CHEMICAL REVIEWS

pubs.acs.org/CR

Review

Machine Learning for Chemical Reactions

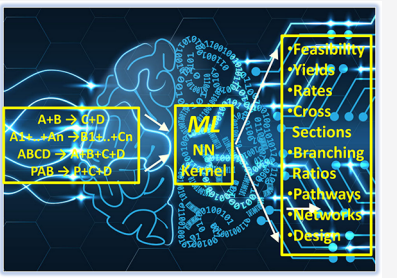
Markus Meuwly*

Cite This: https://doi.org/10.1021/acs.chemrev.1c00033

Read Online

ACCESS | Metrics & More | Article Recommendations

ABSTRACT: Machine learning (ML) techniques applied to chemical reactions have a long history. The present contribution discusses applications ranging from small molecule reaction dynamics to computational platforms for reaction planning. ML-based techniques can be particularly relevant for problems involving both computation and experiments. For one, Bayesian inference is a powerful approach to develop models consistent with knowledge from experiments. Second, ML-based methods can also be used to handle problems that are formally intractable using conventional approaches, such as exhaustive characterization of state-to-state information in reactive collisions. Finally, the explicit simulation of reactive networks as they occur in combustion has become possible using machine-learned neural network potentials. This review provides an overview of the questions that can and have been addressed using machine learning techniques, and an outlook discusses challenges in this diverse and stimulating field. It is concluded that ML applied to chemistry problems as practiced and conceived today has the potential to transform the way with which the field approaches problems involving chemical reactions, in both research and academic teaching.



CONTENTS

1. INTRODUCTION

The prediction of chemical reaction outcomes in terms of products, yields, or reaction rates from computations is a formidable undertaking. Characterizing and understanding chemical transformations is at the heart of chemistry and provides the necessary information about mechanisms and efficiencies of such processes. A quantitative understanding of the speed and efficiency of chemical reactions pertains to, but is not limited to, fields as diverse as atmospheric,[1] combustion,[2] or astrophysical reactions[3] to explosions and enzymatic[4] or organic reactions.[5] It is of interest both to unravel mechanistic underpinnings of known and well-characterized reactions and to predict reaction outcomes based on physically based or physics-inspired models[6] or on high-accuracy PESs.[7−9]

If one is interested in following bond breaking and bond formation in time and space, there are in general two possibilities based on dynamics simulations: those directly and explicitly using quantum mechanical (QM) calculations for the total (electronic) energy and methods based on representations of these energies such as (parametrized) empirical force fields or—as done more recently—based on machine learning. Empirical energy functions represent the potential energy surface (PES) as a computable function for given atomic coordinates whereas

Special Issue: Machine Learning at the Atomic Scale

Received: January 12, 2021

QM-based methods solve the electronic Schrödinger equation directly for every configuration $\vec{x}$ of the system for which it is required. QM-based methods[10] provide the most general framework for investigating the dynamics of chemical reactivity without preconceived molecular structures or reaction mechanisms. However, limitations in direct application of such *ab initio* MD methods[11] are due both (a) to the computational approach *per se* (speed and efficiency) and (b) to practical aspects of quantum chemistry such as the convergence of the Hartree−Fock wave function to the desired electronic state for arbitrary geometries, including corrections due to basis set superposition errors, or the choice of a suitable active space independent of molecular geometry.

For larger systems, such as reactions involving biomolecules, mixed quantum mechanics/molecular mechanics (QM/MM) treatments have become popular.[12] In this approach the system is decomposed into a "reactive region" which is treated with a quantum chemical (or semiempirical) method whereas the environment is described by an empirical force field. This makes the investigation of certain processes feasible that would not be amenable to a full quantum treatment such that even free energy simulations in multiple dimensions can be carried out.[13] One of the current open questions concerns the size of the QM region required for converged results which was recently considered for catechol-*O*-methyltransferase.[14,15] While one study reported that convergence of activation energies required up to one-third of the atoms to be included in the QM part,[14] another study for the same system found that the size of the QM region had a rather moderate effect on activation energies but a somewhat larger influence on the reaction free energy.[15] One possible source for the differences may be the fact that one study carried out umbrella sampling simulations[15] whereas the other did not.[14]

