matteo Aldeghi is a postdoctoral fellow at the Vector Institute for Artificial Intelligence and the University of Toronto.

Riley J. Hickman is a PhD student at the University of Toronto.

Mario Krenn is an Erwin Schrödinger postdoctoral fellow at the University of Toronto and the Vector Institute for Artificial Intelligence.

Cyrille Lavigne is a postdoctoral fellow at the University of Toronto.

Michael Lindner-D'Addario is a PhD student at the University of Toronto.

AkshatKumar Nigam is a researcher at the University of Toronto.

Cher Tian Ser is a PhD student at the University of Toronto.

Zhenpeng Yao is a postdoctoral fellow at the University of Toronto.

Alán Aspuru-Guzik is a Professor of Chemistry and Computer Science at the University of Toronto, a Canada 150 Research Chair in Theoretical Chemistry, a Canada CIFAR AI Chair at the Vector Institute, a CIFAR Lebovic Fellow in the Biologically Inspired Solar Energy program, and a Google Industrial Research Chair in Quantum Computing.

ACKNOWLEDGMENTS

We thank all our co-workers and collaborators who contributed to the projects highlighted in this account. R.P. acknowledges funding through a Postdoc.Mobility fellowship by the Swiss National Science Foundation (SNSF, Project No. 191127). G.P.G gratefully acknowledges the Natural Sciences and Engineering Research Council of Canada (NSERC) for the Banting Postdoctoral Fellowship. R.J.H. gratefully acknowledges NSERC for provision of the Postgraduate Scholarships-Doctoral Program (PGSD3-534584-2019). M.K. acknowledges support from the Austrian Science Fund (FWF) through the Erwin Schrödinger fellowship No. J4309. M.L.-D. gratefully acknowledges the Fonds de Recherche Quebec Nature et Technologies (FRQNT) for the B1X Master's Scholarship. M.L.-D. also acknowledges support from the Queen Elizabeth II Graduate Scholarship in Science and Technology (QEII-GSST). We acknowledge the Defense Advanced Research Projects Agency (DARPA) under the Accelerated Molecular Discovery Program under Cooperative Agreement No. HR00111920027 dated

https://dx.doi.org/10.1021/acs.accounts.0c00785
Acc. Chem. Res. XXXX, XXX, XXX−XXX

August 1, 2019. The content of the information presented in this work does not necessarily reflect the position or the policy of the Government. A.A.-G. thanks Anders G. Frøseth for his generous support. A.A.-G. also acknowledges the generous support of Natural Resources Canada and the Canada 150 Research Chairs program. We also acknowledge the Department of Navy award (N00014-19-1-2134) issued by the Office of Naval Research. The United States Government has a royalty-free license throughout the world in all copyrightable material contained herein. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the Office of Naval Research.

REFERENCES

(1) Gómez-Bombarelli, R.; Aguilera-Iparraguirre, J.; Hirzel, T. D.; Duvenaud, D.; Maclaurin, D.; Blood-Forsythe, M. A.; Chae, H. S.; Einzinger, M.; Ha, D.-G.; Wu, T.; Markopoulos, G.; Jeon, S.; Kang, H.; Miyazaki, H.; Numata, M.; Kim, S.; Huang, W.; Hong, S. I.; Baldo, M.; Adams, R. P.; Aspuru-Guzik, A. Design of Efficient Molecular Organic Light-Emitting Diodes by a High-Throughput Virtual Screening and Experimental Approach. *Nat. Mater.* **2016**, *15* (10), 1120−1127.

(2) Yao, Z.; Sanchez-Lengeling, B.; Bobbitt, N. S.; Bucior, B. J.; Kumar, S. G. H.; Collins, S. P.; Burns, T.; Woo, T. K.; Farha, O.; Snurr, R. Q.; Aspuru-Guzik, A. Inverse Design of Nanoporous Crystalline Reticular Materials with Deep Generative Models. *Nat. Mach. Intell.* **2021**, *3*, 76.

