CHEMICAL REVIEWS

Review

Cite This: *Chem. Rev.* 2019, 119, 6595−6612

pubs.acs.org/CR

Search for Catalysts by Inverse Design: Artificial Intelligence, Mountain Climbers, and Alchemists

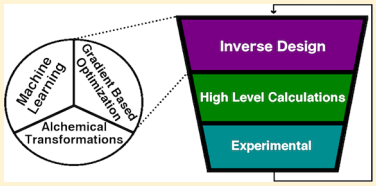
Jessica G. Freeze,[†,‡] H. Ray Kelly,[†,‡] and Victor S. Batista*[,§,‡]

[†]Department of Chemistry, Yale University, New Haven, Connecticut 06520, United States

[‡]Energy Sciences Institute, Yale University, West Haven, Connecticut 06516, United States

[§]Department of Chemistry, Yale University, P.O. Box 208107, New Haven, Connecticut 06520, United States

ABSTRACT: *In silico* catalyst design is a grand challenge of chemistry. Traditional computational approaches have been limited by the need to compute properties for an intractably large number of possible catalysts. Recently, inverse design methods have emerged, starting from a desired property and optimizing a corresponding chemical structure. Techniques used for exploring chemical space include gradient-based optimization, alchemical transformations, and machine learning. Though the application of these methods to catalysis is in its early stages, further development will allow for robust computational catalyst design. This review provides an overview of the evolution of inverse design approaches and their relevance to catalysis. The strengths and limitations of existing techniques are highlighted, and suggestions for future research are provided.



CONTENTS

1. INTRODUCTION

The chemical space of all possible compounds is almost inconceivably vast, with just the number of small organic molecules estimated to be $>10^{60}$.[1] Clearly, it is impossible to synthesize and characterize all possible compounds. So, there is great interest in the development of efficient methods to search for molecular compounds and materials with desirable properties without having to test all possible structures. In particular, the search for efficient catalysts is central to a wide range of chemical processes, with 80% of all manufacturing requiring catalysis at one or more steps of the production mechanisms.[2] The importance of catalysis has led to the selection of *in silico* catalyst design as a holy grail in chemistry.[3] Here, we focus on the emergence of computational methods based on inverse design for the search of molecular and heterogeneous catalysts.

Traditionally, the search for molecular compounds and materials with desired properties has been based on the so-called *direct method* where a library of promising candidates is generated and then experimentally screened to identify compounds with suitable properties.[4−7] However, the number of possible molecules in the library grows exponentially with the number of sites that could be modified, so the cost and time needed for synthesis and testing can be massive. Computational methods can reduce the experimental burden by narrowing the range of possibilities. One approach involves virtual screening by implementing the direct approach *in silico* to narrow the range of promising candidates, as already successfully applied for the search of catalysts for methanation and hydrogen evolution.[8−10] Virtual screening can assess a large number of possibilities rather quickly since it can be trivially parallelized for distributed computing. However, it has also been limited by the exponential scaling and thus is

Special Issue: Computational Design of Catalysts from Molecules to Materials

Received: December 9, 2018
Published: May 6, 2019

DOI: 10.1021/acs.chemrev.8b00759
Chem. Rev. 2019, 119, 6595−6612

ineffective as the dimensionality of the chemical space increases.

The *inverse design* strategy is the reverse of the direct method since it starts with a desired target property and tailors a structure with that property. Often, an initial reference structure is gradually changed by following the gradients of the expectation value of the property with respect to the parameters that define the chemical identity.[11−13] The exponential scaling of computational cost is avoided since the expectation values (and gradients) are computed in polynomial time.

Early applications of inverse methods included design of solid-state materials with a desired band structure.[14−16] Subsequent implementations have successfully optimized the visible-light absorption properties of dye sensitizers for photocatalysis and dye-sensitized solar cells.[17] Other applications were focused on the hyperpolarizability of large aromatic structures,[18] specific material hardness,[19] volumetric properties in metal−organic frameworks,[20] binding in metal clusters,[21] nitrogen-fixating catalysts,[22] and binding energies of adsorbates to catalytic metallic nanoparticles and slabs.[23,24] Nevertheless, inverse design methods remain in the limited realm of a few groups involved in method development and have yet to gain widespread application to the design of molecular and heterogeneous catalysts.

Here, we review algorithms that have been developed for inverse design of molecules and materials with emphasis on methods relevant to catalysts and/or catalytic properties. We focus on three major techniques, including gradient-based methods such as the linear combination of atomic potentials,[17,18,25−33] alchemical transformations,[23,24,34−38] and machine learning techniques.[39−47] We also offer our perspective on future directions for inverse design of catalysts.

2. GRADIENT-BASED OPTIMIZATION

2.1. Mapping Inverse Design to Optimization

The inverse design problem can be mapped into a gradient-based optimization when the property of interest is a smooth function of the parameters that define the chemical identity. Similarly to a traveler climbing to the top of a mountain range (A, Figure 1), an initial reference structure B can be changed toward the top by moving along the negative gradients of the property to be optimized with respect to the parameters that define the chemical composition. Fortunately, the optimization of chemical properties tends to be more efficient than one might expect from the vastness of chemical space. Property optimization has been formulated as an optimal control problem with a fitness landscape that relates an objective (e.g., catalytic activity) to a set of input parameters.[48,49] These landscapes were mathematically shown to contain no suboptimal traps given a few physical assumptions, though such traps can be created by constraints imposed by the input variables.[50] Fitness landscapes were generated for over 100 sets of experimental data, and the vast majority was shown to be free of traps. In particular, activity landscapes were produced for over 30 different types of solid-state catalysts with varying elemental composition. Traps, which only appeared in four of the catalyst sets, were said to arise from an insufficient number of variables to find the optimal solution (i.e., constrained elemental composition and/or experimental conditions).[50] The surprising efficiency of chemical property optimization enables the successful use of gradient-driven methods for inverse design.

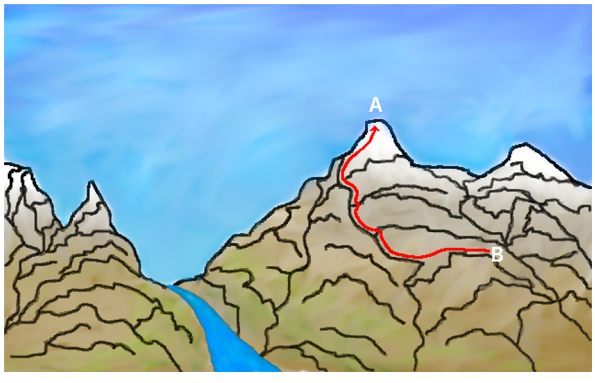


Figure 1. Mountain range representing a property of interest in chemical space, optimized from B to A by following the negative gradients as in the GdMC method.

The *gradient-driven molecule construction* (GdMC) method is a representative example of inverse design for construction of a molecule with a desired property based on gradient optimization methods.[22] The GdMC method requires one to start with a predefined molecular fragment and builds a molecule by gradient optimization while keeping the initial fragment fixed. The molecular scaffold (i.e., everything outside of the constrained fragment) is initially represented by a potential, referred to as a "jacket potential", which is optimized by minimizing the geometry gradient. This potential can consist of anything from a group of point charges to a full quantum model, depending on the requirements of the optimization problem. Following the optimization process, the potential is converted to a molecular structure. A model system using an abstract grid-based potential was explored, but representation of the potential as a molecular structure was nearly impossible. GdMC was applied to the construction of a molybdenum complex for nitrogen fixation, based on the well-known Schrock[51] complex. Starting from a fixed Mo−N_2 structural fragment, a jacket potential was used to represent the ligand environment and its interaction with the Mo−N_2 fragment. The potential was optimized to reduce the geometry gradient and subsequently converted to a molecular ligand environment. Several representations of the ligands were considered. Single point charges, multiple point charges, and nuclei and electrons were used as models. Additionally, ligands were represented directly in terms of an extra potential in the Kohn−Sham equations,[52] which added the challenge of ensuring that the potential corresponded to an actual ligand environment. It was suggested that the linear combination of atomic potentials (LCAP)[32] method described in section 2.2 could be used to address this issue.[22] Further development of the GdMC method could lead to an effective method to aid in the design of catalysts from a known reference fragment such as Mo−N_2 that can be used to constrain the catalytic binding site, while the rest of the complex is optimized.

The GdMC method has the potential to impact the field of inverse design. However, it shares the same problems that plague most gradient descent techniques. Namely, finding a desired global optimum can be challenging or computationally expensive.[22] One reason for this hindrance is that, like mountain ranges, the chemical space is often a nonconvex surface leading to the identification of only local optima. This is easily seen in Figure 1 as the optimization would have a

DOI: 10.1021/acs.chemrev.8b00759
Chem. Rev. 2019, 119, 6595−6612

much easier time finding a local cusp rather than the global optimum. Another problem that affects gradient-based solutions is the issue of sampling discrete spaces, as can be seen by the pillars in Figure 2. The difficulty arises from determining which points will have data that can be sampled to determine the gradient. From Figure 2, it makes as much sense to check point (0.234,5.6) as it does (0,1). Even if restricted to integers, there is no guarantee that every integer position in space has data. Thus, solutions to the discreteness problem had to be devised.

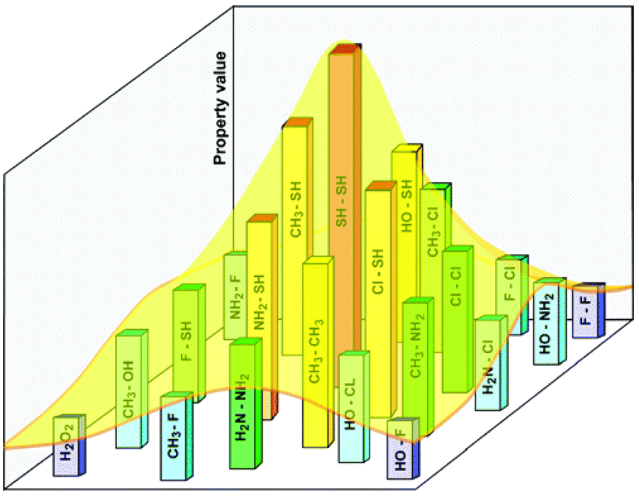


Figure 2. Linear combination of atomic potential algorithm produces a continuous surface from discrete possible structures with varying polarizabilities. Reproduced with permission from ref 32. Copyright 2006 American Chemical Society.

2.2. Linear Combination of Atomic Potentials

The LCAP method, first implemented by David N. Beratan and Weitao Yang,[32] is inspired by Kirkwood's coupling parameter method[53−56] in thermodynamic integration (and other alchemical transformation methods discussed in section 3.1). Continuous alchemical interpolation transforms the analysis of a discrete space of possible structures into the analysis of a continuous function for the electronic−nuclear potential, defined as a linear combination of atomic potentials (eq 1):

$$v(r) = \sum_{\mathrm{R,A}} b_{\mathrm{A}}^{\mathrm{R}} v_{\mathrm{A}}^{\mathrm{R}}(r) \tag{1}$$

$v(r)$: Linear combination of atomic potentials. $b_{\mathrm{A}}^{\mathrm{R}}$: Atom A exists at location R with $b_{\mathrm{A}}^{\mathrm{R}}$ probability between 0 and 1, where

$$\sum_{\mathrm{A}} b_{\mathrm{A}}^{\mathrm{R}} = 1$$

$$v_{\mathrm{A}}^{\mathrm{R}}(r) = \begin{cases} \text{Atomic potential of substituent A at position R} \\ \text{A may be an atom or a substituentmade up of atoms B} \sum_{\mathrm{B}} v_{\mathrm{B}}(r) \end{cases}$$

The expectation value of the property of interest is optimized relative to the coefficients $b_{\mathrm{A}}^{\mathrm{R}}$. Upon convergence, the largest coefficient b_{A} is rounded for each site R to define the favored substituent for that site. Figure 2 shows continuous structure generated by LCAP.[32] Through the development of LCAP, Wang et al. were able to optimize the composition of a molecule separately from geometry, thereby finding a structure with optimal polarizability and hyperpolarizability more quickly than enumeration with standard DFT calculations.[32]

Getting stuck at a local optimum is a risk common to most gradient-based optimization methods, including the LCAP optimization in chemical space.[27] In addition, optimization in the continuous chemical space can lead to a local optimum of a nonphysical alchemical species, instead of converging into a real molecule corresponding to the global optimum. To address these challenges, a gradient-directed jump method was introduced.[27] Starting from an initial real molecule, LCAP is applied only to find the gradients necessary to "jump" to a nearby real molecule along the direction of the gradients. In the analogous mountain range picture, one can imagine the jump corresponding to a traveler hopping to a nearby hill along the direction of the negative gradients. The resulting gradient-directed jump method has already been used to develop a molecule with a lower LUMO energy for the purpose of discovering new n-type semiconductors.[57] While semiconductors and catalysts may sometimes be thought of as separate fields, the interfacing between the two often becomes quite pertinent, particularly in the development of solar fuels and water splitting.[57−63]

The gradient-directed Monte Carlo (MC) optimization method is another approach to address the challenge of local minima in nonconvex surfaces.[64] Escape from a minimum is induced through the implementation of MC with Metropolis conditions for randomly accepting a move. While this method was initially applied to porphyrins in the context of nonlinear optics,[64] porphyrins have also been extensively investigated as catalysts in a myriad of applications.[65−67] Gradient-directed MC was later combined with a best-first search algorithm to optimize the functionalization of diamondoids by minimizing/maximizing the HOMO−LUMO gap of materials for potential optoelectronic applications.[68] An alternative approach to avoid alchemical local minima has been proposed by Keinan, Therian, Beratan, and Yang.[30] The scheme applied a finite difference method to update the substituents at alterable positions on the compound based on calculations of the hyperpolarizability. From a library of 940 800 possible molecules, ten runs of the optimization algorithm found six unique porphyrin structures for nonlinear optical chromophores. Remarkably, the optimization found a new class of T-shaped structures,[30] demonstrating the power of LCAP to expand diversity in a group of chemicals that offer new synthetic opportunities.