Another approach that has been used for chemically reactive systems is the fragment molecular orbital (FMO) method.[16,17] FMO starts from a partitioning of a molecule into building blocks and assigns electrons to each of the subunits. Next, the molecular orbitals for each fragment and for all fragment pairs are determined from which the total energy of the molecule is obtained from diagonalization of the approximate Hamiltonian. Application to ethanol found equilibrium bond lengths to within 0.04 Å and bond angles to within 5° compared with the reference values from rigorous MO calculations.[16] More recently, FMO has also been applied to chemical reactions, such as for S_N2 reactions in explicit solvent, or for the free energy of binding in a protein−ligand complex involving the Trp cage.[17] Other fragmentation methods[18] are molecular tailoring,[19] systematic fragmentation (STM),[20] the kernel energy method (KEM),[21] or molecular fractionation with conjugate caps (MFCC).[22] It is also possible to run MD simulations with methods such as the force balanced generalized molecular fractionation with conjugate caps (FB-GMFCC) method as has been done for Ace-$(ALA)_9$-NME.[23]

The use of empirical force fields to follow chemical reactions dates back at least 50 years.[24−26] Such an approach has seen various incarnations, including the theory of diatomics in molecules,[24,25] empirical valence bond (EVB)[27] with its extension to several bonding patterns specifically for proton transport in water,[28] the ReaxFF force field,[29] the reactive molecular dynamics force field (RMDff) initially developed for polymers and based on the concept of bond order,[30,31] or adiabatic reactive molecular dynamics (ARMD)[32,33] together with its multistate variants.[34−36] There is a broad range of reviews on the subject of investigating chemical reactions based on established treatments[37] of the potential energy surfaces with applications in gas phase,[38,39] solution,[40,41] and enzymatic reaction[42−46] dynamics, and the technology has also been extended to coarse-grained simulations.[47]

The present work focuses on more recent developments and applications of machine learning techniques applied to problems involving reactions in the gas phase, in solution, and in enzymes. Most problems concerning the *representation* of the underlying potential energy surfaces are excluded, as these are already well covered by accompanying contributions to this special issue.[7,8] Rather, the present work focuses on the application of ML-based techniques to problems that cannot otherwise be exhaustively sampled, on the interplay between experimental observables and computations and how to exploit this for improved understanding of intermolecular interactions, and on dealing with entire reaction networks.

Machine learning (ML) is a data-driven method based on statistical learning theory to generate numerical models that generalize to new data, not used in the learning process.[48] Ideally, ML models interpolate and extrapolate to new data but in general this needs to be verified for every task. Historically, ML can be traced back to work on "Turing's Learning Machine". In 1951 the "Stochastic Neural Analog Reinforcement Calculator" (SNARC) was built by Minsky and Edmonds as a summer research project which is considered one of the first "artificial neural networks" (ANNs). Limitations of the learning capabilities of ANNs were described in "Perceptron",[49] and convolutional (CNN)[50] and recurrent (RNN)[51] NNs were developed subsequently. In the 1970s automatic differentiation was developed[52] which eventually lead to backpropagation that allows us to learn internal representations[53] and was first applied to NNs in the behavioral sciences.[54] With the possibility to compute extensive reference data sets, application of ML-based techniques to questions of chemical reactivity has become a powerful complement to established methods.

One aspect of particular interest in the present work concerns the interplay between observation and information obtained from a computer model. "Observation" in the present context can either be an experimental observation or one from another computation. As an example, the PES can be a parametrized or nonparametric representation of *ab initio* computed energies and the observables can be obtained from either a classical, quasiclassical, or quantum nuclear dynamics simulation such as an inelastic scattering cross section. Conversely, it is also possible to start with a set of experimental observations, for example reaction rates, and aim at determining the corresponding underlying PES that supports these experimental observables.