(3) Nigam, A.; Friederich, P.; Krenn, M.; Aspuru-Guzik, A. Augmenting Genetic Algorithms with Deep Neural Networks for Exploring the Chemical Space. In *International Conference on Learning Representations*; 2020.

(4) Häse, F.; Roch, L. M.; Kreisbeck, C.; Aspuru-Guzik, A. Phoenics: A Bayesian Optimizer for Chemistry. *ACS Cent. Sci.* **2018**, *4* (9), 1134−1145.

(5) Hey, T.; Tansley, S.; Tolle, K. *The Fourth Paradigm: Data-Intensive Scientific Discovery*; Microsoft Research: Redmond, WA, 2009.

(6) Schindler, S. Scientific Discovery: That-Whats and What-Thats. *Ergo, an Open Access Journal of Philosophy* **2015**, *2*, 123−148.

(7) Kuhn, T. S. Historical Structure of Scientific Discovery. *Science* **1962**, *136* (3518), 760−764.

(8) March, S. T.; Smith, G. F. Design and Natural Science Research on Information Technology. *Decision Support Systems* **1995**, *15* (4), 251−266.

(9) Ostroverkhova, O. Organic Optoelectronic Materials: Mechanisms and Applications. *Chem. Rev.* **2016**, *116* (22), 13279−13412.

(10) Hedley, G. J.; Ruseckas, A.; Samuel, I. D. W. Light Harvesting for Organic Photovoltaics. *Chem. Rev.* **2017**, *117* (2), 796−837.

(11) Liu, Q.; Jiang, Y.; Jin, K.; Qin, J.; Xu, J.; Li, W.; Xiong, J.; Liu, J.; Xiao, Z.; Sun, K.; Yang, S.; Zhang, X.; Ding, L. 18% Efficiency Organic Solar Cells. *Science Bulletin* **2020**, *65* (4), 272−275.

(12) Hachmann, J.; Olivares-Amaya, R.; Atahan-Evrenk, S.; Amador-Bedolla, C.; Sánchez-Carrera, R. S.; Gold-Parker, A.; Vogt, L.; Brockway, A. M.; Aspuru-Guzik, A. The Harvard Clean Energy Project: Large-Scale Computational Screening and Design of Organic Photovoltaics on the World Community Grid. *J. Phys. Chem. Lett.* **2011**, *2* (17), 2241−2251.

(13) Olivares-Amaya, R.; Amador-Bedolla, C.; Hachmann, J.; Atahan-Evrenk, S.; Sánchez-Carrera, R. S.; Vogt, L.; Aspuru-Guzik, A. Accelerated Computational Discovery of High-Performance Materials for Organic Photovoltaics by Means of Cheminformatics. *Energy Environ. Sci.* **2011**, *4* (12), 4849−4861.

(14) Hachmann, J.; Olivares-Amaya, R.; Jinich, A.; Appleton, A. L.; Blood-Forsythe, M. A.; Seress, L. R.; Román-Salgado, C.; Trepte, K.; Atahan-Evrenk, S.; Er, S.; Shrestha, S.; Mondal, R.; Sokolov, A.; Bao, Z.; Aspuru-Guzik, A. Lead Candidates for High-Performance Organic Photovoltaics from High-Throughput Quantum Chemistry − the Harvard Clean Energy Project. *Energy Environ. Sci.* **2014**, *7* (2), 698−704.

(15) Sokolov, A. N.; Atahan-Evrenk, S.; Mondal, R.; Akkerman, H. B.; Sánchez-Carrera, R. S.; Granados-Focil, S.; Schrier, J.; Mannsfeld, S. C. B.; Zoombelt, A. P.; Bao, Z.; Aspuru-Guzik, A. From Computational Discovery to Experimental Characterization of a High Hole Mobility Organic Crystal. *Nat. Commun.* **2011**, *2* (1), 437.

(16) Lopez, S. A.; Sanchez-Lengeling, B.; de Goes Soares, J.; Aspuru-Guzik, A. Design Principles and Top Non-Fullerene Acceptor Candidates for Organic Photovoltaics. *Joule* **2017**, *1* (4), 857−870.