LCAP was further improved to drastically reduce the "cost" of computations by reducing the number of steps necessary for design *in silico*. The combination of tight binding and LCAP was able to explore a space of 10^4 π-conjugated structures with only 40 property calculations. The tight binding method was then implemented into LCAP, utilizing an independent particle Hückel−Hamiltonian matrix to optimize the matrix coefficients which in turn optimize the electronic structure.[18] A tutorial for that implementation has been developed.[33]

Xiao et al. extended the LCAP method to design photoabsorbers for dye-sensitized solar cells,[11,17] using the phenyl-acetylacetonate anchor as a starting point for linking chromophores to TiO_2. From 144 possible molecules, the 3-acac-pyran-2-one anchor was discovered using the tight-binding (TB) LCAP Hamiltonian and subsequently synthesized and tested. The predicted properties of improved photoabsorption and electron transfer, when compared to the initial reference adsorbate, have been experimentally

DOI: 10.1021/acs.chemrev.8b00759
Chem. Rev. 2019, 119, 6595−6612

confirmed.[11,17] It is therefore expected that these methods should be particularly useful in the design of other semiconductor surfaces that might be doped or functionalized with molecular adsorbates as in dye-sensitized solar cells and photocatalytic applications. Recently, we extended the TB-LCAP method to the inverse design of molecular catalysts.[69] The ligand composition of Ni(II) transition metal catalysts for CO/CO_2 conversion[70,71] has been optimized to reduce the activation energy of the rate-limiting step of the underlying catalytic reaction mechanism. Optimization of this energy difference using the TB-LCAP method resulted in an improved catalyst with a DFT-estimated 2 orders of magnitude improvement in turnover frequency.[69] This methodology could be extended to cover multiple descriptors of catalytic reactivity.

In summary, the LCAP method has already been shown to address the challenges faced by inverse design based on gradient optimization methods. The LCAP optimization is able to traverse the space of possible chemical structures to find the ones with desired properties. Advances using gradient-directed jumps and tight binding model Hamiltonians led to increased robustness toward nonconvex surfaces and greater screening efficiency. With uses in semiconductors, porphyrins, and chromophores, LCAP stands to have great promise in the development of catalysts.

3. ALCHEMICAL DERIVATIVES

3.1. Alchemical Transformations

Alchemical transformations rely on the concept of state functions describing certain properties of systems at equilibrium. Changes in these properties are independent of the path taken between the initial and final states, enabling the selection of arbitrary paths between any pair of states regardless of whether these transformations can be observed in the real world. Transitions occurring through experimentally inaccessible paths are known as *alchemical* transformations. Typically, alchemical transformations are described by a model Hamiltonian defined as a linear interpolation of Hamiltonians for the initial and final states, weighted by the coupling parameter λ, as follows

$$H(\lambda) = (1 - \lambda)H_A + \lambda H_B \tag{2}$$

where λ is 0 at state A and 1 at state B. As λ is not required to be an integer, values of λ between 0 and 1 may suggest a nonphysical mixture of states. Alchemical methods have been most frequently used for calculations of free energy changes, as implemented in the thermodynamic integration method proposed by Kirkwood in 1935.[53−56] Practical aspects of various computational methods for determining free energy changes have previously been reviewed.[72,73] A key focus of this review is a discussion of recent research efforts focused on the application of alchemical transformations for inverse design of catalysts and other materials.

3.2. Alchemical Potentials

As mentioned previously, the representation of chemical space as a continuum is crucial to inverse design algorithms. One such representation within grand-canonical ensemble DFT was proposed by von Lilienfeld et al., who devised a variational particle number approach for the inverse design of molecules.[34,35] While conventional DFT methods involve the optimization of electronic structure and nuclear positions, this approach allows for chemical composition to also be varied and optimized. From the Hohenberg−Kohn theorem,[74] it was demonstrated that the external potential v_{ext} uniquely determines the electron density $\rho(r)$ for constant number of electrons N_e, where v_{ext} is also a functional of the nuclear charge distribution $Z(r)$.[34] Therefore, any ground-state observable O can be written as a function of N_e and a functional of $Z(r)$, i.e., $O[Z(r)](N_e)$. This led to the creation of a penalty functional to be minimized for the purpose of property optimization

$$P[Z](N_e) = |O[Z](N_e) - O_{ref}|^2 \tag{3}$$

where O_{ref} is the reference value to which the observable should be optimized.[34] Equation 3 allows for the inverse design of compounds to be approached as a minimization problem in terms of electrons and nuclei which can be accelerated using gradient descent methods. The nuclear chemical potential μ_n was presented as a function of space which corresponds to the tendency of the molecule to undergo changes in the nuclear charge distribution Z. An equation for the first-order approximation of μ_n was introduced as the electrostatic field $E^{(1)}$ at a position r

$$\mu_n(r) \approx E^{(1)}(r) = -\int dr' \frac{\rho(r')}{|r - r'|} + \sum_I \frac{Z_I}{|r - R_I|} \tag{4}$$

where R_I is the position of the nucleus of atom I.[34] At a nuclear position, $r = R_I$, μ_n is referred to as the alchemical potential as it corresponds to the propensity for atom I to mutate into another element. At other positions, μ_n is related to the proton affinity at position r. Using the gradient of the penalty functional outlined in eq 3 with respect to Z, a nonpeptidic anticancer drug candidate was identified.[34] Subsequently, von Lilienfeld and Tuckerman rigorously described chemical space using a molecular grand canonical ensemble DFT framework.[35] This description of chemical space allows for the consideration of alchemical changes in molecular composition within a DFT framework, which is needed for inverse design optimizations. A new expression for the nuclear potential as a modified electrostatic potential was proposed

$$\mu_n(r) = \int dr' \frac{Z(r')\mathrm{erf}[\sigma|r - r'|] - \rho(r')}{|r - r'|} \tag{5}$$

where the error function serves to eliminate the effect of intranuclear proton repulsion with a sufficiently small σ.[35] Various relations were obtained within this framework, including those for the nuclear hardness and molecular Fukui function.[35] It was noted that a coupling parameter λ could be used to move through alchemical paths in chemical space. The effect of these alchemical transformations on interaction energies was later investigated by the same authors by transmuting between neutral ten-electron molecules (CH_4,NH_3, H_2O, and HF) as they interacted with formic acid.[36] The transformations were performed both with a rigid structure and while allowing for relaxations, observing a significant difference in these interaction energies even for a transformation between CH_4 and H_2O.[36] These deviations underscored the need to identify alchemical paths that enable accurate prediction of molecular properties.

3.3. Predictions of Energy Changes Using Alchemical Derivatives

Alchemical derivatives, which correspond to the change in the energy with respect to changes in the nuclear charge distribution, show great promise for estimating energy

DOI: 10.1021/acs.chemrev.8b00759
Chem. Rev. 2019, 119, 6595−6612

differences between compounds. These estimates could be used to guide a gradient-based optimization for inverse design of catalysts or simply to make a large number of energy estimates with only a few calculations. Importantly, an analytic alchemical derivative can be obtained from a single-point calculation. Based on the Hellmann−Feynman theorem,[75] von Lilienfeld proposed a method for accurate energy predictions for isoelectronic compounds.[37] For linearly interpolated transformations between compounds A and B, along the λ parameter, the derivative of the energy E is given by

$$\frac{dE(\lambda)}{d\lambda} = \langle H_B - H_A \rangle_\lambda = \int dr \rho_\lambda(r) \cdot [v_B(r) - v_A(r)] \quad (6)$$

where $v(r)$ corresponds to the Coulomb potential.[37] These derivatives were computed for various transmutations between 10-electron molecules, obtaining good agreement with finite differences. However, the energetic paths were not linear in λ. Ideally, one would like to be able to truncate $\frac{dE(\lambda)}{d\lambda}$ to first order in Taylor series so that accurate predictions for many transmutations could be made from only one DFT calculation. The concept of a linearizing coefficient was introduced to address this issue. The coefficient is calculated from a reference transmutation and used to make more accurate predictions of the energy change. The improvement is illustrated in Figure 3, where inclusion of a linearizing coefficient improved the predictions of HOMO eigenvalues for all cases.[37]

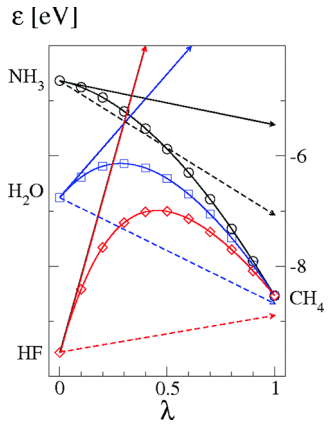


Figure 3. Predicted HOMO eigenvalues for CH_4 from the first-order derivatives of NH_3, H_2O, and HF. The solid lines correspond to predictions made using just the derivatives, while the dotted lines included linearization coefficients obtained by using reference compounds. The reference pairs used were CH_3-CH_3 with each of CH_3-NH_3, CH_3-OH, and CH_3-F. Note the nonlinearity of the HOMO eigenvalue with respect to = λ for each of these transformations. Reproduced with permission from ref 37. Copyright 2009 AIP Publishing.

Alchemical derivatives were later used to estimate energy barriers for simple reactions.[23] The energy derivative for isoelectronic transformations with relaxed geometries can be expressed to first-order precision, as follows

$$\frac{\partial E}{\partial \lambda} = \int dr \mu_n(r) \frac{\partial Z(r)}{\partial \lambda} \quad (7)$$

where the nuclear chemical potential is defined by eq 5.[23] It is important to note that the requirement for isoelectronic transformations is less restrictive when pseudopotentials are used, as a constant number of valence electrons is satisfactory. Two model systems were analyzed showing agreement between alchemical analytical and finite difference derivatives, including the umbrella flipping of ammonia and the protonation of small molecules.[23] Although there was generally good agreement, some errors resulted from the nonlinearity of the properties along the λ path. This is shown for the activation energy of the umbrella flipping of NH_3 in Figure 4 where the

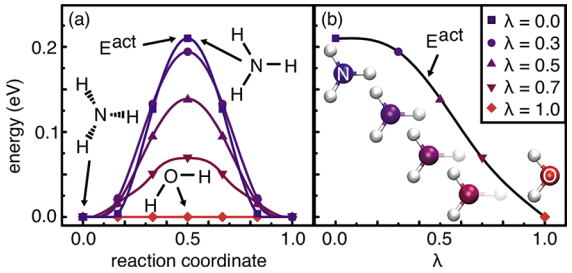


Figure 4. (a) Energy along the reaction coordinate for the umbrella flipping of NH_3 shown for values of λ as the molecule is transformed into H_2O. (b) Activation energy, E^{act}, of umbrella flipping for different values of λ. It can be seen that E^{act} is not linear in λ and that the derivative at H_2O will better predict the energy at NH_3 than vice versa. Reproduced with permission from ref 23. Copyright 2010 AIP Publishing.

derivative at H_2O would give a good estimate of the barrier for ammonia, but the derivative at NH_3 would strongly overestimate the barrier for water. This underscores the need for finding paths in λ such that the first-order derivative remains valid.

Alchemical derivatives have also been used to predict the behavior of catalytic materials, including Pd nanoparticle catalysts for oxygen reduction.[23] As the activity of oxygen reduction catalysts has been correlated with the binding energy of O,[76] this property can be optimized to reduce the barrier of a reaction that limits the efficiency of fuel cells with proton exchange membranes.[77] In the Pd case, the alchemical term in eq 7 was isolated by relaxing all geometries and only allowing isoelectronic transformations.[23] For every atom that was mutated from Pd to Ag, another was simultaneously mutated to Rh to maintain N_e. The predicted changes in binding energy from alchemical derivatives was compared to those calculated by DFT (Figure 5), demonstrating the ability to obtain sensible binding energy estimates from only the three DFT calculations needed for a routine binding energy calculation of oxygen on a Pd nanoparticle.[23] The results demonstrate that once the accuracy has been confirmed for predictions based on alchemical derivatives rapid identification of improved catalysts can be accomplished with only a few DFT calculations. In an inverse design scheme, catalysts that showed binding energies near the optimum value could be selected for further computational and eventually experimental study.

More recently, first-order alchemical derivatives were used in a similar fashion to make predictions of binding energies for oxygen reduction on periodic metal surfaces.[24] The binding energies of O*, OH*, and OOH* adsorbates were computed on alchemically perturbed Pt(111), Pd(111), and Ni(111) slabs to assess the catalytic activity of these alloys. The slabs consisted of four layers of four atoms, and all alchemical transformations were isoelectronic and resulted in no change to the total atomic number of the slab. The computation of the alchemical derivatives was straightforward. First, a standard

DOI: 10.1021/acs.chemrev.8b00759
Chem. Rev. 2019, 119, 6595−6612

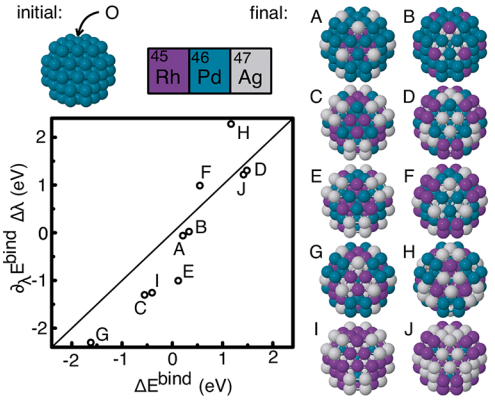


Figure 5. Binding energies of oxygen from the first-order alchemical derivative, $\partial_\lambda E^{bind}\Delta\lambda$, where $\Delta\lambda = 1$, compared to DFT-calculated energy differences, ΔE^{bind}, for isoelectronic variations of a 79 atom Pd nanoparticle. Reproduced with permission from ref 23. Copyright 2010 AIP Publishing.

DFT binding energy calculation was made requiring three optimizations. For each atom I, the difference in the nuclear potential between the two states was computed and placed in an array $\Delta\mu_{nI}$. Alchemical predictions could then be computed for each transformation by taking the dot product of an array of the changes in nuclear charges with $\Delta\mu_{nI}$, an operation with negligible computational cost. Surface alloys for the three metals were considered such that transmutations occurred only in the top and bottom layers, where the atoms in the bottom layer have minimal impact on the binding energy but maintain N_e. Additionally, mutations of skin alloys of Ni with the form M_3Ni were considered where M was either Pt or Pd, and half of the second layer atoms were Ni. Figure 6[24] shows the predicted energies for 360 alloys of the two forms. Those predictions required only 15 DFT calculations, while the direct DFT binding energies needed 720 calculations. In general, errors were less than 0.1 eV, with major exceptions when the Ni in the skin alloys was mutated to Mn, Fe, or Zn. With further study to identify and overcome the cases where the approximation breaks down, first-order alchemical derivatives could be used to rapidly identify improved catalytic materials at low computational cost.