This review focuses on machine learning for chemical reactions. First, ML to generate models for or use results of experimental observables is discussed. This is followed by the computation of reaction rates and pathways from reactive potential energy surfaces. The generation of such surfaces also benefits from ML techniques such as transfer learning and Δ-learning which are discussed there. In section 4 the application of ML to reactions involving enzymes and their evolution is discussed. Next, ML in the context of organic reactions, specifically retrosynthesis, and for unimolecular decomposition reactions using mass spectrometry as the detection technique is reviewed. Finally, an overview of the application of machine learning techniques to entire reaction

https://doi.org/10.1021/acs.chemrev.1c00033
Chem. Rev. XXXX, XXX, XXX−XXX

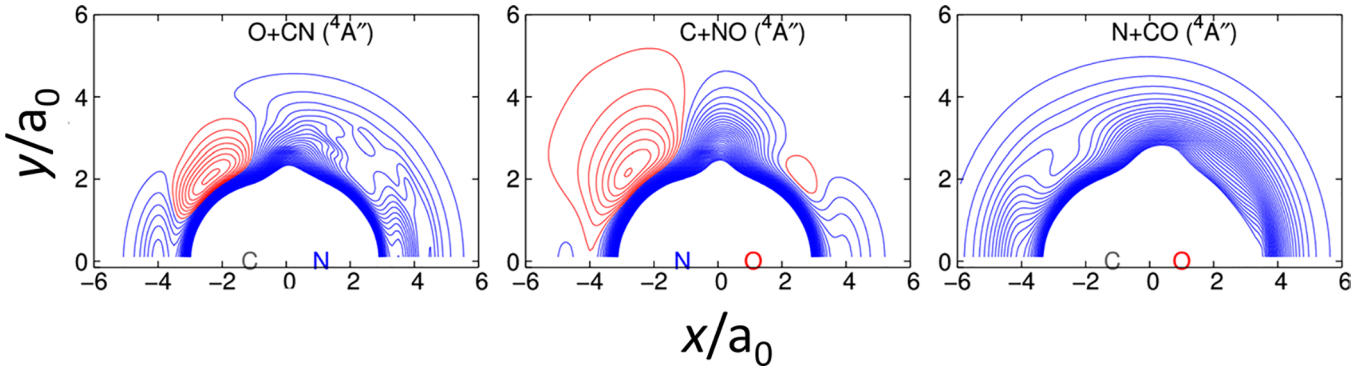


Figure 1. Contour plots of the $^4A''$ PESs for the [CNO] system represented as a reproducing kernel Hilbert space.[74,75] The diatoms (CN, NO, and CO from left to right) are at their equilibrium structures (2.234, 2.192, and 2.150 a_0, respectively). The spacing between the isocontours is 0.2 eV with red lines corresponding to negative energies (−0.1, −0.3, −0.5 eV and lower) and blue lines to positive energies (0.1, 0.3, 0.5 eV and higher). Adapted with permission from ref 73. Copyright 2018 American Chemical Society.

networks is given. The manuscript closes with possible future developments, including visualization and data archiving.

2. MACHINE LEARNING OF REACTION OBSERVABLES

Machine learning techniques can be used to develop comprehensive models for observables that are directly related to experiments. Examples include prediction of thermal[55−57] or quantum reaction rates,[57] state-to-state cross sections,[58] mapping initial to final state distributions,[59] or even chemical reaction yields.[60]

2.1. Learning Reaction Rates

Chemical reactions involve bond-breaking and bond-formation processes. Hence, if the explicit nuclear dynamics describing the transition between reactant and one or several possible products is of interest, the motion between the two involves transgressing a barrier on the multidimensional PES. Empirical force fields, originally based on experimental information such as structure, spectroscopy, and thermodynamics,[61−64] are not suitable for following chemical reactions as the bonding pattern between the atoms is fixed. With the advent of efficient electronic structure methods, attention has shifted either to complement existing parametrizations with information from (high-level) electronic structure data[6,65−69] or to develop models that are entirely based on quantum chemical calculations.[70]