(17) Zou, S.-J.; Shen, Y.; Xie, F.-M.; Chen, J.-D.; Li, Y.-Q.; Tang, J.-X. Recent Advances in Organic Light-Emitting Diodes: Toward Smart Lighting and Displays. *Mater. Chem. Front.* **2020**, *4* (3), 788−820.

(18) Tabor, D. P.; Gomez-Bombarelli, R.; Tong, L.; Gordon, R. G.; Aziz, M. J.; Aspuru-Guzik, A. Mapping the Frontiers of Quinone Stability in Aqueous Media: Implications for Organic Aqueous Redox Flow Batteries. *J. Mater. Chem. A* **2019**, *7* (20), 12833−12841.

(19) Luo, J.; Hu, B.; Hu, M.; Zhao, Y.; Liu, T. L. Status and Prospects of Organic Redox Flow Batteries toward Sustainable Energy Storage. *ACS Energy Lett.* **2019**, *4* (9), 2220−2240.

(20) Lin, K.; Gómez-Bombarelli, R.; Beh, E. S.; Tong, L.; Chen, Q.; Valle, A.; Aspuru-Guzik, A.; Aziz, M. J.; Gordon, R. G. A Redox-Flow Battery with an Alloxazine-Based Organic Electrolyte. *Nature Energy* **2016**, *1* (9), 1−8.

(21) Er, S.; Suh, C.; Marshak, M. P.; Aspuru-Guzik, A. Computational Design of Molecules for an All-Quinone Redox Flow Battery. *Chemical Science* **2015**, *6* (2), 885−893.

(22) Yang, Z.; Tong, L.; Tabor, D. P.; Beh, E. S.; Goulet, M.-A.; De Porcellinis, D.; Aspuru-Guzik, A.; Gordon, R. G.; Aziz, M. J. Alkaline Benzoquinone Aqueous Flow Battery for Large-Scale Storage of Electrical Energy. *Adv. Energy Mater.* **2018**, *8* (8), 1702056.

(23) Kwabi, D. G.; Lin, K.; Ji, Y.; Kerr, E. F.; Goulet, M.-A.; De Porcellinis, D.; Tabor, D. P.; Pollack, D. A.; Aspuru-Guzik, A.; Gordon, R. G.; Aziz, M. J. Alkaline Quinone Flow Battery with Long Lifetime at PH 12. *Joule* **2018**, *2* (9), 1894−1906.

(24) Goulet, M.-A.; Tong, L.; Pollack, D. A.; Tabor, D. P.; Odom, S. A.; Aspuru-Guzik, A.; Kwan, E. E.; Gordon, R. G.; Aziz, M. J. Extending the Lifetime of Organic Flow Batteries via Redox State Management. *J. Am. Chem. Soc.* **2019**, *141* (20), 8014−8019.

(25) Hoober-Burkhardt, L.; Krishnamoorthy, S.; Yang, B.; Murali, A.; Nirmalchandar, A.; Prakash, G. K. S.; Narayanan, S. R. A New Michael-Reaction-Resistant Benzoquinone for Aqueous Organic Redox Flow Batteries. *J. Electrochem. Soc.* **2017**, *164* (4), A600.

(26) Arroyo-de Dompablo, M. E.; Ponrouch, A.; Johansson, P.; Palacín, M. R. Achievements, Challenges, and Prospects of Calcium Batteries. *Chem. Rev.* **2020**, *120* (14), 6331−6357.

(27) Ponrouch, A.; Frontera, C.; Bardé, F.; Palacín, M. R. Towards a Calcium-Based Rechargeable Battery. *Nat. Mater.* **2016**, *15* (2), 169−172.

(28) Wang, M.; Jiang, C.; Zhang, S.; Song, X.; Tang, Y.; Cheng, H.-M. Reversible Calcium Alloying Enables a Practical Room-Temperature Rechargeable Calcium-Ion Battery with a High Discharge Voltage. *Nat. Chem.* **2018**, *10* (6), 667−672.