Alchemical derivatives have been applied to problems beyond catalysis, including covalent bonding,[78] BN-doped carbon materials,[79,80] semiconductor band structures,[81] binding in metal clusters,[21,82−84] and bulk material properties.[85,86] Terms related to electrostatic, polarization, and electron transfer effects were applied to Al_{13} clusters doped with four or fewer Si atoms to test the limits of alchemical derivatives.[87] The doped Al cluster was challenging because none of these effects were dominant in determining the relative energies. Agreement with *ab initio* calculations was obtained for two-atom isoelectronic mutations and one-atom nonisoelectronic mutations. Higher-order alchemical derivatives could be used to make more accurate energy predictions although at a higher computational cost.[78,80,88,89] If the calculations of these derivatives become too costly, the advantage over brute-force DFT calculations is less apparent. Thus, there has been an effort to linearize properties, such as energy, with respect to the coupling parameter λ for isoelectronic transformations.[38] The linearization enables the use of first-order alchemical derivatives to compute exact energies, though no general method for linearizing energy in λ exists without relying on specific system information.[90] However, as shown previously, accurate results can be obtained using the first-order derivative for certain systems. Improvements in numerical methods for determining paths in λ for which the first-order perturbation is valid, combined with guidelines when the approximation fails, could lead to the rapid identification of improved catalysts at minimal computational cost.

4. MACHINE LEARNING

Machine learning (ML) as an artificial intelligence approach has been around since 1959.[91] Over the past few decades, there has been a flurry of activity in the field with the development of algorithms and software packages for efficient parametrization of artificial neural network (ANN) and powerful classification methods.[92−94] Great advances in image recognition,[95,96] language processing,[97−99] and optimal path finding[100] have led to many useful new technologies. In chemistry, ML has

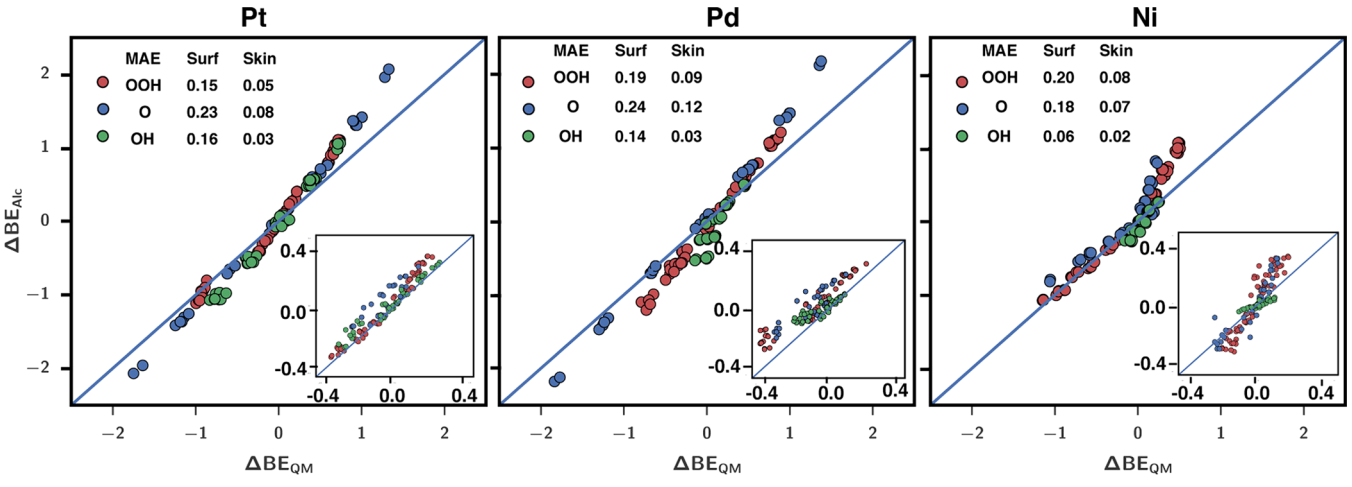


Figure 6. Comparison of alchemically predicted, ΔBE_{Alc}, and DFT calculated, ΔBE_{QM}, OH*, OOH*, and O* binding energies on alloys of Pt(111) (left), Pd(111) (middle), and Ni(111) (right) slabs with transmutations such that $|\Delta Z| = 1$. The data in the figure correspond to surface alloys, while that in the inset is for skin alloys. Mean absolute errors (MAEs) are displayed in eV. Reproduced with permission from ref 24. Copyright 2017 American Chemical Society.

DOI: 10.1021/acs.chemrev.8b00759
Chem. Rev. 2019, 119, 6595−6612

already been explored in a wide range of applications, including force field development,[101,102] DFT calculations,[103] drug discovery,[104] new materials development,[105] toxicity probing,[106] prediction of chemical reactions,[107] and reaction yields,[108] to mention a few. With the overall interest in ML having increased at least 10-fold over the past decade,[109] we anticipate ML will become a valuable computational tool for the discovery and selection of catalytic materials.[110−112] Therefore, we provide a brief overview of a few ideas in ML that could be combined with inverse design methods. We note that the field of ML is vast, allowing for many other ways of combining inverse design with the ANN and classification methods beyond the scope of our review.[47]

We also emphasize that ML has been mostly limited to data analysis methods that compute the structure of correlations in data sets and make predictions based on the extracted correlations. Unfortunately, correlation does not necessarily imply causation.[113] Thus, it remains an outstanding challenge to gain a fundamental understanding of systems from a causal inference of the observational data produced by ML models parametrized with empirical data.[114]

As a pretext, it is perhaps useful to mention what machine learning is capable of doing at its current stage. Largely, machine learning is capable of solving classification and regression problems. This means that if one wants to use machine learning for chemistry they must first phrase their problem either as separating data into classes based on differences in the values of descriptors used to describe that data (classification problem) or looking for a relationship between an input set of features and an output (regression problem). As a trivial example, a classification problem may be inputting free energies and separating reactions into spontaneous and nonspontaneous. A regression problem may be relating ligand-withdrawing capability to reaction rate. Ensuring that the problem fits the capabilities of machine learning is the first hurdle to applying these methods.

One of the most common limiting factors in applications of ML to chemistry is the availability of enough data for reliable parametrization of ANN or classification models. Databases such as ChemDB,[115] ChemSpider,[116] The PubChem Project,[117] and The National Chemical Database Service hosted by the Royal Society of Chemistry provide valuable repositories of structures and data. Expanding past these databases often requires a great deal of computer science skill, though some have attempted to make this process easier. One such example was the development of Algorithm for Chemical Space Exploration with Stochastic Search (ACSESS).[118−120] ACSESS allows for the systematic identification of missing components of already explored chemical space and the expansion into unknown regions to generate new libraries. Extensions to the algorithm have added preference toward the exploration of diverse molecules with desired properties.[119] Though this method has only yet been used to explore small organic molecules, it presents a promising start for those wishing to explore the frontiers of chemical space. Another way to generate data is through the use of a genetic algorithms (GAs).

Genetic algorithms provide powerful methods for generating and assessing structures over which ML models can be trained and tested. GAs are evolutionary methods that mimic the process of biological evolution as exhibited by successive generations that adapt in response to changes in environmental conditions. Analogously, new molecules can be formed based on those features from previous generations found to be correlated with favorable performance.

4.1. Genetic Algorithms

GAs applied to molecular design seek to evolve molecules and improve their properties as determined by changes in structural and functional properties, often encrypted in binary form.[121,122] For example, the presence (or absence) of a specific functional group could be represented by 1 (or 0) of a variable artificial "gene". The list of activated/deactivated genes constitutes the molecular "chromosome" typically initialized with random values, as shown in Figure 7 for a binary depiction of the chromosome of a representative system.[123]

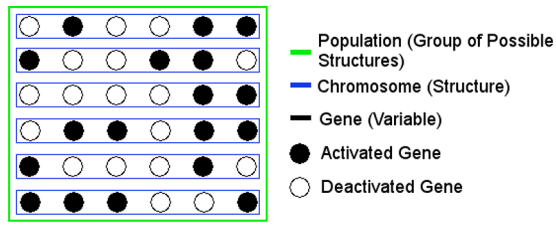


Figure 7. Binary representation of activated genes (solid) in the chromosome of features of a molecular system that affect the property of interest.

Once the chromosome is initialized, the gene variables are evolved according to the following steps:

1. *Fitness Testing*: The chromosome "fitness" is assessed by evaluating the property of interest, usually defined as a simple function of the gene variables.
2. *Parents Selection*: The selection of "parents" corresponding to the current chromosome can be performed in multiple ways, though the most popular approach is to select a chromosome section for replacement at random, as with a roulette wheel (Figure 8).[123] Other methods of selection are highlighted in ref 124 with the number of parents historically set to two although there could be more.[125,126]
3. *Crossover of Parent Genes*: Crossover is the process of combining the parent genes to form the chromosomes of children. One common method involves selecting one or

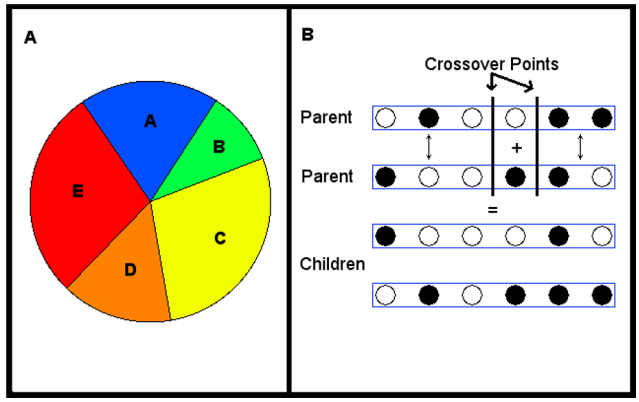


Figure 8. Genetic Algorithm Steps. Panel A shows a circle broken into segments based on the fitness of each parent A through E. Bigger segments mean higher fitness. Panel B shows multipoint crossover where parent's genes are swapped in odd segments and kept in even segments to make children.

DOI: 10.1021/acs.chemrev.8b00759
Chem. Rev. 2019, 119, 6595−6612

more crossover points and swapping the genes on one side of the point but leaving them unchanged on the other side, as depicted in Figure 8A.[123,127] Another option is to translate those genes that match between parents, thereby enforcing more commonly found genes under the assumption that better genes will appear more frequently. The genes that do not match between parents are randomly selected from the possible gene expressions, with the more common methods discussed in ref 127. Crossover occurs according to random chance, and in the event that it does not occur, the parent chromosome may be included in the next generation.

4. *Mutation of Child Genes*: A small mutation chance is permitted for each gene to sample genes beyond those of the parents.
5. Repeat Steps 1−4 Until Convergence of Fitness.

4.2. Genetic Algorithms in Inverse Design

The combination of inverse design and GAs has been employed for the discovery of materials with hardness >40 GPa,[19] for exploring the full range of band gaps for AlGaAs alloys,[128] and for deriving understanding of structural motifs that led to desirable band structures in GaP alloys with nitrogen impurities.[15] These studies are relevant to catalytic applications since the band gap energy has been identified as a descriptor of catalytic activity in various semiconductor materials, including mixed metal oxides.[129]

Froemming and Henkelman combined GA with DFT to optimize core−shell metal nanoparticles for the catalysis of oxygen reduction.[130] Previously, Hammer and Nørskov demonstrated the correlation of the center of the d-band with catalytic activity.[131,132] Therefore, the chosen fitness function was the difference between the center of the d-band of nanoparticles and that of Pt(111), a known oxygen reduction catalyst.[130] GA was applied to the initial random generation of 30 core−shell nanoparticles, with the constraint that each d-block metal was used once as a core and once as a shell. DFT optimizations and calculations of the electronic density of states were performed to estimate the fitness according to the difference between the d-band center and that of Pt(111). The particles were then ranked according to their fitness and underwent a breeding process to produce the next generation of particles as depicted in Figure 9.[130] Breeding probabilities were chosen such that the particle with the highest ranking was 10 times more likely to breed than the lowest ranked one.

Random mutations of the parent particles were allowed with 10% probability to ensure ample sampling of the chemical space. Single-point, double-point, and inversion mutations occurred with equal probability (3.3%). The GA was compared to Metropolis Monte Carlo and brute-force procedures, with improved efficiency as shown in Figure 10. While the 38- to 79-atom particles considered in this study might not be readily synthesizable, this GA method can provide insight into the types of metals that might be combined to create improved catalysts. We anticipate that further refinement of the fitness functions to better match catalytic activity could result in the applicability of this type of sampling to more complex catalytic systems.

As detailed by Sokalski,[133,134] an electrostatic field can be used as an abstract representation of a catalytic environment. This approximation is useful when electrostatic effects are dominant in determining activity. Dittner and Hartke used

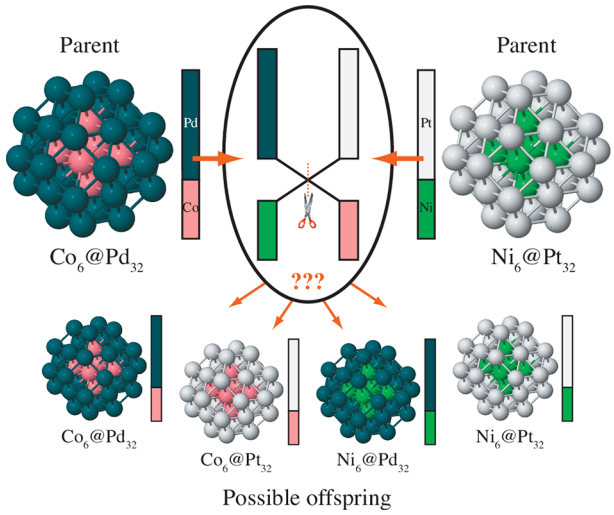


Figure 9. Depiction of the breeding process used in the GA of ref 130 for two example parents $Co_6@Pd_{32}$ and $Ni_6@Pt_{32}$. The single offspring of each breeding event was chosen randomly from the possible offspring. Reproduced with permission from ref 130. Copyright 2009 AIP Publishing.