From a microscopic perspective suitable methods to follow chemical transformations—e.g. quasiclassical trajectory or quantum simulations—at a molecular length scale are sensitive to the *entire* PES and require a global, reactive PES. Representing such a PES can be very challenging even for triatomic systems[71,72] because their topographies can be rather complex.[73] An example for such a PES for the [CNO] reactive system is reported in Figure 1. These 3-dimensional PESs were represented as a reproducing kernel Hilbert space which exactly matches the reference *ab initio* calculations on the reference points.[74,75] Finding parametrized functions with similar performance is in general very challenging. An alternative is permutationally invariant polynomials.[76]

The determination of accurate reaction rates from computations is a formidable task even for seemingly simple A + BC → AB + C atom exchange reactions[72] such as the O + CN → [C + NO, N + CO] reaction for which PESs are illustrated in Figure 1. Modern approaches for gas phase reactions are based on (a) the computation of thousands of energies at a high level of quantum chemical theory (e.g., coupled cluster (CCSD(T)) or multireference (MRCI) treatments) using large basis sets, (b) the representation of these energies either as a parametrized function[76] or using machine learning techniques such as neural networks,[7,77−80] reproducing kernel Hilbert space methods,[74,75,81,82] and (c) following the nuclear dynamics either using classical mechanics (quasiclassical trajectory calculations (QCT))[26,83] or solving the three-dimensional nuclear Schrödinger equation.[84,85] Such an approach is overall very computationally demanding, and the quality of the computed rates depends sensitively on the accuracy of the underlying PES.[86]

In an effort to alleviate this problem, a recent effort[55] explored the possibility to use Gaussian process (GP) regression to train a correction $\chi(T)$ to predict thermal rates $k(T)$ in

$$k(T) = [\kappa_{\mathrm{ECK}}(T)k^{\mathrm{TST}}(T)]\chi(T) \quad (1)$$

Here, $k(T)$ was determined for ∼50 different reactions based primarily on collinear collisions—which is a simplification but does not limit the general applicability of the approach—and the rates $\kappa_{\mathrm{ECK}}(T)$ and $k^{\mathrm{TST}}(T)$ are those from a simplified treatment of the Eckart tunneling correction (ECK) to conventional transition state theory (TST). Training was done for 13 reactions, and the model was tested on 40 reactions. Between 3 and 5 descriptors were chosen to represent $\chi(T)$, and it was reported that judicious choice of the descriptors can lead to marked improvements in the model performance.

Compared with either TST or the Eckart treatment, the machine-learned model performed best on the training set, on systems with symmetrical and asymmetrical barriers. In all cases, the error of the learned model ranged from 10% to 120% when compared with the exact data. This is a marked improvement over errors ranging from 80% to 180% (for ECK) and 180% to 760% (for TST). For reactions involving hydrogen abstraction from CH_4, the Eckart model was best, followed by the GP-trained model and TST. This study also highlights a critical need for both the quality and quantity of reliable reference data in training such models. One surprising finding is the fact that using a model trained on collinear reactions ("2d") but applying to a data set for reactions in full dimensionality ("3d") can perform quite well.

A recent application of this model concerned the $O(^3P)$ + HCl → OH + Cl reaction[56] which can be considered as a particularly challenging example due to the large reaction barrier, the presence of low-energy reactive resonances, and the heavy−light−heavy character of the system. Future applications of such an approach may be possible on sufficiently extensive