(29) Yao, Z.; Hegde, V. I.; Aspuru-Guzik, A.; Wolverton, C. Discovery of Calcium-Metal Alloy Anodes for Reversible Ca-Ion Batteries. *Adv. Energy Mater.* **2019**, *9* (9), 1802994.

(30) Belsky, A.; Hellenbrandt, M.; Karen, V. L.; Luksch, P. New Developments in the Inorganic Crystal Structure Database (ICSD): Accessibility in Support of Materials Research and Design. *Acta Crystallogr., Sect. B: Struct. Sci.* **2002**, *58* (3−1), 364−369.

(31) Lyu, H.; Ji, Z.; Wuttke, S.; Yaghi, O. M. Digital Reticular Chemistry. *Chem.* **2020**, *6* (9), 2219−2241.

(32) Bucior, B. J.; Rosen, A. S.; Haranczyk, M.; Yao, Z.; Ziebel, M. E.; Farha, O. K.; Hupp, J. T.; Siepmann, J. I.; Aspuru-Guzik, A.; Snurr, R. Q. Identification Schemes for Metal−Organic Frameworks To Enable Rapid Search and Cheminformatics Analysis. *Cryst. Growth Des.* **2019**, *19* (11), 6682−6697.

(33) Chung, Y. G.; Camp, J.; Haranczyk, M.; Sikora, B. J.; Bury, W.; Krungleviciute, V.; Yildirim, T.; Farha, O. K.; Sholl, D. S.; Snurr, R. Q. Computation-Ready, Experimental Metal−Organic Frameworks: A

https://dx.doi.org/10.1021/acs.accounts.0c00785
Acc. Chem. Res. XXXX, XXX, XXX−XXX