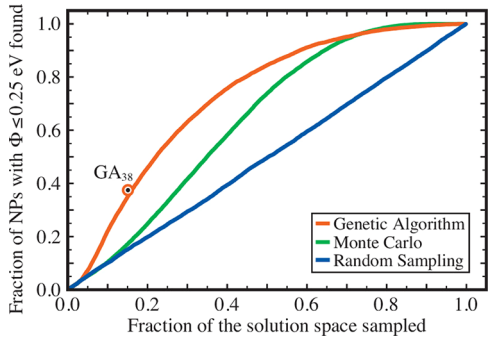


Figure 10. Comparison of GA with Monte Carlo and random sampling for optimization of core−shell nanoparticles. The fraction of nanoparticles found with energies within 0.25 eV of the optimal d-band is plotted against the total fraction sampled. At the point marked, the GA had found more than twice as many particles with the desired fitness as the other sampling methods (37% of fit particles after sampling 15% of space). Reproduced with permission from ref 130. Copyright 2009 AIP Publishing.

global optimization methods to allow for the inverse design of these fields.[135] Starting with a preoptimized reaction path, an optimal electrostatic model of a catalyst is identified. These models, which are produced by GA, can provide insight into important catalytic effects and guide the design of new catalysts. However, there is not yet a clear way to translate them to a molecular structure. Currently, this method can be used to find the optimal catalytic field (which is constant over the fixed reaction path) for one reaction step. Future work could eliminate some of the approximations used so that this method can allow for a more robust inverse design of catalysts.[135]

GAs allow for the rapid identification of molecules with desired properties. Springborg et al. used an inverse design method based on GA to optimize mixed Si−Ge clusters for solar energy harvesting.[136] A performance function was defined which included several criteria corresponding to the suitability of molecules for solar cell applications. Using computationally efficient density-functional tight binding (DFTB), the proper-

DOI: 10.1021/acs.chemrev.8b00759
Chem. Rev. 2019, 119, 6595−6612

ties of interest were rapidly computed to give a performance score. A GA was used in which those members with the better score were kept for future generations. It was noted that many of the optimal Si−Ge clusters did not follow typical chemical intuition, which necessitates the use of inverse design.[136] This work was expanded upon with the creation of the poor man's materials optimization (PooMa) approach.[137] This method allows for the optimization of some property given a particular molecular backbone structure. Such a scheme could be used for the optimization of heterogeneous catalysts, in which the atoms at particular sites are varied. PooMa consists of a GA for guiding site transformations, a subroutine for molecular construction based on typical structures, a method for electronic structure calculations, and a performance function corresponding to the desired property. This scheme is sufficiently general to allow for the adjustment of the electronic structure method and performance function to fit any application. The DFTB calculations employed by the authors are computationally feasible on a standard laptop or desktop computer.[137] GAs have similarly been used to guide the computational screening of organic polymers for photovoltaic applications.[138,139] After the pool of candidates had been reduced by GA, more rigorous calculations were used to create a computational pipeline for the identification of promising compounds.[139] Recently, a GA was used to identify fluorophores for organic light-emitting diode applications.[140] A subspace of 1.26 million molecules was reduced to 3792 promising candidates after performing only 7518 DFT calculations. Molecules were represented as tree structures in which each node corresponded to a molecular fragment from a predefined library. This methodology could be applied to catalyst design with a predefined library consisting of molecules thought to be synthetically viable. Further development of GA-based methods will enable the rapid identification of catalytic molecules and materials with desired properties.

It is important to note that GAs can be used to guide both experiments and computations. The use of evolutionary algorithms for material discovery, including catalytic materials, has recently been reviewed.[141] These algorithms have been used to direct the synthesis of materials, beginning with an influential paper by Wolf et al. in which a heterogeneous catalyst for the catalysis of propane to propene was optimized according to an experimental fitness function.[142]

Similar to the use of GAs, the best-first search (BFS) can be used as a discrete method of moving through chemical space.[25] This method is best suited for identifying molecules with an optimal property within a moderate sample space. A BFS scheme was used to optimize the acidity of 2-napthol derivatives in both the ground and excited states.[26] Results for test cases were compared to synthetic known values and found to be within 0.15 eV. For a molecule with N sites on which the substituents will be altered, site 1 is first altered with each possible substituent, and the value of the desired property is computed for each variation. The molecule with the optimal substituent is kept, and site 2 is varied such that the property is computed for each substituent. This process is continued through site N and is repeated for a chosen number of iterations. Convergence to a local optimum is mitigated by randomly generating several initial structures.[26] Similarly, BFS was used to design intrinsically stable thiadiazinyl radicals and optimize their electrophilicity and nucleophilicity.[143,144] A gradient-driven Monte Carlo[28,64] component was later added to BFS to avoid getting trapped in local minima.[68] One could imagine using a similar BFS algorithm to optimize a catalytic property. This would be especially effective for a well-defined catalyst structure on which the substituents need to be tuned to improve catalytic activity.

4.3. Machine Learning for Inverse Design?

Having discussed ML and GA, one may wonder whether inverse design could benefit from ML. A capability that ML can bring to inverse design methods is the ability to establish patterns of correlation that might not be otherwise evident from a cursory examination of the data or through traditional analysis using general trends and chemical principles. Such capabilities of ML have been recently demonstrated in the field of game theory through the discovery of new strategies for winning the game of Go that human players had not found in the 2500 years since the game's inception.[145] The ability to determine new patterns may even lead to new chemical understanding of structure−property relationships. Furthermore, the combination of ML with principal component analysis could pick out the features that most drastically affect the property of interest and reduce the vastness of chemical space to a tractable subspace.

Kulik and co-workers, for example, have already utilized GA to generate structures at a specific distance that was defined by metal proximal effects from other structures trained by an ANN. Combining GA with DFT calculations revealed new spin-crossover complexes through reduction of the spin-state splitting energy.[39] A comparison of the DFT-driven GA method versus the ANN found that two-thirds of the structures identified by the ANN method were spin-crossover complexes as verified by DFT. The ANN method was able to boil down more than 5500 possible complexes to 51, yet only took seconds for evaluation of a compound as compared to days by DFT.[39] These studies demonstrated that ANN could identify candidate molecules with the properties of interest with high accuracy and much faster than accomplished with brute force high-throughput methods.[39]

Gómez-Bombarelli et al. used ML to aid the virtual screening of organic light-emitting diode molecules.[146] Using ML, a search space of 1.6 million molecules was reduced to 400 000 for evaluation by time-dependent DFT (TD-DFT). This space was comprised of molecules fitting a maximum mass of 1100 g/mol and fitting the formula donor−$(\text{bridge})_x$−acceptor with x values between 0 and 2. A random set of 40 000 molecules were treated with TD-DFT to train the neural network, and subsequent molecules were selected for TD-DFT simulations based on ANN predictions. The ANN was retrained to incorporate additional TD-DFT calculations as they were performed and was found to produce more accurate predictions than a linear regression model. Hundreds of promising molecules were identified, and a select few were chosen for synthesis. Experimental results showed external quantum efficiencies of up to 22%, demonstrating the power of combining ML with quantum calculations.[146]

Quantitative structure−activity relationship (QSAR) models also establish correlations between molecular structure and molecular properties.[147] The origin of structure−functional models can be traced back at least to the Hammett equation which correlates the reactivity of aromatic molecules with the nature of the substituent groups described by empirical parameters (e.g., reaction rates of benzoic acid derivatives).[148] Modern QSAR methods involve the use of informatics and ML techniques to identify relationships between molecular

DOI: 10.1021/acs.chemrev.8b00759
Chem. Rev. 2019, 119, 6595−6612

structures and properties as implemented in a wide range of applications including medicinal chemistry,[149] polymer materials for solar cells applications,[147] and the assessment of toxicity of pesticide metabolites in food.[150] However, QSAR models rely on the "activity landscape" of the property of interest being relatively smooth so that small changes in structure result in small changes in activity. Unfortunately, many applications exhibit "activity cliffs" where similar molecules have very different properties.[151] Additionally, molecules used in training sets may not adequately describe novel molecules beyond the types of molecules used for parametrization of the models. Some of these drawbacks can be overcome by combining QSAR models with ML for the development of scoring functions.

In fact, Google's DeepMind did just that with the new AlphaFold. AlphaFold inputs a protein sequence into neural networks trained on amino acid pair distances and bonding angles between amino acids.[152] By generating a scoring function from the output of these predictions and performing gradient descent, a three-dimensional protein structure is formed. Though currently far from perfect, AlphaFold beat out all competitors at the 2018 Critical Assessment of Structure Prediction, correctly predicting the structures of 25 of the possible 43 proteins.[152,153] This method, while no doubt reliant on the high computing power of Google's infrastructure, presents hope for solving chemistry's most complex systems with machine learning.

4.4. Machine Learning for Catalyst Inverse Design

Few studies have as of yet been reported as using ML for inverse design of catalysts.[40,41,154] Most of these efforts have been limited to prediction of catalytic activity[41,154] and catalytic reaction pathways.[40] As the range of possible pathways based on available sites, possible structures, and experimental conditions involved in a typical reaction can be enormous, it makes sense to invert design techniques to explore them. For example, the reaction of carbon monoxide with hydrogen gas has more than 2000 possible pathways as estimated by Ulissi et al.[40] Nevertheless, a significant reduction in the range of possibilities was achieved by using ML in conjunction with principal component analysis and with the assistance of group additivity. Free energies were predicted cheaply, in comparison with traditional DFT, and subsequently used to rule out unfavorable pathways. These free energies were predicted with a Gaussian process regression with input features consisting of the largest ten principal components of fragment-based fingerprints. The method of Gaussian process regression builds a normal distribution over each feature, offering the benefit of understanding the uncertainty in predictions. Therefore, through the use of ML, a screening of pathways at lower computational cost was achieved, while viable pathways for catalytic reaction were identified for additional study, with some aspects of selectivity already confirmed in experimental literature.[40]

Even when a reaction pathway is known, the catalyst which will assist the most in lowering the reaction activation barrier remains to be determined. This problem was addressed for the Suzuki cross-coupling reaction[155] using machine learning to predict the reaction energies of catalysts and plotting them on a reference volcano plot.[154,156] The goal of volcano plots is to identify catalysts that bind substrates strongly, but not too strongly, thus setting them at the activity peak of the "volcano". This ML method varied both the ligands and metals and discovered 557 catalysts that fit in the Goldilocks region of the volcano plot. Starting from a database of 25 116 possible realistic structure candidates, this study revealed the ability of ML to generalize patterns across varied metal and ligand types, even when some of the test ligands were not present in the training sets.[154] Further studies using this technique could work on eliminating the reliance on reference volcano plots.

Looking practically from the experimental side, reaction yield can be of great importance not only for turnover numbers but also especially for multistep reactions where product may be lost at every step. Ahneman et al. examine the use of high-throughput experimentation for generating output labels in the form of reaction yield for the Pd-catalyzed Buchwald−Hartwig C−N cross-coupling reaction the presence of isoxazoles.[157] Using publicly available reagent geometries and scripts, catalyst descriptors were generated and used as input for multiple ML methods including KNN, neural networks, random forest, linear regression, and more. The research discovered that random forests performed the best at predicting reaction yield with an RMSE of 7.8.[157] Performing additional statistics to verify generalization uncovered the importance of understanding the underlying chemistry in train/test set splitting and found that active splitting performed better in generalization tests than random splitting.[158] This additional testing is the result of communication with data scientists and highlights the crucial nature of applying null hypotheses and other statistical methods when utilizing machine learning[159]

In an interesting twist on the concept of inverse design, Jinnouchi and Asahi computed the binding energies of NO on small-crystal slabs with DFT and then compared larger nanoparticles to the slabs by using ML techniques.[41] This research aimed at describing the catalytic activity of large heterogeneous crystal structures, an open question in surface chemistry. Specifically focusing on structures of the type $Rh_{1-x}Au_x$, this work studied the direct decomposition of NO. The advantage of ML relative to the heavy cost of DFT calculations was that it allowed for the analysis of large nanoparticles, which would otherwise be computationally intractable. The binding energies were then used to calculate catalytic activities of RhAu alloy nanoparticles. The method used a similarity kernel called Smooth Overlap Atomic Position (SOAP)[42] to calculate the overlap integrals of two single-crystal three-dimensional representations. The overlaps K_{IJ} were used to obtain the binding energy, as follows

$$E_{IJ} = \sum_J w_J K_{IJ} \tag{8}$$

where w_J were determined by Bayesian linear regression with DFT binding energies calculated for the slabs.[41] The model was then used to extrapolate to the nanoparticle by applying the similarity kernel to the Ith nanoparticle and the Jth single crystal to find the new K_{IJ} which is plugged back into the above equation. The Jth w_J is then used to determine the binding energy of NO on the nanoparticle. Addressing the important topic of catalytic conversion of nitrous oxide to nitrogen and oxygen gas in cars, this study showed that the method was practically applicable for active site surface analysis, turnover frequency, and analysis of size and composition. It was also suggested that the methodology could be expanded to examine diffusion barriers and lateral binding energies.[41] It is important to mention that kernel methods are well-defined approaches for extracting nonlinear relationships from data and are specifically designed not to depend on the particular feature

DOI: 10.1021/acs.chemrev.8b00759
Chem. Rev. 2019, 119, 6595−6612

space being examined. This ML technique is therefore apt to address a wide range of chemical problems beyond those presented in this review.[160]

While the SOAP method is a step away from inverse design, the use of slabs with known energies to form a larger structure with desired properties is similar to searching in property space to find a solicited property and then generating the corresponding structure. The use of the single-crystal structures may assist in the inverse design process as they can be used to explore the configurational space for systems with similar properties.

From the examples above, it is clear that using ML for inverse design of catalysts or understanding catalytic pathways shows promise. Therefore, it is natural to anticipate that applications of ML to catalysis and inverse design in general will continue to provide valuable insights into future developments.