https://doi.org/10.1021/acs.chemrev.1c00033
Chem. Rev. XXXX, XXX, XXX−XXX

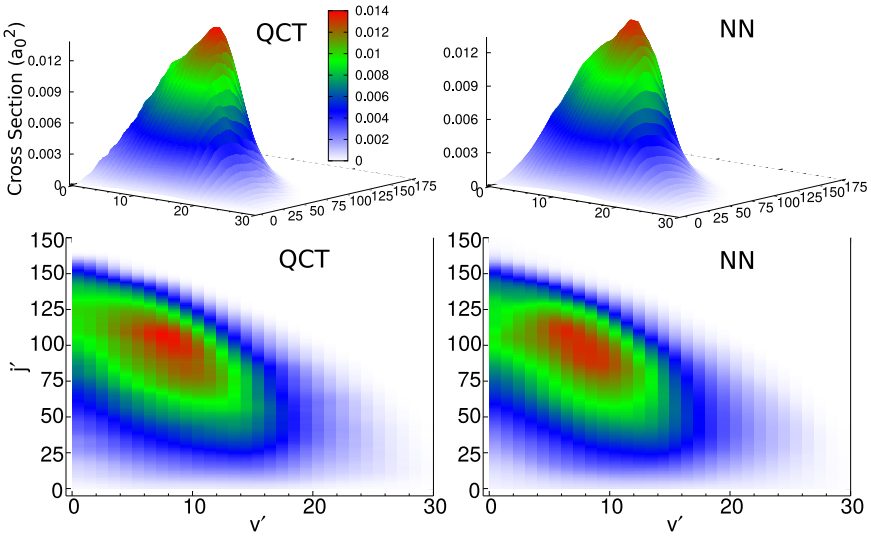


Figure 2. QCT-calculated (left) and STS-predicted (right) state-to-state cross sections for the N + NO($v = 6, j = 30$) → O + $N_2(v', j')$ reaction at E_t = 2.5 eV. The top row reports a 3D surface and the bottom row a contour color map of the data. Adapted with permission from ref 58. Copyright 2019 American Chemical Society.

and curated data sets from actual experiments. As the reference data originated from a broad range of PESs it would also be of interest to determine whether model prediction can be further improved if all PESs are based on the same or a similar level of theory.

Machine learning approaches were also extended to one-dimensional quantum reaction rates.[57] Based on a large number of quantum rates $k_Q(T)$, a deep neural network (DNN) was trained. The potential energy surfaces considered include single and double barriers and symmetric and asymmetric shapes. The optimized DNN, trained on ∼1.5 million data points, was then applied to predicted rates for a range of gas phase and surface reactions. The overall accuracy on the test set for log $k_Q(T)$ was 1.1%, and even at temperatures below 300 K for which tunneling effects are expected to become important, the relative error was only ∼30%.

2.2. State-to-State Models and Rates

Maintaining the full dimensionality of the problem, a NN-based, machine-learned model was developed for the state-to-state (STS) cross sections of the $N(^4S) + NO(^2\Pi) \rightarrow O(^3P) + N_2(X^1\Sigma_g^+)$ reaction.[58] This and other atom + diatom reactions ($O(^3P) + NO(^2\Pi) \rightarrow O_2(X^3\Sigma_g^-) + N(^4S)$,[87−89] $C(^3P) + O_2(^3\Sigma_g^-) \leftrightarrow CO_2 \leftrightarrow CO(^1\Sigma^+) + O(^1D)/O(^3P)$,[90,91] and $C(^3P) + NO(X^2\Pi) \rightarrow O(^3P) + CN(X^2\Sigma^+)$, $N(^2D)/N(^4S) + CO(X^1\Sigma^+)$[73]) are relevant in the atmosphere, in combustion, and for hypersonic flight.[72,92] For the $^3A'$ state the total state space of the $N(^4S) + NO(^2\Pi) \rightarrow O(^3P) + N_2(X^1\Sigma_g^+)$ reaction involves a maximum of 47 and 57 vibrational states for NO and N_2, and the maximum rotational quantum numbers for NO and N_2 are j = 241 and 273, respectively. Overall, there are 6329 an 8733 rovibrational states for the N + NO and O + N_2 channels, respectively. To converge one specific state-to-state cross section $\sigma_{v,j\rightarrow v'j'}(E_t)$ for given translational energy E_t, typically 10^4 to 10^5 QCT simulations need to be run. Doing this for the $\sim 10^4$ initial states going into all $\sim 10^4$ final states requires an estimated 10^{12} to 10^{13} QCT simulations for directly sampling all available rovibrational states which is computationally impractical. For diatom−diatom collisions the number of transitions increases to $\sim 10^{15}$ and the necessary number of QCT simulations approaches 10^{20}.[93] Instead of using explicit QCT simulations to generate the necessary data for reaction network modeling, a ML-based state-to-state (STS) approach using a neural network was developed to map initial to final states.[58]