Tool To Enable High-Throughput Screening of Nanoporous Crystals. *Chem. Mater.* **2014**, *26* (21), 6185−6192.
(34) Pyzer-Knapp, E. O.; Suh, C.; Gómez-Bombarelli, R.; Aguilera-Iparraguirre, J.; Aspuru-Guzik, A. What Is High-Throughput Virtual Screening? A Perspective from Organic Materials Discovery. *Annu. Rev. Mater. Res.* **2015**, *45* (1), 195−216.
(35) Sanchez-Lengeling, B.; Aspuru-Guzik, A. Inverse Molecular Design Using Machine Learning: Generative Models for Matter Engineering. *Science* **2018**, *361* (6400), 360−365.
(36) Schapira, M.; Raaka, B. M.; Das, S.; Fan, L.; Totrov, M.; Zhou, Z.; Wilson, S. R.; Abagyan, R.; Samuels, H. H. Discovery of Diverse Thyroid Hormone Receptor Antagonists by High-Throughput Docking. *Proc. Natl. Acad. Sci. U. S. A.* **2003**, *100* (12), 7354−7359.
(37) Ceder, G.; Chiang, Y.-M.; Sadoway, D. R.; Aydinol, M. K.; Jang, Y.-I.; Huang, B. Identification of Cathode Materials for Lithium Batteries Guided by First-Principles Calculations. *Nature* **1998**, *392* (6677), 694−696.
(38) de Silva, P. Inverted Singlet−Triplet Gaps and Their Relevance to Thermally Activated Delayed Fluorescence. *J. Phys. Chem. Lett.* **2019**, *10* (18), 5674−5679.
(39) Ehrmaier, J.; Rabe, E. J.; Pristash, S. R.; Corp, K. L.; Schlenker, C. W.; Sobolewski, A. L.; Domcke, W. Singlet−Triplet Inversion in Heptazine and in Polymeric Carbon Nitrides. *J. Phys. Chem. A* **2019**, *123* (38), 8099−8108.
(40) Pollice, R.; Friederich, P.; Lavigne, C.; dos Passos Gomes, G.; Aspuru-Guzik, A. Organic Molecules with Inverted Gaps between First Excited Singlet and Triplet States and Appreciable Fluorescence Rates. *ChemRxiv*, October 29, 2020, ver. 1. DOI: 10.26434/chemrxiv.13087319.v1.
(41) Chen, J.; Chan, B.; Shao, Y.; Ho, J. How Accurate Are Approximate Quantum Chemical Methods at Modelling Solute−Solvent Interactions in Solvated Clusters? *Phys. Chem. Chem. Phys.* **2020**, *22* (7), 3855−3866.
(42) Häse, F.; Roch, L. M.; Aspuru-Guzik, A. Next-Generation Experimentation with Self-Driving Laboratories. *TRECHEM* **2019**, *1* (3), 282−291.
(43) Flores-Leonar, M. M.; Mejía-Mendoza, L. M.; Aguilar-Granda, A.; Sanchez-Lengeling, B.; Tribukait, H.; Amador-Bedolla, C.; Aspuru-Guzik, A. Materials Acceleration Platforms: On the Way to Autonomous Experimentation. *Current Opinion in Green and Sustainable Chemistry* **2020**, *25*, 100370.
(44) Gómez-Bombarelli, R.; Wei, J. N.; Duvenaud, D.; Hernández-Lobato, J. M.; Sánchez-Lengeling, B.; Sheberla, D.; Aguilera-Iparraguirre, J.; Hirzel, T. D.; Adams, R. P.; Aspuru-Guzik, A. Automatic Chemical Design Using a Data-Driven Continuous Representation of Molecules. *ACS Cent. Sci.* **2018**, *4* (2), 268−276.