4.5. Machine Learning for Configurational Sampling

One of the important challenges of computational modeling of catalytic systems is the description of the relative stability of configurations. ML could significantly improve the computational efficiency of configurational sampling. In fact, ML has already been applied to a wide range of studies where the relative energies of different configurations are critical, including the description of reactive gas-surface dynamics of N_2 on Ru(0001),[43] CO_2 adsorption on Au/Cu alloy surfaces,[44,161] and formation energies of elpasolites made from all main-group elements up to Bi.[162]

In the study of N_2 on Ru(0001), an ANN was parametrized to reproduce DFT calculations of energies and forces based on 25 000 configurations. The ANN enabled efficient calculations of the dissociation of nitrogen gas from ruthenium slabs at the DFT level of accuracy. This would have otherwise required roughly 10^5 trajectory calculations to reach convergence with the same level of error as DFT. This method produced phonon modes that closely modeled experimental results from the literature.[43]

ANN has also been used to generate DFT-grade potentials in the study of oxygen and CO_2 adsorption onto Au/Cu alloy surfaces containing up to 3915 atoms.[163] The studies were able to determine the environmental conditions that led to changes in structure, size, and composition responsible for differences in reactivity.[44] Further study of those nanoparticles found promising structures that were proposed for experimental testing.[161]

A study of elpasolites made from main-group elements developed a ML model to estimate DFT energies from descriptors based on atom types and static energy contributions.[162] The energies of 2 million elpasolite configurations were estimated using ridge regression based on the aforementioned ML model. Configurations with negative formation energies were identified, including 2133 thermodynamically allowed crystal structures determined through the phase diagram analysis. DFT revealed that 90 of those structures were stable and suitable for further analysis, demonstrating significant efficiency gains when combined with the preliminary analysis based on the ML model. The ML method explores elapsolite crystals based on formation energies, finding these energies with similar or better accuracies than DFT as compared to experiment.[162]

ML has been able to detect patterns of correlation between molecular structures and properties, as illustrated by the examples mentioned above as well as many other studies.[164−166] ML can provide gains in computational efficiency that are critical to general-purpose applications, including catalyst inverse design and beyond. In the context of polymer chemistry, a combination of GA and ML was found to be useful for the design of polymer dielectrics with a desired band gap and target dielectric constant. Polymers in this space contained four units in the repeating block that were built from the subunits of CH_2, NH, CO, C_6H_4, C_4H_2S, CS, and O. For computational efficiency, a kernel ridge regression (KRR) algorithm was utilized to map the GA structures to the property of interest, defined by the following fitness function:

$$F = (\epsilon - \epsilon_{\text{target}})^2 + (E_{\text{gap}} - E_{\text{target gap}})^2 \quad (9)$$

A KRR maps a given problem into a higher-dimensional *kernel* space where a linear regression of the data can be found. The KRR model was trained at the DFT level, using the rPW86 functional and the DFT-DF2 van der Waals correction, sampling polymer crystal structures generated from the Minima Hopping[167,168] prediction scheme. Comparison to DFT and experimental results for bandgap and total dielectric constant shows similar results, though errors are not systematic and offer an area for improvement.[45]

Combining ML and inverse design can be a powerful approach to guide the development of catalysts. However, the machine learning process can be slow when the property of interest is matched by multiple structures. That difficulty typically arises when the property of interest is not very sensitive to structural changes, when data used to train the ANN have inadequate precision, or when there is a lack of data for the ANN to pick out the differences that would lead to different property value. In any of these cases, it can be challenging for the algorithm to resolve which structure should be generated from the property. A solution to this dilemma has been explored using a combination of forward traversing and inverse direction ANN.[46] First, the forward model was trained to go from structure to property using input and labels from full-wave electromagnetic simulations. This kind of network always produces a unique property from a given structure. The inverse network was then connected to the forward model to take a spectrum as input and produced a unique design as output which in turn produced a spectrum. The final spectrum output was used to calculate the cost to update the weights in the inverse network. In that way, the inverse network was trained only for unique structure−property relations. To achieve prediction of spectral response for any hole vector within the 2^{400} combinatorial possibilities, only 20 000 simulations were needed for training.[46]

Transfer learning is a ML technique to repurpose a model trained on one task for a second related objective. In the simplest form, transfer learning adds on layers to a network pretrained for a similar task to fit a new problem while retaining the lessons from the previous purpose. Transfer learning is an excellent way to save computational time by generating structures from potentials rather than using the forward methodology.[43] At the same time, this avoids repetition of previous calculations by tailoring previous work to new research objectives. Transfer learning is one method to bridge the bottlenecks of machine learning in chemistry, many of which have been interwoven into the discussion above. Explicitly, these bottlenecks are access to data, covering enough of a property subspace to describe a desired

DOI: 10.1021/acs.chemrev.8b00759
Chem. Rev. 2019, 119, 6595−6612

relationship, complexity of chemical interactions increasing size of model needed, generalizability to vast chemical space, and representation of system. Data are the key unlocking relationships with machine learning. The generation of large data sets of chemical data, similar to those generated in the biochemistry field but particularly for catalysis, is needed for continued development of catalysts with machine learning. Such data sets must reach widely across chemical space but also deeply into the pockets of space that are examined so that complex chemical relationships can be unwrapped. In particular, relating catalytic performance to molecular properties is an open problem that nonlinear regression methods may be able to successfully tackle given enough well-selected data. Additionally, more complex relationships often require deeper models, which increase memory and processing requirements. This is perhaps why Google is able to surpass individual research teams that had been working on the protein folding problem for many more years. With the improved data sets described above, models could be trained to be generalizable to wider spaces of data, increasing the utility of individual models. Finally, representation of systems in machine learning remains an open question. The creation of a compact data structure containing all pertinent chemical interactions offers the promise of model training and preprocessing speedup, allowing individual researchers to compete more evenly with industrial giants. Each of these bottlenecks presents opportunities for research in chemistry as well as math and computer science. In the interim though, it is clear that advances in machine learning for the inverse design of catalysts have laid a foundation on which further research may be built.

5. CONCLUSIONS AND OUTLOOK

The development of new catalysts is critical for a wide range of applications, including the generation of sustainable energy. Recent developments in inverse design offer new opportunities for identifying catalysts using computationally efficient methods that bypass the need for high-throughput screening and narrow the range of compounds and catalytic materials to only those candidates with properties of interest. These can then be the subject of synthetic and experimental work. From this review, it is clear that while many advancements are tested against experimental results in the literature, few have initiated collaboration to experimentally test the validity of entirely new predicted molecules. The authors encourage such follow-up studies and think such studies are key to determining which methods are the most useful. In the very far future, these methods could be used to completely automate the design and synthesis of compounds according to the user's needs.[169] Catalytic mechanisms could even be explored by inverse design.[170] Here, we have reviewed several inverse design methods that are pertinent to catalyst development. As these methods have not yet been extensively applied to catalysis, there is ample room for new and exciting developments in this growing field.

Inverse design could be applied to modulate catalytic activity through changes in the first and second coordination spheres of the catalyst binding site. Those effects are known to be critical for functionality of catalytic cofactors in enzymes, often responsible for efficiency and selectivity.[171] In fact, such ideas have been the motivation for efforts to design molecular catalysts with structures similar to those of enzyme active sites.[172−174]

Along the same lines, in real-world applications, catalysis does not occur in a vacuum. The environment has a major impact on catalytic activity. In particular, the solvent in which catalysis occurs can play a major role in reactivity. Solvents can increase the rate constant of a reaction significantly by stabilizing the transition state or can undesirably hinder the reaction rate.[175−177] Recent efforts have focused on computationally designing solvents for improving the rates of reactions.[178] Including solvent effects in existing methods and using inverse design to identify improved solvents for catalysis will contribute to the development of more efficient catalytic systems.

From a practical standpoint, the primary objective of inverse design is to identify promising structures for further computational and experimental study. Therefore, it is important to find compounds that are straightforward to synthesize. A possible approach is to use the synthetic accessibility score,[179] typically applied for drug molecules, as part of the scoring functions of inverse design that would ensure synthetic feasibility. One method of doing such design efficiently may rely on the use of empirical parameters to describe molecules without the cost of using three-dimensional coordinates for a whole structure and without relying on a model to determine the complex interactions indirectly from the geometry.

At the computational level, progress in inverse design relies critically upon optimization algorithms. These algorithms, which dictate how an answer space is explored, could ensure higher rates of success in finding parameters that optimize the value of a scoring function. Methods, such as the recently developed Classical Optimal Control Optimization (COCO) algorithm,[180] for global energy minimization could be particularly valuable. COCO is based on diffeomorphic modulation under the observable-response-preserving homotopy (DMORPH) algorithm[181−186] and guides the classical dynamics of a probe particle. These dynamics are driven by an external field to reach the global optimum of a multidimensional function by iteratively adapting field control parameters along the direction of the gradient of the scoring function with respect to the controls. The global minimum is typically found, even for initial states far from the minimum, as long as the field has enough control parameters. While COCO has been demonstrated for model systems, an outstanding challenge is its implementation for scoring functions in inverse design applications.

Inverse design methods can benefit from ML techniques and the development of scoring functions from data analysis methods that provide patterns of correlation. These scoring functions could correlate molecular descriptors to catalytic properties to find catalysts through gradient-based optimization. Still to overcome are the difficulties previously discussed in section 4.3 that might plague the resulting scoring functions as well. As an example, very similar molecules often have very different catalytic activity due to subtle effects that must be captured by scoring functions. Such subtleties may be captured by pairing with experimentation to construct training sets of systems that differ in values of the properties in order to determine feature sets that are able to capture these properties. Through these negative tests, features can be ruled independent of the processes in question, developing new chemical understanding. ML could similarly be used to determine performance scores for GA-based methods such as PooMa.[136,137] In the realm of ML, autoencoders have been applied to transform SMILES representations of chemicals to a

DOI: 10.1021/acs.chemrev.8b00759
Chem. Rev. 2019, 119, 6595−6612

continuous latent space to optimize chemical properties, such as the synthetic accessibility score and Quantitative Estimation of Drug Likeness.[187,188] Using gradient-based methods, the latent space could be traversed to predict novel structures which could then be synthesized and tested. Such a combination of inverse design with gradient-driven optimization and ML is a prime example of using artificial intelligence methods to address the challenge of catalyst discovery.

Similarly, a wide range of related methods that have been developed to explore the chemical space for drug development could be implemented for catalyst discovery. Such methodology could be applied to examine catalytic reactivity based on empirical parameters, such as the Hammett[148] and Lever[189] ligand parameters, which provide a linear relationship between the Gibbs free energy change and reaction rates. As many catalysts consist of substituted phenyls and metal centers with varying oxidation states, Hammett and Lever parameters are expected to be particularly valuable as molecular descriptors to scan across the chemical space for ligand design. These and other empirical parameters could be used as inputs for parametrization of ANN. The labels used for output comparison could be a molecular representation or property that enables the discovery of relationships between the molecular descriptors and specific properties of interest. Such patterns could give rise to a multivariate level of understanding of catalytic reactivity. Further development of such networks could employ transfer learning to examine the effects of the environment on catalytic reactivity. As a final comment, we note that the use of structural parameters and empirical descriptors could provide significant gains in computational efficiency while at the same time narrowing the range of promising candidates to be analyzed.

Inverse design shows great promise for the development of molecular and heterogeneous catalysts. It is natural to anticipate that the early efforts highlighted in this review will certainly be expanded upon to create more robust and efficient inverse design methods. Catalysts identified through inverse design can be subjected to high-level computations followed by synthesis and experimental analysis, reducing the overall time and cost needed for discovery and development. Furthermore, inverse design will allow for the discovery of improved catalysts through rigorous multivariate analyses beyond the capabilities of the traditional approach of intuition-driven trial-and-error.

AUTHOR INFORMATION

Corresponding Author

*Phone: 203-432-6672. E-mail: victor.batista@yale.edu.

ORCID

Victor S. Batista: 0000-0002-3262-1237

Notes

The authors declare no competing financial interest.

Biographies

Jessica G. Freeze received her B.S. in Chemistry and B.A. in Computer Science from the University of Rochester. She has performed research under Prof. James Foresman, Dr. David Mathews, and Prof. Christopher Cramer and her senior thesis under Prof. David McCamant. Research topics have included metric study for automated single nucleotide polymorphism detection, high-throughput catalyst screening for transesterification of caprilactone, and study of dye-sensitized solar cells via diffuse reflectance IR Fourier transform spectroscopy and induced electron transfer calculations. Jessica is currently a PhD candidate at Yale University's Chemistry Department under the guidance of Prof. Victor Batista.

H. Ray Kelly received his B.S. in Chemistry and Computer Science from the University of Miami where he performed computational biochemistry research under the guidance of Prof. Rajeev Prabhakar. He is presently a Ph.D. student in the Chemistry Department and Energy Sciences Institute at Yale University supervised by Prof. Victor S. Batista. His current research interests include catalysis on metal surfaces and mechanistic and spectroscopic studies of CO_2 reduction catalysts in solution and attached to surfaces.

Victor S. Batista (1966, Buenos Aires, Argentina) received his Lic. Ciencias Quimicas degree from Universidad de Buenos Aires, Argentina (1989), and the Sugata Ray Award (1995) and a Ph.D. degree in Theoretical Chemistry (1996) from Boston University. After completing postdoctoral programs with William H. Miller at the University of California, Berkeley (1997−1999), and Paul Brumer at the University of Toronto (2000), he joined the Yale faculty in 2001, where he has received the ACS PRF-G6 Award (2002), the Research Corporation Innovation Award (2002), the NSF Career Award (2004), the Sloan Fellowship (2005−2006), and the Camille Dreyfus Teacher-Scholar Award (2005).

ACKNOWLEDGMENTS

The authors acknowledge Benjamin Rudshteyn for his assistance with the first stages of this review and his guidance in resource gathering. This material is based upon work supported by AFOSR Grant #FA9550-17-0198. V.S.B. acknowledges high performance computing time from both NERSC and the Yale University Faculty of Arts and Sciences High Performance Computing Center, whose acquisition was partially funded by the National Science Foundation under grant number CNS08-21132. B.R. gratefully acknowledges support from the NSF Graduate Research Fellowship under Grant No. DGE-1122492. The authors thank Prof. Dequan Xiao (University of New Haven) for writing the initial version of the EHT-LCAP code in our lab as well as Profs. David Beratan and Weitao Yang (Duke University) for creating the LCAP algorithm.

REFERENCES

(1) Kirkpatrick, P.; Ellis, C. Chemical Space. *Nature* **2004**, *432*, 823.
(2) Catlow, C. R.; Davidson, M.; Hardacre, C.; Hutchings, G. J. Catalysis Making the World a Better Place. *Philos. Trans. R. Soc., A* **2016**, *374*, 20150089.
(3) Poree, C.; Schoenebeck, F. A Holy Grail in Chemistry: Computational Catalyst Design: Feasible or Fiction? *Acc. Chem. Res.* **2017**, *50*, 605−608.
(4) Houk, K.; Cheong, P. H.-Y. Computational Prediction of Small-Molecule Catalysts. *Nature* **2008**, *455*, 309−313.
(5) Nørskov, J. K.; Abild-Pedersen, F.; Studt, F.; Bligaard, T. Density Functional Theory in Surface Chemistry and Catalysis. *Proc. Natl. Acad. Sci. U. S. A.* **2011**, *108*, 937−943.
(6) Thiel, W. Computational Catalysis−Past, Present, and Future. *Angew. Chem., Int. Ed.* **2014**, *53*, 8605−8613.
(7) Sperger, T.; Sanhueza, I. A.; Schoenebeck, F. Computation and Experiment: A Powerful Combination to Understand and Predict Reactivities. *Acc. Chem. Res.* **2016**, *49*, 1311−1319.
(8) Greeley, J.; Jaramillo, T. F.; Bonde, J.; Chorkendorff, I.; Nørskov, J. K. Computational High-Throughput Screening of Electrocatalytic Materials for Hydrogen Evolution. *Nat. Mater.* **2006**, *5*, 909−913.
(9) Andersson, M. P.; Bligaard, T.; Kustov, A.; Larsen, K. E.; Greeley, J.; Johannessen, T.; Christensen, C. H.; Nørskov, J. K.