2.2.1. State-to-State Model. The NN architecture for this application was based on ResNet[94] which uses identity shortcut connections to alleviate the vanishing gradient problem[95] that slows down the learning capacity with increasing depth of the NN. The input is transformed through four identical residual blocks,[94] followed by a linear transformation using a scaled sigmoid function to project onto the final output.[58] The weight matrices $\mathbf{W} \in \mathbb{R}_{F\times F}$ (weight matrix) and $\mathbf{b} \in \mathbb{R}_F$ (bias vector) contain the parameters to be optimized. The residual blocks consist of two dense layers with the same number of nodes and transform their input $\mathbf{x}^l$ according to

$$\mathbf{x}^{l+2} = \mathbf{x}^l + \mathrm{ReLU}[\mathbf{W}^{l+1}\mathrm{snasinh}(\mathbf{W}^l\mathbf{x}^l + \mathbf{b}^l) + \mathbf{b}^{l+1}] \quad (2)$$

for layer l. Two different activation functions (a rectified linear unit (ReLU)[96] and a self-normalizing[97] inverse hyperbolic sine (snasinh)[98]) are used in the residual blocks. The final output is obtained from

$$y = C \times \mathrm{sig}(\mathbf{W}^o\mathbf{x}^l + b^o) \quad (3)$$

where C = 0.4 is a scaling constant and $\mathrm{sig}(x) = (1 + e^{-x})^{-1}$ is the sigmoid function. Superscripts o and l refer to the "output" and "last" hidden layer.

To train such a network, features $\mathbf{f}$ such as the internal energy, vibrational and rotational quantum numbers, or the relative velocity of the reactants need to be chosen. Overall, 12 such features are chosen in this STS ML-based model for final state prediction.[58] To determine the progress of the optimization, a loss function (L_f) is required

$$L_f = \frac{1}{N}\sum_1^N [\log(y' + 1.0) - \log(y' + |y - y'| + 1.0)]^2 \quad (4)$$

where y' and y are the reference (QCT) and predicted values (NN), respectively. All parameters of the NN are initialized according to the Glorot initialization scheme[95] and optimized using Adam optimization.[99]