(45) Guimaraes, G. L.; Sanchez-Lengeling, B.; Outeiral, C.; Farias, P. L. C.; Aspuru-Guzik, A. Objective-Reinforced Generative Adversarial Networks (ORGAN) for Sequence Generation Models. *arXiv (Machine Learning)*, February 7, 2018, 1705.10843, ver. 3.
(46) Weininger, D. SMILES, a Chemical Language and Information System. 1. Introduction to Methodology and Encoding Rules. *J. Chem. Inf. Model.* **1988**, *28* (1), 31−36.
(47) Krenn, M.; Hase, F.; Nigam, A.; Friederich, P.; Aspuru-Guzik, A. Self-Referencing Embedded Strings (SELFIES): A 100% Robust Molecular String Representation. *Mach. Learn.: Sci. Technol.* **2020**, *1*, 045024.
(48) Aspuru-Guzik-Group/Selfies: aspuru-guzik-group/selfies. *GitHub*, 2020. https://github.com/aspuru-guzik-group/selfies.
(49) Häse, F.; Aldeghi, M.; Hickman, R. J.; Roch, L. M.; Christensen, M.; Liles, E.; Hein, J. E.; Aspuru-Guzik, A. Olympus: A Benchmarking Framework for Noisy Optimization and Experiment Planning. *arXiv (Machine Learning)*, October 8, 2020, 2010.04153, ver. 1.
(50) Shahriari, B.; Swersky, K.; Wang, Z.; Adams, R. P.; de Freitas, N. Taking the Human Out of the Loop: A Review of Bayesian Optimization. *Proc. IEEE* **2016**, *104* (1), 148−175.
(51) Häse, F.; Roch, L. M.; Aspuru-Guzik, A. Gryffin: An Algorithm for Bayesian Optimization for Categorical Variables Informed by Physical Intuition with Applications to Chemistry. *arXiv (Machine Learning)*, March 26, 2020, 2003.12127, ver 1.
(52) Häse, F.; Roch, L. M.; Aspuru-Guzik, A. Chimera: Enabling Hierarchy Based Multi-Objective Optimization for Self-Driving Laboratories. *Chem. Sci.* **2018**, *9* (39), 7642−7655.
(53) Roch, L. M.; Häse, F.; Kreisbeck, C.; Tamayo-Mendoza, T.; Yunker, L. P. E.; Hein, J. E.; Aspuru-Guzik, A. ChemOS: Orchestrating Autonomous Experimentation. *Sci. Rob.* **2018**, *3* (19), eaat5559.
(54) Roch, L. M.; Häse, F.; Kreisbeck, C.; Tamayo-Mendoza, T.; Yunker, L. P. E.; Hein, J. E.; Aspuru-Guzik, A. ChemOS: An Orchestration Software to Democratize Autonomous Discovery. *PLoS One* **2020**, *15* (4), e0229862.
(55) MacLeod, B. P.; Parlane, F. G. L.; Morrissey, T. D.; Häse, F.; Roch, L. M.; Dettelbach, K. E.; Moreira, R.; Yunker, L. P. E.; Rooney, M. B.; Deeth, J. R.; Lai, V.; Ng, G. J.; Situ, H.; Zhang, R. H.; Elliott, M. S.; Haley, T. H.; Dvorak, D. J.; Aspuru-Guzik, A.; Hein, J. E.; Berlinguette, C. P. Self-Driving Laboratory for Accelerated Discovery of Thin-Film Materials. *Science Advances* **2020**, *6* (20), eaaz8867.
(56) Langner, S.; Häse, F.; Perea, J. D.; Stubhan, T.; Hauch, J.; Roch, L. M.; Heumueller, T.; Aspuru-Guzik, A.; Brabec, C. J. Beyond Ternary OPV: High-Throughput Experimentation and Self-Driving Laboratories Optimize Multicomponent Systems. *Adv. Mater.* **2020**, *32* (14), 2070110.