DOI: 10.1021/acs.chemrev.8b00759
Chem. Rev. 2019, 119, 6595−6612

Toward Computational Screening in Heterogeneous Catalysis: Pareto-Optimal Methanation Catalysts. *J. Catal.* **2006**, *239*, 501−506.
(10) Sehested, J.; Larsen, K. E.; Kustov, A. L.; Frey, A. M.; Johannessen, T.; Bligaard, T.; Andersson, M. P.; Nørskov, J. K.; Christensen, C. H. Discovery of Technical Methanation Catalysts Based on Computational Screening. *Top. Catal.* **2007**, *45*, 9−13.
(11) Xiao, D.; Warnke, I.; Bedford, J.; Batista, V. S. *Chemical Modelling*; The Royal Society of Chemistry, 2014; Vol. *10*, pp 1−31.
(12) Weymuth, T.; Reiher, M. Inverse Quantum Chemistry: Concepts and Strategies for Rational Compound Design. *Int. J. Quantum Chem.* **2014**, *114*, 823−837.
(13) von Lilienfeld, O. A. *Many-Electron Approaches in Physics, Chemistry and Mathematics*; Springer, 2014; pp 169−189.
(14) Franceschetti, A.; Zunger, A. The Inverse Band-Structure Problem of Finding an Atomic Configuration with Given Electronic Properties. *Nature* **1999**, *402*, 60−63.
(15) Dudiy, S. V.; Zunger, A. Searching for Alloy Configurations with Target Physical Properties: Impurity Design via a Genetic Algorithm Inverse Band Structure Approach. *Phys. Rev. Lett.* **2006**, *97*, No. 046401.
(16) Piquini, P.; Graf, P. A.; Zunger, A. Band-Gap Design of Quaternary (In,Ga)(As,Sb) Semiconductors via the Inverse-Band-Structure Approach. *Phys. Rev. Lett.* **2008**, *100*, 186403.
(17) Xiao, D.; Martini, L. A.; Snoeberger, R. C., III; Crabtree, R. H.; Batista, V. S. Inverse Design and Synthesis of acac-Coumarin Anchors for Robust TiO_2 Sensitization. *J. Am. Chem. Soc.* **2011**, *133*, 9014−9022.
(18) Xiao, D.; Yang, W.; Beratan, D. N. Inverse Molecular Design in a Tight-Binding Framework. *J. Chem. Phys.* **2008**, *129*, No. 044106.
(19) Cerqueira, T. F. T.; Sarmiento-Pérez, R.; Amsler, M.; Nogueira, F.; Botti, S.; Marques, M. A. L. Materials Design On-The-Fly. *J. Chem. Theory Comput.* **2015**, *11*, 3955−3960.
(20) Martin, R. L.; Haranczyk, M. Optimization-Based Design of Metal-Organic Framework Materials. *J. Chem. Theory Comput.* **2013**, *9*, 2816−2825.
(21) Weigend, F.; Schrodt, C.; Ahlrichs, R. Atom Distributions in Binary Atom Clusters: A Perturbational Approach and its Validation in a Case Study. *J. Chem. Phys.* **2004**, *121*, 10380−10384.
(22) Weymuth, T.; Reiher, M. Gradient-Driven Molecule Construction: An Inverse Approach Applied to the Design of Small-Molecule Fixating Catalysts. *Int. J. Quantum Chem.* **2014**, *114*, 838−850.
(23) Sheppard, D.; Henkelman, G.; von Lilienfeld, O. A. Alchemical Derivatives of Reaction Energetics. *J. Chem. Phys.* **2010**, *133*, No. 084104.
(24) Saravanan, K.; Kitchin, J. R.; von Lilienfeld, O. A.; Keith, J. A. Alchemical Predictions for Computational Catalysis: Potential and Limitations. *J. Phys. Chem. Lett.* **2017**, *8*, 5002−5007.
(25) Balamurugan, D.; Yang, W.; Beratan, D. N. Exploring Chemical Space with Discrete, Gradient, and Hybrid Optimization Methods. *J. Chem. Phys.* **2008**, *129*, 174105.
(26) De Vleeschouwer, F.; Yang, W.; Beratan, D. N.; Geerlings, P.; De Proft, F. Inverse Design of Molecules with Optimal Reactivity Properties: Acidity of 2-Naphthol Derivatives. *Phys. Chem. Chem. Phys.* **2012**, *14*, 16002−16013.
(27) Keinan, S.; Hu, X.; Beratan, D. N.; Yang, W. Designing Molecules with Optimal Properties Using the Linear Combination of Atomic Potentials Approach in an AM1 Semiempirical Framework. *J. Phys. Chem. A* **2007**, *111*, 176−181.
(28) Hu, X.; Beratan, D. N.; Yang, W. Emergent Strategies for Inverse Molecular Design. *Sci. China, Ser. B: Chem.* **2009**, *52*, 1769−1776.
(29) Keinan, S.; Paquette, W. D.; Skoko, J. J.; Beratan, D. N.; Yang, W.; Shinde, S.; Johnston, P. A.; Lazo, J. S.; Wipf, P. Computational Design, Synthesis and Biological Evaluation of Para-Quinone-Based Inhibitors for Redox Regulation of the Dual-Specificity Phosphatase Cdc25B. *Org. Biomol. Chem.* **2008**, *6*, 3256−3263.
(30) Keinan, S.; Therien, M. J.; Beratan, D. N.; Yang, W. Molecular Design of Porphyrin-Based Nonlinear Optical Materials. *J. Phys. Chem. A* **2008**, *112*, 12203−12207.
(31) Kuhn, C.; Beratan, D. N. Inverse Strategies for Molecular Design. *J. Phys. Chem.* **1996**, *100*, 10595−10599.
(32) Wang, M.; Hu, X.; Beratan, D. N.; Yang, W. Designing Molecules by Optimizing Potentials. *J. Am. Chem. Soc.* **2006**, *128*, 3228−3232.
(33) Xiao, D.; Hu, R. A Tutorial of the Inverse Molecular Design Theory in Tight-Binding Frameworks and Its Applications. *Handbook of Green Chemistry, Tools for Green Chemistry*; American Cancer Society, 2017; Vol. *10*, pp 169−188.
(34) von Lilienfeld, O. A.; Lins, R. D.; Rothlisberger, U. Variational Particle Number Approach for Rational Compound Design. *Phys. Rev. Lett.* **2005**, *95*, 153002.
(35) von Lilienfeld, O. A.; Tuckerman, M. E. Molecular Grand-Canonical Ensemble Density Functional Theory and Exploration of Chemical Space. *J. Chem. Phys.* **2006**, *125*, 154104.
(36) von Lilienfeld, O. A.; Tuckerman, M. E. Alchemical Variations of Intermolecular Energies According to Molecular Grand-Canonical Ensemble Density Functional Theory. *J. Chem. Theory Comput.* **2007**, *3*, 1083−1090.
(37) Anatole von Lilienfeld, O. Accurate Ab Initio Energy Gradients in Chemical Compound Space. *J. Chem. Phys.* **2009**, *131*, 164102.
(38) von Lilienfeld, O. A. First Principles View on Chemical Compound Space: Gaining Rigorous Atomistic Control of Molecular Properties. *Int. J. Quantum Chem.* **2013**, *113*, 1676−1689.
(39) Janet, J. P.; Chan, L.; Kulik, H. J. Accelerating Chemical Discovery with Machine Learning: Simulated Evolution of Spin Crossover Complexes with an Artificial Neural Network. *J. Phys. Chem. Lett.* **2018**, *9*, 1064−1071.
(40) Ulissi, Z. W.; Medford, A. J.; Bligaard, T.; Nørskov, J. K. To Address Surface Reaction Network Complexity Using Scaling Relations Machine Learning and DFT Calculations. *Nat. Commun.* **2017**, *8*, 14621.
(41) Jinnouchi, R.; Asahi, R. Predicting Catalytic Activity of Nanoparticles by a DFT-Aided Machine-Learning Algorithm. *J. Phys. Chem. Lett.* **2017**, *8*, 4279−4283.
(42) Bartók, A. P.; Kondor, R.; Csányi, G. On Representing Chemical Environments. *Phys. Rev. B: Condens. Matter Mater. Phys.* **2013**, *87*, 184115.
(43) Shakouri, K.; Behler, J.; Meyer, J.; Kroes, G.-J. Accurate Neural Network Description of Surface Phonons in Reactive Gas-Surface Dynamics: N2 + Ru(0001). *J. Phys. Chem. Lett.* **2017**, *8*, 2131−2136.
(44) Artrith, N.; Kolpak, A. M. Understanding the Composition and Activity of Electrocatalytic Nanoalloys in Aqueous Solvents: A Combination of DFT and Accurate Neural Network Potentials. *Nano Lett.* **2014**, *14*, 2670−2676.
(45) Mannodi-Kanakkithodi, A.; Pilania, G.; Huan, T. D.; Lookman, T.; Ramprasad, R. Machine Learning Strategy for Accelerated Design of Polymer Dielectrics. *Sci. Rep.* **2016**, *6*, 20952.
(46) Liu, D.; Tan, Y.; Khoram, E.; Yu, Z. Training Deep Neural Networks for the Inverse Design of Nanophotonic Structures. *ACS Photonics* **2018**, *5*, 1365−1369.
(47) Sanchez-Lengeling, B.; Aspuru-Guzik, A. Inverse Molecular Design Using Machine Learning: Generative Models for Matter Engineering. *Science* **2018**, *361*, 360−365.
(48) Moore, K. W.; Pechen, A.; Feng, X.-J.; Dominy, J.; Beltrani, V.; Rabitz, H. Universal Characteristics of Chemical Synthesis and Property Optimization. *Chem. Sci.* **2011**, *2*, 417−424.
(49) Moore, K. W.; Pechen, A.; Feng, X.-J.; Dominy, J.; Beltrani, V. J.; Rabitz, H. Why is Chemical Synthesis and Property Optimization Easier Than Expected? *Phys. Chem. Chem. Phys.* **2011**, *13*, 10048−10070.
(50) Tibbetts, K. M.; Feng, X.-J.; Rabitz, H. Exploring Experimental Fitness Landscapes for Chemical Synthesis and Property Optimization. *Phys. Chem. Chem. Phys.* **2017**, *19*, 4266−4287.
(51) Schrock, R. R. Catalytic Reduction of Dinitrogen to Ammonia at a Single Molybdenum Center. *Acc. Chem. Res.* **2005**, *38*, 955−962.