https://doi.org/10.1021/acs.chemrev.1c00033
Chem. Rev. XXXX, XXX, XXX−XXX

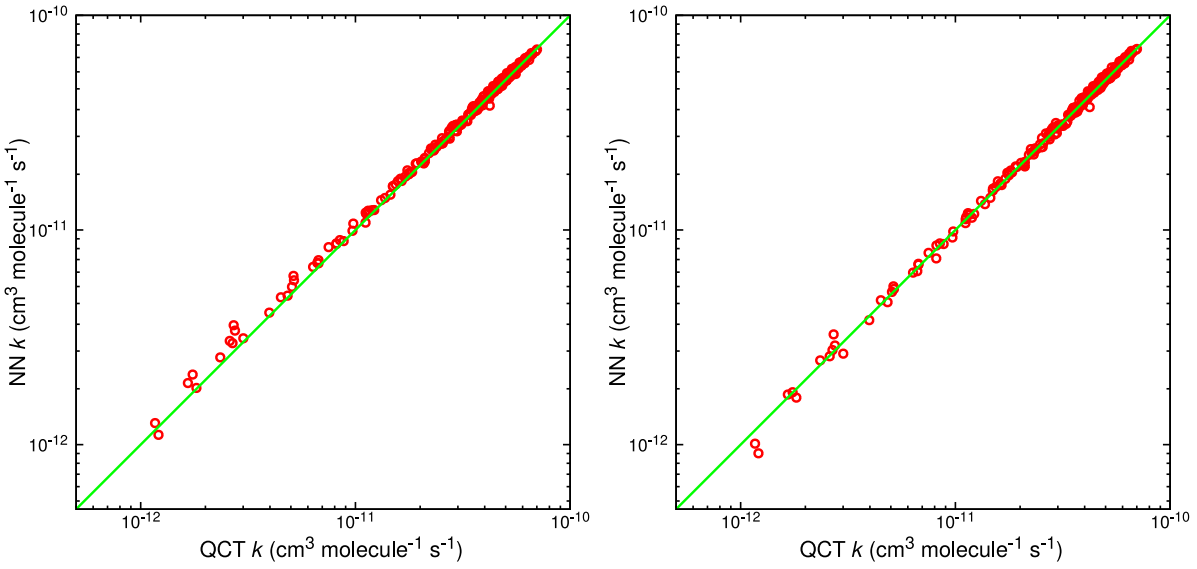


Figure 3. Correlation between the QCT-calculated and NN-predicted initial state selected rates. Left panel: Initial state selected QCT vs NN-STS rates for $v = 5$, 10, 15, and 20, $j = 20$, 25, 40, 60, 85, and 110, and at $T = 2000$, 4000, 6000, 8000, 10000, 12000, 14000, 16000, and 18000 K. Right panel: Total QCT vs NN-Tot rates at $T = 1000$, 2000, 3000, 4000, 5000, 8000, 10000, 12000, 15000, 18000, and 20000 K. Diagonals are shown as solid lines. Adapted with permission from ref 58. Copyright 2019 American Chemical Society.