(57) Christensen, M.; Yunker, L. P. E.; Adedeji, F.; Häse, F.; Roch, L. M.; Gensch, T.; dos Passos Gomes, G.; Zepel, T.; Sigman, M. S.; Aspuru-Guzik, A. Data-Science Driven Autonomous Process Optimization. *ChemRxiv* November 2, 2020, ver 1. DOI: 10.26434/chemrxiv.13146404.v1.
(58) Butler, K. T.; Davies, D. W.; Cartwright, H.; Isayev, O.; Walsh, A. Machine Learning for Molecular and Materials Science. *Nature* **2018**, *559* (7715), 547−555.
(59) Goodfellow, I.; Bengio, Y.; Courville, A. *Deep Learning*; MIT Press, 2016.
(60) Duvenaud, D. K.; Maclaurin, D.; Iparraguirre, J.; Bombarell, R.; Hirzel, T.; Aspuru-Guzik, A.; Adams, R. P. Convolutional Networks on Graphs for Learning Molecular Fingerprints. In *Advances in Neural Information Processing Systems 28*; Cortes, C., Lawrence, N. D., Lee, D. D., Sugiyama, M., Garnett, R., Eds.; Curran Associates, Inc., 2015; pp 2224−2232.
(61) Flam-Shepherd, D.; Wu, T.; Friederich, P.; Aspuru-Guzik, A. Neural Message Passing on High Order Paths. *arXiv (Machine Learning)*, February 24, 2020, 2002.10413, ver. 1.
(62) Pyzer-Knapp, E. O.; Li, K.; Aspuru-Guzik, A. Learning from the Harvard Clean Energy Project: The Use of Neural Networks to Accelerate Materials Discovery. *Adv. Funct. Mater.* **2015**, *25* (41), 6495−6502.
(63) Roelofs, R.; Shankar, V.; Recht, B.; Fridovich-Keil, S.; Hardt, M.; Miller, J.; Schmidt, L. A Meta-Analysis of Overfitting in Machine Learning. In *Advances in Neural Information Processing Systems 32*; Wallach, H., Larochelle, H., Beygelzimer, A., Alché-Buc, F., Fox, E., Garnett, R., Eds.; Curran Associates, Inc., 2019; pp 9179−9189.
(64) Cawley, G. C.; Talbot, N. L. C. On Over-Fitting in Model Selection and Subsequent Selection Bias in Performance Evaluation. *J. Mach. Learn. Res.* **2010**, *11* (70), 2079−2107.
(65) Ambroise, C.; McLachlan, G. J. Selection Bias in Gene Extraction on the Basis of Microarray Gene-Expression Data. *Proc. Natl. Acad. Sci. U. S. A.* **2002**, *99* (10), 6562−6566.
(66) Dupuy, A.; Simon, R. M. Critical Review of Published Microarray Studies for Cancer Outcome and Guidelines on Statistical Analysis and Reporting. *J. Natl. Cancer Inst* **2007**, *99* (2), 147−157.
(67) Shi, L.; Campbell, G.; Jones, W. D.; Campagne, F.; Wen, Z.; Walker, S. J.; Su, Z.; Chu, T.-M.; Goodsaid, F. M.; Pusztai, L.; Shaughnessy, J. D.; Oberthuer, A.; Thomas, R. S.; Paules, R. S.; Fielden, M.; Barlogie, B.; Chen, W.; Du, P.; Fischer, M.; Furlanello, C.; Gallas, B. D.; Ge, X.; Megherbi, D. B.; Symmans, W. F.; Wang, M. D.; Zhang, J.; Bitter, H.; Brors, B.; Bushel, P. R.; Bylesjo, M.; Chen, M.; Cheng, J.; Cheng, J.; Chou, J.; Davison, T. S.; Delorenzi, M.; Deng, Y.; Devanarayan, V.; Dix, D. J.; Dopazo, J.; Dorff, K. C.; Elloumi, F.; Fan, J.; Fan, S.; Fan, X.; Fang, H.; Gonzaludo, N.; Hess, K. R.; Hong, H.;