DOI: 10.1021/acs.chemrev.8b00759
Chem. Rev. 2019, 119, 6595−6612

(52) Kohn, W.; Sham, L. J. Self-Consistent Equations Including Exchange and Correlation Effects. *Phys. Rev.* **1965**, *140*, A1133−A1138.
(53) Kirkwood, J. G. Statistical Mechanics of Fluid Mixtures. *J. Chem. Phys.* **1935**, *3*, 300−313.
(54) Zwanzig, R. W. High-Temperature Equation of State by a Perturbation Method.I. Nonpolar Gases. *J. Chem. Phys.* **1954**, *22*, 1420−1426.
(55) Jorgensen, W. L.; Ravimohan, C. Monte Carlo Simulation of Differences in Free Energies of Hydration. *J. Chem. Phys.* **1985**, *83*, 3050−3054.
(56) van Gunsteren, W. F.; Berendsen, H. J. C. Thermodynamic Cycle Integration by Computer Simulation as a Tool for Obtaining Free Energy Differences in Molecular Chemistry. *J. Comput.-Aided Mol. Des.* **1987**, *1*, 171−176.
(57) Yang, J.; Cooper, J. K.; Toma, F. M.; Walczak, K. A.; Favaro, M.; Beeman, J. W.; Hess, L. H.; Wang, C.; Zhu, C.; Gul, S.; et al. A Multifunctional Biphasic Water Splitting Catalyst Tailored for Integration with High-Performance Semiconductor Photoanodes. *Nat. Mater.* **2017**, *16*, 335−341.
(58) Chen, J.; Wu, K.; Rudshteyn, B.; Jia, Y.; Ding, W.; Xie, Z.-X.; Batista, V. S.; Lian, T. Ultrafast Photoinduced Interfacial Proton Coupled Electron Transfer from CdSe Quantum Dots to 4,4'-Bipyridine. *J. Am. Chem. Soc.* **2016**, *138*, 884−892.
(59) Sheehan, S. W.; Thomsen, J. M.; Hintermair, U.; Crabtree, R. H.; Brudvig, G. W.; Schmuttenmaer, C. A. A Molecular Catalyst for Water Oxidation That Binds to Metal Oxide Surfaces. *Nat. Commun.* **2015**, *6*, 6469.
(60) Windle, C. D.; Pastor, E.; Reynal, A.; Whitwood, A. C.; Vaynzof, Y.; Durrant, J. R.; Perutz, R. N.; Reisner, E. Improving the Photocatalytic Reduction of CO_2 to CO through Immobilisation of a Molecular Re Catalyst on TiO_2. *Chem. - Eur. J.* **2015**, *21*, 3746−3754.
(61) Zhao, Y.; Yang, K. R.; Wang, Z.; Yan, X.; Cao, S.; Ye, Y.; Dong, Q.; Zhang, X.; Thorne, J. E.; Jin, L.; et al. Stable Iridium Dinuclear Heterogeneous Catalysts Supported on Metal-Oxide Substrate for Solar Water Oxidation. *Proc. Natl. Acad. Sci. U. S. A.* **2018**, *115*, 2902−2907.
(62) Nellist, M. R.; Laskowski, F. A. L.; Lin, F.; Mills, T. J.; Boettcher, S. W. Semiconductor-Electrocatalyst Interfaces: Theory, Experiment, and Applications in Photoelectrochemical Water Splitting. *Acc. Chem. Res.* **2016**, *49*, 733−740.
(63) Willkomm, J.; Orchard, K. L.; Reynal, A.; Pastor, E.; Durrant, J. R.; Reisner, E. Dye-Sensitised Semiconductors Modified with Molecular Catalysts for Light-Driven H_2 Production. *Chem. Soc. Rev.* **2016**, *45*, 9−23.
(64) Hu, X.; Beratan, D. N.; Yang, W. A Gradient-Directed Monte Carlo Approach to Molecular Design. *J. Chem. Phys.* **2008**, *129*, No. 064102.
(65) Barona-Castaño, J. C.; Carmona-Vargas, C. C.; Brocksom, T. J.; de Oliveira, K. T. Porphyrins as Catalysts in Scalable Organic Reactions. *Molecules* **2016**, *21*, 310.
(66) Oldacre, A. N.; Friedman, A. E.; Cook, T. R. A Self-Assembled Cofacial Cobalt Porphyrin Prism for Oxygen Reduction Catalysis. *J. Am. Chem. Soc.* **2017**, *139*, 1424−1427.
(67) Rybicka-Jasinska, K.; Shan, W.; Zawada, K.; Kadish, K. M.; Gryko, D. Porphyrins as Photoredox Catalysts: Experimental and Theoretical Studies. *J. Am. Chem. Soc.* **2016**, *138*, 15451−15458.
(68) Teunissen, J. L.; De Proft, F.; De Vleeschouwer, F. Tuning the HOMO−LUMO Energy Gap of Small Diamondoids Using Inverse Molecular Design. *J. Chem. Theory Comput.* **2017**, *13*, 1351−1365.
(69) Chang, A. M.; Rudshteyn, B.; Warnke, I.; Batista, V. S. Inverse Design of a Catalyst for Aqueous CO/CO_2 Conversion Informed by the Ni^{II}-Iminothiolate Complex. *Inorg. Chem.* **2018**, *57*, 15474−15480.
(70) Lu, Z.; White, C.; Rheingold, A. L.; Crabtree, R. H. Functional Modeling of CO Dehydrogenase: Catalytic Reduction of Methylviologen by CO/H_2O with an N, O, S-Ligated Nickel Catalyst. *Angew. Chem., Int. Ed. Engl.* **1993**, *32*, 92−94.
(71) Lu, Z.; Crabtree, R. H. A Functional Modeling Study of the CO Oxidation Site of Nickel CO Dehydrogenase. *J. Am. Chem. Soc.* **1995**, *117*, 3994−3998.
(72) Kollman, P. Free Energy Calculations: Applications to Chemical and Biochemical Phenomena. *Chem. Rev.* **1993**, *93*, 2395−2417.
(73) Hansen, N.; van Gunsteren, W. F. Practical Aspects of Free-Energy Calculations: A Review. *J. Chem. Theory Comput.* **2014**, *10*, 2632−2647.
(74) Hohenberg, P.; Kohn, W. Inhomogeneous Electron Gas. *Phys. Rev.* **1964**, *136*, B864−B871.
(75) Feynman, R. P. Forces in Molecules. *Phys. Rev.* **1939**, *56*, 340−343.
(76) Nørskov, J. K.; Rossmeisl, J.; Logadottir, A.; Lindqvist, L.; Kitchin, J. R.; Bligaard, T.; Jónsson, H. Origin of the Overpotential for Oxygen Reduction at a Fuel-Cell Cathode. *J. Phys. Chem. B* **2004**, *108*, 17886−17892.
(77) Sui, S.; Wang, X.; Zhou, X.; Su, Y.; Riffat, S.; Liu, C.-J. A Comprehensive Review of Pt Electrocatalysts for the Oxygen Reduction Reaction: Nanostructure, Activity, Mechanism and Carbon Support in PEM Fuel Cells. *J. Mater. Chem. A* **2017**, *5*, 1808−1825.
(78) Chang, K. Y. S.; Fias, S.; Ramakrishnan, R.; von Lilienfeld, O. A. Fast and Accurate Predictions of Covalent Bonds in Chemical Space. *J. Chem. Phys.* **2016**, *144*, 174110.
(79) Al-Hamdani, Y. S.; Michaelides, A.; von Lilienfeld, O. A. Exploring Dissociative Water Adsorption on Isoelectronically BN Doped Graphene Using Alchemical Derivatives. *J. Chem. Phys.* **2017**, *147*, 164113.
(80) Balawender, R.; Lesiuk, M.; De Proft, F.; Geerlings, P. Exploring Chemical Space with Alchemical Derivatives: BN-Simultaneous Substitution Patterns in C_{60}. *J. Chem. Theory Comput.* **2018**, *14*, 1154−1168.
(81) Chang, K. Y. S.; von Lilienfeld, O. A. Al_xGa_{1-x} As Crystals with Direct 2 eV Band Gaps from Computational Alchemy. *Phys. Rev. Materials.* **2018**, *2*, No. 073802.
(82) Weigend, F.; Schrodt, C. Atom-Type Assignment in Molecules and Clusters by Perturbation Theory−A Complement to X-ray Structure Analysis. *Chem. - Eur. J.* **2005**, *11*, 3559−3564.
(83) Weigend, F. Extending DFT-Based Genetic Algorithms by Atom-to-Place Re-Assignment via Perturbation Theory: A Systematic and Unbiased Approach to Structures of Mixed-Metallic Clusters. *J. Chem. Phys.* **2014**, *141*, 134103.
(84) Seifried, C.; Longo, L.; Pollak, P.; Weigend, F. The Chemical Space of $Pb_{N-n}Bi_n$ and $(Pb_{N-n}Bi_n)^+$: A Systematic Study for N = 3−13. *J. Chem. Phys.* **2017**, *146*, No. 034304.
(85) to Baben, M.; Achenbach, J. O.; von Lilienfeld, O. A. Guiding Ab Initio Calculations by Alchemical Derivatives. *J. Chem. Phys.* **2016**, *144*, 104103.
(86) Solovyeva, A.; von Lilienfeld, O. A. Alchemical Screening of Ionic Crystals. *Phys. Chem. Chem. Phys.* **2016**, *18*, 31078−31091.
(87) Munoz, M.; Cardenas, C. How Predictive Could Alchemical Derivatives Be? *Phys. Chem. Chem. Phys.* **2017**, *19*, 16003−16012.
(88) Lesiuk, M.; Balawender, R.; Zachara, J. Higher Order Alchemical Derivatives from Coupled Perturbed Self-Consistent Field Theory. *J. Chem. Phys.* **2012**, *136*, No. 034104.
(89) Balawender, R.; Welearegay, M. A.; Lesiuk, M.; De Proft, F.; Geerlings, P. Exploring Chemical Space with the Alchemical Derivatives. *J. Chem. Theory Comput.* **2013**, *9*, 5327−5340.
(90) Miranda-Quintana, R. A.; Ayers, P. W. Interpolating Hamiltonians in Chemical Compound Space. *Int. J. Quantum Chem.* **2017**, *117*, No. e25384.
(91) Samuel, A. L. In *Computer Games I*; Levy, D. N. L., Ed.; Springer New York: New York, NY, 1988; pp 335−365.
(92) Michalski, R.; Carbonell, J.; Mitchell, T. *Machine Learning: An Artificial Intelligence Approach*; Morgan Kaufmann Publ Inc., 1984.
(93) Nielsen, M. A. *Neural Networks and Deep Learning*, 1970 (accessed Dec 4, 2018). http://neuralnetworksanddeeplearning.com/.

DOI: 10.1021/acs.chemrev.8b00759
Chem. Rev. 2019, 119, 6595−6612

(94) Anzai, Y. *Pattern Recognition and Machine Learning*; Academic Press: Boston, 1992.
(95) Krizhevsky, A.; Sutskever, I.; Hinton, G. E. In *Advances in Neural Information Processing Systems 25*; Pereira, F., Burges, C. J. C., Bottou, L., Weinberger, K. Q., Eds.; Curran Associates, Inc., 2012; pp 1097−1105.
(96) Shi, J.; Malik, J. Normalized Cuts and Image Segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **2000**, *22*, 888−905.
(97) Collobert, R.; Weston, J. A Unified Architecture for Natural Language Processing: Deep Neural Networks with Multitask Learning. *Proceedings of the 25th International Conference on Machine Learning*; Association for Computing Machinery: New York, NY, USA, 2008; pp 160−167.
(98) Sebastiani, F. Machine Learning in Automated Text Categorization. *ACM Comput. Surv.* **2002**, *34*, 1−47.
(99) Pang, B.; Lee, L.; Vaithyanathan, S. Thumbs Up?: Sentiment Classification Using Machine Learning Techniques. *Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2002; Vol. *10*, pp 79−86.
(100) Pohl, I. Heuristic Search Viewed as Path Finding in a Graph. *Artificial Intelligence* **1970**, *1*, 193−204.
(101) Smith, J. S.; Isayev, O.; Roitberg, A. E. ANI-1: An Extensible Neural Network Potential with DFT Accuracy at Force Field Computational Cost. *Chem. Sci.* **2017**, *8*, 3192−3203.
(102) Li, Y.; Li, H.; Pickard, F. C.; Narayanan, B.; Sen, F. G.; Chan, M. K. Y.; Sankaranarayanan, S. K. R. S.; Brooks, B. R.; Roux, B. Machine Learning Force Field Parameters from Ab Initio Data. *J. Chem. Theory Comput.* **2017**, *13*, 4492−4503.
(103) Brockherde, F.; Vogt, L.; Li, L.; Tuckerman, M. E.; Burke, K.; Müller, K.-R. Bypassing the Kohn-Sham Equations with Machine Learning. *Nat. Commun.* **2017**, *8*, 872.
(104) Chen, H.; Engkvist, O.; Wang, Y.; Olivecrona, M.; Blaschke, T. The Rise of Deep Learning in Drug Discovery. *Drug Discovery Today* **2018**, *23*, 1241−1250.
(105) Liu, Y.; Zhao, T.; Ju, W.; Shi, S. Materials Discovery and Design Using Machine Learning. *J. Materiomics* **2017**, *3*, 159−177.
(106) Lei, T.; Chen, F.; Liu, H.; Sun, H.; Kang, Y.; Li, D.; Li, Y.; Hou, T. ADMET Evaluation in Drug Discovery. Part 17: Development of Quantitative and Qualitative Prediction Models for Chemical-Induced Respiratory Toxicity. *Mol. Pharmaceutics* **2017**, *14*, 2407−2421.
(107) Wei, J. N.; Duvenaud, D.; Aspuru-Guzik, A. Neural Networks for the Prediction of Organic Chemistry Reactions. *ACS Cent. Sci.* **2016**, *2*, 725−732.
(108) Yada, A.; Nagata, K.; Ando, Y.; Matsumura, T.; Ichinoseki, S.; Sato, K. Machine Learning Approach for Prediction of Reaction Yield with Simulated Catalyst Parameters. *Chem. Lett.* **2018**, *47*, 284−287.
(109) Google. Google Trends - Machine Learning (accessed Dec 4, 2018). https://trends.google.com/trends/explore?date=all&q=machinelearning.
(110) O'Connor, N. J.; Jonayat, A. S. M.; Janik, M. J.; Senftle, T. P. Interaction Trends Between Single Metal Atoms and Oxide Supports Identified with Density Functional Theory and Statistical Learning. *Nat. Catal.* **2018**, *1*, 531−539.
(111) Goldsmith, B. R.; Esterhuizen, J.; Liu, J.-X.; Bartel, C. J.; Sutton, C. Machine Learning for Heterogeneous Catalyst Design and Discovery. *AIChE J.* **2018**, *64*, 2311−2323.
(112) Jinnouchi, R.; Asahi, R. Predicting Catalytic Activity of Nanoparticles by a DFT-Aided Machine-Learning Algorithm. *J. Phys. Chem. Lett.* **2017**, *8*, 4279−4283.
(113) Reichenbach, H. *The Direction of Time*; Dover: Boston, 1956.
(114) Bontempi, G.; Flauder, M. From Dependency to Causality: A Machine Learning Approach. *J. Mach. Learn. Res.* **2015**, *16*, 2437−2457.
(115) Chen, J.; Swamidass, S. J.; Dou, Y.; Bruand, J.; Baldi, P. ChemDB: A Public Database of Small Molecules and Related Chemoinformatics Resources. *Bioinformatics* **2005**, *21*, 4133−4139.
(116) Pence, H. E.; Williams, A. ChemSpider: An Online Chemical Information Resource. *J. Chem. Educ.* **2010**, *87*, 1123−1124.
(117) Kim, S.; Thiessen, P.; Bolton, E.; Chen, J.; Fu, G.; Gindulyte, A.; Han, L.; He, J.; He, S.; Shoemaker, B.; et al. PubChem Substance and Compound Databases. *Nucleic Acids Res.* **2016**, *44*, D1202−13.
(118) Virshup, A. M.; Contreras-Garcia, J.; Wipf, P.; Yang, W.; Beratan, D. N. Stochastic Voyages Into Uncharted Chemical Space Produce a Representative Library of All Possible Drug-Like Compounds. *J. Am. Chem. Soc.* **2013**, *135*, 7296−7303.
(119) Rupakheti, C.; Virshup, A.; Yang, W.; Beratan, D. N. Strategy to Discover Diverse Optimal Molecules in the Small Molecule Universe. *J. Chem. Inf. Model.* **2015**, *55*, 529−537.
(120) Rupakheti, C.; Al-Saadon, R.; Zhang, Y.; Virshup, A. M.; Zhang, P.; Yang, W.; Beratan, D. N. Diverse Optimal Molecular Libraries for Organic Light-Emitting Diodes. *J. Chem. Theory Comput.* **2016**, *12*, 1942−1952.
(121) Eremeev, A. V. A Genetic Algorithm with a Non-Binary Representation for the Set Covering Problem. *Operations Research Proceedings* **1999**, *1999*, 175−181.
(122) Mirshekarian, S.; Süer, G. A. Experimental Study of Seeding in Genetic Algorithms with Non-Binary Genetic Representation. *J. Intell. Manuf.* **2018**, *29*, 1637−1646.
(123) McCall, J. Genetic Algorithms for Modelling and Optimisation. *J. Comp. Appl. Math.* **2005**, *184*, 205−222.
(124) Thierens, D.; Goldberg, D. Convergence Models of Genetic Algorithm Selection Schemes. *Parallel Problem Solving from Nature — PPSN III*; Berlin, Heidelberg, 1994; pp 119−129.
(125) Phyu, S. P. T. P.; Srijuntongsiri, G. Effect of the Number of Parents on the Performance of Multi-Parent Genetic Algorithm. *2016 11th International Conference on Knowledge, Information and Creativity Support Systems (KICSS)*. 2016; pp 1−6.
(126) Eiben, A. E.; Raué, P. E.; Ruttkay, Z. Genetic Algorithms with Multi-Parent Recombination. *Parallel Problem Solving from Nature — PPSN III*; Berlin, Heidelberg, 1994; pp 78−87.
(127) Hasançebi, O.; Erbatur, F. Evaluation of Crossover Techniques in Genetic Algorithm Based Optimum Structural Design. *Comput. Struct.* **2000**, *78*, 435−448.
(128) Kim, K.; Graf, P. A.; Jones, W. B. A Genetic Algorithm Based Inverse Band Structure Method for Semiconductor Alloys. *J. Comput. Phys.* **2005**, *208*, 735−760.
(129) Getsoian, A. B.; Zhai, Z.; Bell, A. T. Band-Gap Energy as a Descriptor of Catalytic Activity for Propene Oxidation over Mixed Metal Oxide Catalysts. *J. Am. Chem. Soc.* **2014**, *136*, 13684−13697.
(130) Froemming, N. S.; Henkelman, G. Optimizing Core-shell Nanoparticle Catalysts with a Genetic Algorithm. *J. Chem. Phys.* **2009**, *131*, 234103.
(131) Hammer, B.; Nørskov, J. Why Gold is the Noblest of All the Metals. *Nature* **1995**, *376*, 238.
(132) Hammer, B.; Nørskov, J. Electronic Factors Determining the Reactivity of Metal Surfaces. *Surf. Sci.* **1995**, *343*, 211−220.
(133) Sokalski, W. A. Theoretical Model for Exploration of Catalytic Activity of Enzymes and Design of New Catalysts: CO_2 Hydration Reaction. *Int. J. Quantum Chem.* **1981**, *20*, 231−240.
(134) Sokalski, W. A. Nonempirical Modeling of the Static and Dynamic Properties of the Optimum Environment for Chemical Reactions. *J. Mol. Struct.: THEOCHEM* **1986**, *138*, 77−87.
(135) Dittner, M.; Hartke, B. Globally Optimal Catalytic Fields - Inverse Design of Abstract Embeddings for Maximum Reaction Rate Acceleration. *J. Chem. Theory Comput.* **2018**, *14*, 3547−3564.
(136) Springborg, M.; Kohaut, S.; Dong, Y.; Huwig, K. Mixed Si-Ge Clusters, Solar-Energy Harvesting, and Inverse-Design Methods. *Comput. Theor. Chem.* **2017**, *1107*, 14−22.
(137) Huwig, K.; Fan, C.; Springborg, M. From Properties to Materials: An Efficient and Simple Approach. *J. Chem. Phys.* **2017**, *147*, 234105.
(138) O'Boyle, N. M.; Campbell, C. M.; Hutchison, G. R. Computational Design and Selection of Optimal Organic Photovoltaic Materials. *J. Phys. Chem. C* **2011**, *115*, 16200−16210.