https://dx.doi.org/10.1021/acs.accounts.0c00785
Acc. Chem. Res. XXXX, XXX, XXX−XXX

983 Huan, J.; Irizarry, R. A.; Judson, R.; Juraeva, D.; Lababidi, S.; Lambert,
984 C. G.; Li, L.; Li, Y.; Li, Z.; Lin, S. M.; Liu, G.; Lobenhofer, E. K.; Luo, J.;
985 Luo, W.; McCall, M. N.; Nikolsky, Y.; Pennello, G. A.; Perkins, R. G.;
986 Philip, R.; Popovici, V.; Price, N. D.; Qian, F.; Scherer, A.; Shi, T.; Shi,
987 W.; Sung, J.; Thierry-Mieg, D.; Thierry-Mieg, J.; Thodima, V.; Trygg, J.;
988 Vishnuvajjala, L.; Wang, S. J.; Wu, J.; Wu, Y.; Xie, Q.; Yousef, W. A.;
989 Zhang, L.; Zhang, X.; Zhong, S.; Zhou, Y.; Zhu, S.; Arasappan, D.; Bao,
990 W.; Lucas, A. B.; Berthold, F.; Brennan, R. J.; Buness, A.; Catalano, J. G.;
991 Chang, C.; Chen, R.; Cheng, Y.; Cui, J.; Czika, W.; Demichelis, F.;
992 Deng, X.; Dosymbekov, D.; Eils, R.; Feng, Y.; Fostel, J.; Fulmer-
993 Smentek, S.; Fuscoe, J. C.; Gatto, L.; Ge, W.; Goldstein, D. R.; Guo, L.;
994 Halbert, D. N.; Han, J.; Harris, S. C.; Hatzis, C.; Herman, D.; Huang, J.;
995 Jensen, R. V.; Jiang, R.; Johnson, C. D.; Jurman, G.; Kahlert, Y.; Khuder,
996 S. A.; Kohl, M.; Li, J.; Li, L.; Li, M.; Li, Q.-Z.; Li, S.; Li, Z.; Liu, J.; Liu, Y.;
997 Liu, Z.; Meng, L.; Madera, M.; Martinez-Murillo, F.; Medina, I.;
998 Meehan, J.; Miclaus, K.; Moffitt, R. A.; Montaner, D.; Mukherjee, P.;
999 Mulligan, G. J.; Neville, P.; Nikolskaya, T.; Ning, B.; Page, G. P.; Parker,
1000 J.; Parry, R. M.; Peng, X.; Peterson, R. L.; Phan, J. H.; Quanz, B.; Ren,
1001 Y.; Riccadonna, S.; Roter, A. H.; Samuelson, F. W.; Schumacher, M. M.;
1002 Shambaugh, J. D.; Shi, Q.; Shippy, R.; Si, S.; Smalter, A.; Sotiriou, C.;
1003 Soukup, M.; Staedtler, F.; Steiner, G.; Stokes, T. H.; Sun, Q.; Tan, P.-Y.;
1004 Tang, R.; Tezak, Z.; Thorn, B.; Tsyganova, M.; Turpaz, Y.; Vega, S. C.;
1005 Visintainer, R.; von Frese, J.; Wang, C.; Wang, E.; Wang, J.; Wang, W.;
1006 Westermann, F.; Willey, J. C.; Woods, M.; Wu, S.; Xiao, N.; Xu, J.; Xu,
1007 L.; Yang, L.; Zeng, X.; Zhang, J.; Zhang, L.; Zhang, M.; Zhao, C.; Puri,
1008 R. K.; Scherf, U.; Tong, W.; Wolfinger, R. D.; MAQC Consortium. The
1009 MicroArray Quality Control (MAQC)-II Study of Common Practices
1010 for the Development and Validation of Microarray-Based Predictive
1011 Models. *Nat. Biotechnol.* **2010**, *28* (8), 827−838.
1012 (68) Lavigne, C.; dos Passos Gomes, G.; Pollice, R.; Aspuru-Guzik, A.
1013 Automatic Discovery of Chemical Reactions Using Imposed Activation.
1014 *ChemRxiv*, September 29, 2020, ver.1. DOI: 10.26434/chem-
1015 rxiv.13008500.v1.
1016 (69) Friederich, P.; dos Passos Gomes, G.; Bin, R. D.; Aspuru-Guzik,
1017 A.; Balcells, D. Machine Learning Dihydrogen Activation in the
1018 Chemical Space Surrounding Vaska's Complex. *Chem. Sci.* **2020**, *11*
1019 (18), 4584−4601.
1020 (70) Friederich, P.; Krenn, M.; Tamblyn, I.; Aspuru-Guzik, A.
1021 Scientific Intuition Inspired by Machine Learning Generated
1022 Hypotheses. *arXiv (Machine Learning)*, December 14, 2020,
1023 2010.14236, ver. 2.
1024 (71) Aspuru-Guzik, A.; Lindh, R.; Reiher, M. The Matter Simulation
1025 (R)Evolution. *ACS Cent. Sci.* **2018**, *4* (2), 144−152.
1026 (72) Steiner, S.; Wolf, J.; Glatzel, S.; Andreou, A.; Granda, J. M.;
1027 Keenan, G.; Hinkley, T.; Aragon-Camarasa, G.; Kitson, P. J.; Angelone,
1028 D.; Cronin, L. Organic Synthesis in a Modular Robotic System Driven
1029 by a Chemical Programming Language. *Science* **2019**, *363* (6423),
1030 eaav2211.
1031 (73) Coley, C. W.; Thomas, D. A.; Lummiss, J. A. M.; Jaworski, J. N.;
1032 Breen, C. P.; Schultz, V.; Hart, T.; Fishman, J. S.; Rogers, L.; Gao, H.;
1033 Hicklin, R. W.; Plehiers, P. P.; Byington, J.; Piotti, J. S.; Green, W. H.;
1034 Hart, A. J.; Jamison, T. F.; Jensen, K. F. A Robotic Platform for Flow
1035 Synthesis of Organic Compounds Informed by AI Planning. *Science*
1036 **2019**, *365* (6453), eaax1566.
1037 (74) Burger, B.; Maffettone, P. M.; Gusev, V. V.; Aitchison, C. M.; Bai,
1038 Y.; Wang, X.; Li, X.; Alston, B. M.; Li, B.; Clowes, R.; Rankin, N.; Harris,
1039 B.; Sprick, R. S.; Cooper, A. I. A Mobile Robotic Chemist. *Nature* **2020**,
1040 *583* (7815), 237−241.

https://dx.doi.org/10.1021/acs.accounts.0c00785
Acc. Chem. Res. XXXX, XXX, XXX−XXX