DOI: 10.1021/acs.chemrev.8b00759
Chem. Rev. 2019, 119, 6595−6612

(139) Kanal, I. Y.; Owens, S. G.; Bechtel, J. S.; Hutchison, G. R. Efficient Computational Screening of Organic Polymer Photovoltaics. *J. Phys. Chem. Lett.* **2013**, *4*, 1613−1623.
(140) Shu, Y.; Levine, B. G. Simulated Evolution of Fluorophores for Light Emitting Diodes. *J. Chem. Phys.* **2015**, *142*, 104104.
(141) Le, T. C.; Winkler, D. A. Discovery and Optimization of Materials Using Evolutionary Approaches. *Chem. Rev.* **2016**, *116*, 6107−6132.
(142) Wolf, D.; Buyevskaya, O.; Baerns, M. An Evolutionary Approach in the Combinatorial Selection and Optimization of Catalytic Materials. *Appl. Catal., A* **2000**, *200*, 63−77.
(143) De Vleeschouwer, F.; Chankisjijev, A.; Yang, W.; Geerlings, P.; De Proft, F. Pushing the Boundaries of Intrinsically Stable Radicals: Inverse Design Using the Thiadiazinyl Radical as a Template. *J. Org. Chem.* **2013**, *78*, 3151−3158.
(144) De Vleeschouwer, F.; Chankisjijev, A.; Geerlings, P.; De Proft, F. Designing Stable Radicals with Highly Electrophilic or Nucleophilic Character: Thiadiazinyl as a Case Study. *Eur. J. Org. Chem.* **2015**, *2015*, 506−513.
(145) Silver, D.; Schrittwieser, J.; Simonyan, K.; Antonoglou, I.; Huang, A.; Guez, A.; Hubert, T.; Baker, L.; Lai, M.; Bolton, A.; et al. Mastering the Game of Go Without Human Knowledge. *Nature* **2017**, *550*, 354−359.
(146) Gómez-Bombarelli, R.; Aguilera-Iparraguirre, J.; Hirzel, T. D.; Duvenaud, D.; Maclaurin, D.; Blood-Forsythe, M. A.; Sik Chae, H.; Einzinger, M.; Ha, D.-G.; Wu, T.; et al. Design of Efficient Molecular Organic Light-Emitting Diodes by a High-Throughput Virtual Screening and Experimental Approach. *Nat. Mater.* **2016**, *15*, 1120−1127.
(147) Kar, S.; Sizochenko, N.; Ahmed, L.; Batista, V. S.; Leszczynski, J. Quantitative Structure-Property Relationship Model Leading to Virtual Screening of Fullerene Derivatives: Exploring Structural Attributes Critical for Photoconversion Efficiency of Polymer Solar Cell Acceptors. *Nano Energy* **2016**, *26*, 677−691.
(148) Hammett, L. P. The Effect of Structure upon the Reactions of Organic Compounds. Benzene Derivatives. *J. Am. Chem. Soc.* **1937**, *59*, 96−103.
(149) Cherkasov, A.; Muratov, E. N.; Fourches, D.; Varnek, A.; Baskin, I. I.; Cronin, M.; Dearden, J.; Gramatica, P.; Martin, Y. C.; Todeschini, R.; et al. QSAR Modeling: Where Have You Been? Where Are You Going To? *J. Med. Chem.* **2014**, *57*, 4977−5010.
(150) Worth, A.; Fuart-Gatnik, M.; Lapenna, S.; Serafimova, R. Applicability of QSAR Analysis in the Evaluation of Developmental and Neurotoxicity Effects for the Assessment of the Toxicological Relevance of Metabolites and Degradates of Pesticide Active Substances for Dietary Risk Assessment. *EFSA Supporting Publications* **2011**, *8*, 169E.
(151) Maggiora, G. M. On Outliers and Activity Cliffs - Why QSAR Often Disappoints. *J. Chem. Inf. Model.* **2006**, *46*, 1535−1535.
(152) Hassabis, D.; Jumper, J.; Senior, A.; Evans, R.; Kirkpatrick, J.; Sifre, L.; Green, T.; Qin, C.; Zidek, A.; Nelson, S. et al. AlphaFold: Using AI for Scientific Discovery, 2018 (Accessed Dec 7, 2018). https://deepmind.com/blog/alphafold/.
(153) Stones, J. Google's DeepMind Bests Experts at Predicting 3D Protein Shapes, 2018 (Accessed Dec 7, 2018). https://www.alphr.com/artificial-intelligence/1010276/google-s-deepmind-bests-experts-at-predicting-3d-protein-shapes.
(154) Meyer, B.; Sawatlon, B.; Heinen, S.; von Lilienfeld, O. A.; Corminboeuf, C. Machine Learning Meets Volcano Plots: Computational Discovery of Cross-Coupling Catalysts. *Chem. Sci.* **2018**, *9*, 7069−7077.
(155) Miyaura, N.; Yamada, K.; Suzuki, A. A New Stereospecific Cross-Coupling by the Palladium-Catalyzed Reaction of 1-Alkenylboranes with 1-Alkenyl or 1-Alkynyl Halides. *Tetrahedron Lett.* **1979**, *20*, 3437−3440.
(156) Busch, M.; Wodrich, M. D.; Corminboeuf, C. Linear Scaling Relationships and Volcano Plots in Homogeneous Catalysis - Revisiting the Suzuki Reaction. *Chem. Sci.* **2015**, *6*, 6754−6761.
(157) Ahneman, D. T.; Estrada, J. G.; Lin, S.; Dreher, S. D.; Doyle, A. G. Predicting Reaction Performance in C−N Cross-Coupling Using Machine Learning. *Science* **2018**, *360*, 186−190.
(158) Estrada, J. G.; Ahneman, D. T.; Sheridan, R. P.; Dreher, S. D.; Doyle, A. G. Response to Comment on "Predicting Reaction Performance in C−N Cross-Coupling Using Machine Learning. *Science* **2018**, *362*, eaat8763.
(159) Chuang, K. V.; Keiser, M. J. Comment on "Predicting Reaction Performance in C−N Cross-Coupling Using Machine Learning. *Science* **2018**, *362*, eaat8603.
(160) Shawe-Taylor, J.; Cristianini, N. *Kernel Methods for Pattern Analysis*; Cambridge University Press: New York, NY, USA, 2004.
(161) Artrith, N.; Kolpak, A. M. Grand Canonical Molecular Dynamics Simulations of Cu-Au Nanoalloys in Thermal Equilibrium Using Reactive ANN Potentials. *Comput. Mater. Sci.* **2015**, *110*, 20−28.
(162) Faber, F. A.; Lindmaa, A.; von Lilienfeld, O. A.; Armiento, R. Machine Learning Energies of 2 Million Elpasolite (ABC_2D_6) Crystals. *Phys. Rev. Lett.* **2016**, *117*, 135502.
(163) Kitchin, J. R. Machine Learning in Catalysis. *Nat. Catal.* **2018**, *1*, 230−232.
(164) Musil, F.; De, S.; Yang, J.; Campbell, J. E.; Day, G. M.; Ceriotti, M. Machine Learning for the Structure-Energy-Property Landscapes of Molecular Crystals. *Chem. Sci.* **2018**, *9*, 1289−1300.
(165) King, R. D.; Muggleton, S.; Lewis, R. A.; Sternberg, M. J. Drug Design by Machine Learning: The Use of Inductive Logic Programming to Model the Structure-Activity Relationships of Trimethoprim Analogues Binding to Dihydrofolate Reductase. *Proc. Natl. Acad. Sci. U. S. A.* **1992**, *89*, 11322−11326.
(166) Schütt, K. T.; Glawe, H.; Brockherde, F.; Sanna, A.; Müller, K. R.; Gross, E. K. U. How to Represent Crystal Structures for Machine Learning: Towards Fast Prediction of Electronic Properties. *Phys. Rev. B: Condens. Matter Mater. Phys.* **2014**, *89*, 205118.
(167) Goedecker, S. Minima Hopping: An Efficient Search Method for the Global Minimum of the Potential Energy Surface of Complex Molecular Systems. *J. Chem. Phys.* **2004**, *120*, 9911−9917.
(168) Amsler, M.; Goedecker, S. Crystal Structure Prediction Using the Minima Hopping Method. *J. Chem. Phys.* **2010**, *133*, 224104.
(169) Aspuru-Guzik, A.; Lindh, R.; Reiher, M. The Matter Simulation (R)evolution. *ACS Cent. Sci.* **2018**, *4*, 144−152.
(170) Rangarajan, S.; Maravelias, C. T.; Mavrikakis, M. Sequential-Optimization-Based Framework for Robust Modeling and Design of Heterogeneous Catalytic Systems. *J. Phys. Chem. C* **2017**, *121*, 25847−25863.
(171) Cowan, J. A. Metal Activation of Enzymes in Nucleic Acid Biochemistry. *Chem. Rev.* **1998**, *98*, 1067−1088.
(172) Rakowski DuBois, M.; DuBois, D. L. The Roles of the First and Second Coordination Spheres in the Design of Molecular Catalysts for H_2 Production and Oxidation. *Chem. Soc. Rev.* **2009**, *38*, 62−72.
(173) Shook, R. L.; Borovik, A. S. Role of the Secondary Coordination Sphere in Metal-Mediated Dioxygen Activation. *Inorg. Chem.* **2010**, *49*, 3646−3660.
(174) Shaw, W. J. The Outer-Coordination Sphere: Incorporating Amino Acids and Peptides as Ligands for Homogeneous Catalysts to Mimic Enzyme Function. *Catal. Rev.: Sci. Eng.* **2012**, *54*, 489−550.
(175) Gutmann, V. Solvent Effects on the Reactivities of Organometallic Compounds. *Coord. Chem. Rev.* **1976**, *18*, 225−255.
(176) Klibanov, A. M. Improving Enzymes by Using Them in Organic Solvents. *Nature* **2001**, *409*, 241−246.
(177) Reichardt, C.; Welton, T. *Solvents and Solvent Effects in Organic Chemistry*, 4th ed.; Wiley-VCH: Weinheim, Germany, 2011.
(178) Struebing, H.; Ganase, Z.; Karamertzanis, P. G.; Siougkrou, E.; Haycock, P.; Piccione, P. M.; Armstrong, A.; Galindo, A.; Adjiman, C. S. Quantifying the Chemical Beauty of Drugs. *Nat. Chem.* **2013**, *5*, 952−957.
(179) Ertl, P.; Schuffenhauer, A. Estimation of Synthetic Accessibility Score of Drug-Like Molecules Based on Molecular Complexity and Fragment Contributions. *J. Cheminf.* **2009**, *1*, 8.

DOI: 10.1021/acs.chemrev.8b00759
Chem. Rev. 2019, 119, 6595−6612

(180) Soley, M. B.; Markmann, A.; Batista, V. S. Classical Optimal Control for Energy Minimization Based On Diffeomorphic Modulation under Observable-Response-Preserving Homotopy. *J. Chem. Theory Comput.* **2018**, *14*, 3351−3362.
(181) Rabitz, H. A.; Hsieh, M. M.; Rosenthal, C. M. Quantum Optimally Controlled Transition Landscapes. *Science* **2004**, *303*, 1998−2001.
(182) Rothman, A.; Ho, T.-S.; Rabitz, H. Observable-Preserving Control of Quantum Dynamics over a Family of Related Systems. *Phys. Rev. A: At., Mol., Opt. Phys.* **2005**, *72*, No. 023416.
(183) Rothman, A.; Ho, T.-S.; Rabitz, H. Quantum Observable Homotopy Tracking Control. *J. Chem. Phys.* **2005**, *123*, 134104.
(184) Ho, T.-S.; Rabitz, H. Why Do Effective Quantum Controls Appear Easy to Find? *J. Photochem. Photobiol., A* **2006**, *180*, 226−240.
(185) Hsieh, M.; Wu, R.; Rabitz, H. Topology of the Quantum Control Landscape for Observables. *J. Chem. Phys.* **2009**, *130*, 104109.
(186) Beltrani, V.; Dominy, J.; Ho, T.-S.; Rabitz, H. Exploring the Top and Bottom of the Quantum Control Landscape. *J. Chem. Phys.* **2011**, *134*, 194106.
(187) Bickerton, G. R.; Paolini, G. V.; Besnard, J.; Muresan, S.; Hopkins, A. L. Quantifying the Chemical Beauty of Drugs. *Nat. Chem.* **2012**, *4*, 90−98.
(188) 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*, 268−276.
(189) Lever, A. B. P. Electrochemical Parametrization of Metal Complex Redox Potentials, Using the Ruthenium(III)/Ruthenium-(II) Couple to Generate a Ligand Electrochemical Series. *Inorg. Chem.* **1990**, *29*, 1271−1285.

DOI: 10.1021/acs.chemrev.8b00759
Chem. Rev. 2019, 119, 6595−6612