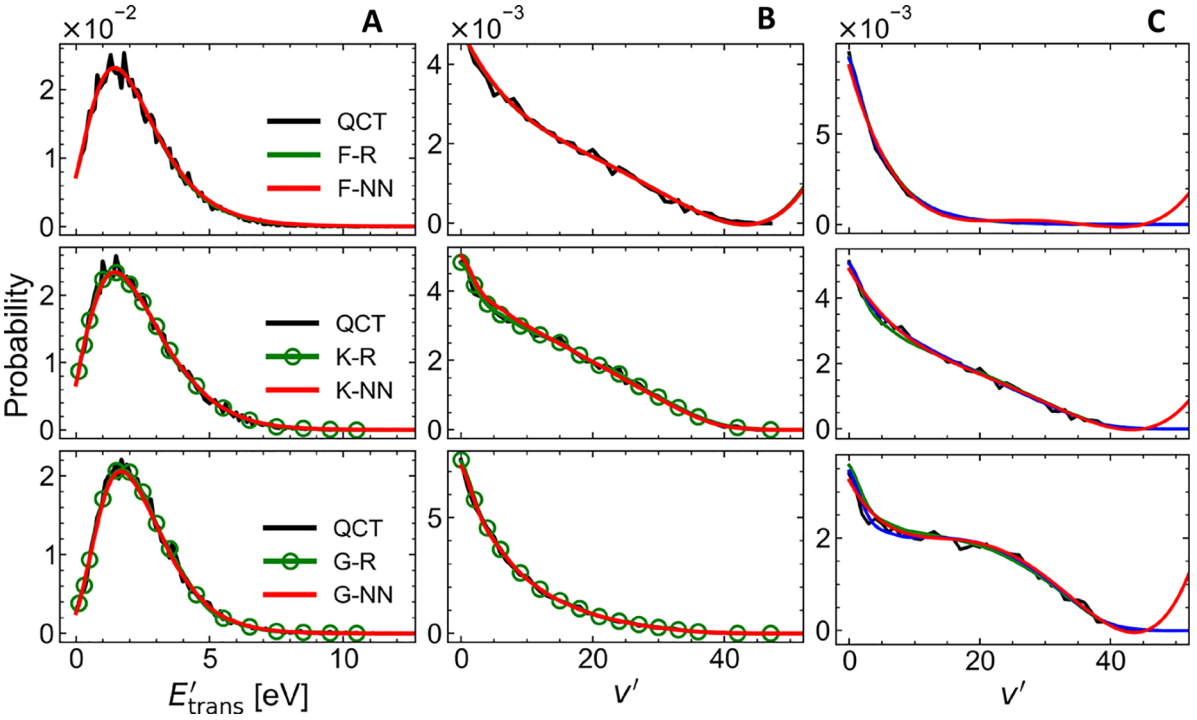


Figure 4. Product state distributions from explicit QCT simulations (QCT, black traces) as well as the corresponding reference data (-R) and the predictions (-NN) from fitting to a funtion (F, top), kernel (K, middle), and grid (G, bottom) representation. Also, the amplitudes to construct the reference RKHS-based representations for (F,K,G) approaches are reported (circles). Panels A and B: QCT simulations with initial conditions sampled at (T_{trans}, T_{vib}, and T_{rot}) for models representative of the average performance of the NN for each of the representations. These temperatures are (9500 K, 16000 K, 16000 K), (10250 K, 19250 K, 19250 K), and (12000 K, 9750 K, 9750 K) and [$RMSD_{NN} = 0.0005$, $R^2_{NN} = 0.9995$], [$RMSD_{NN} = 0.0013$, $R^2_{NN} = 0.9984$], and [$RMSD_{NN} = 0.0009$, $R^2_{NN} = 0.9993$] for F-, K-, and G-based representations, respectively. Panel C: the final vibrational distributions for initial conditions at temperatures (12500 K, 5750 K, 5750 K), (9500 K, 16000 K, 16000 K), and (5750 K, 19250 K, 19250 K) for the top, middle, and bottom panels. Toward the highest values v' the F-based approach (red) is unable to correctly model $P(v')$. Note also the stark contrast to a typical Boltzmann distribution for the vibrational state distribution for the middle and bottom panels in C. Adapted with permission from ref 58. Copyright 2019 American Chemical Society.

To independently test the final NN, additional QCT calculations at fixed E_t were performed for initial conditions not used in the training. Comparison of the explicitly determined QCT cross sections with those predicted by the NN (Figure 2) demonstrates that such an approach is viable and provides an accurate and computationally advantageous

https://doi.org/10.1021/acs.chemrev.1c00033
Chem. Rev. XXXX, XXX, XXX−XXX

alternative for obtaining specific, molecular-level information for such reactions.

To further validate the model, initial state-selected rates were explicitly determined from QCT simulations and compared with predictions from the NN STS model; see Figure 3. Although maximum relative errors of >15% can occur for particular initial states, in most cases the relative errors are <5%. In general, state-specific and temperature-dependent total reaction rates from the NN are in quantitative agreement with explicit QCT simulations.

2.2.2. Distribution-to-Distribution Model. In a recent extension, the STS model was generalized to a distribution-to-distribution model (DTD).[59] For this, the NN-based model was trained to predict final state distributions ($P(E_{trans}')$, $P(v')$, $P(j')$) given reactant state distributions $P(E_{trans})$, $P(v)$, and $P(j)$. Here, $P(v)$ and $P(j)$ are marginal distributions, i.e. $P(v) = \sum_j P(v,j)$ and $P(j) = \sum_v P(v,j)$. Working with the underlying distributions leads to considerably smaller NNs that need to be trained which also speeds up the learning process. For this task, a multilayer perceptron with two hidden layers was used with 10 to 40 input and output nodes depending on the representation of the distributions. The two hidden layers contain between 6 and 12 nodes each.

For representing the distributions, parametrized functions (F), ML-representations based on reproducing kernels (K), and the actual grid points (G) were considered. In general, all three approaches accurately describe the (equilibrium Boltzmann) reactant state distributions. However, the product states are nonequilibrium distributions as they are from high-temperature simulations, ranging from 2000 to 20000 K. Figure 4 reports final translational and vibrational distributions $P(E_{trans}')$ and $P(v')$ from different simulations compared with the machine-learned predictions. For Figures 4A and B the QCT simulations were from conditions representative of the average R^2 over all test data for the respective method, i.e. using F-, K-, or G-based representations of the distributions. For conditions most representative of the average performance, an F-based approach is somewhat better suited (R^2 = 0.999) than a K-based (R^2 = 0.998) or a G-based (R^2 = 0.999) approach.

The K- and G-based approaches reproduce the QCT data very closely; see blue and green traces in Figure 4C. Fitting the product state distributions to parametrized functions leads to differences, in particular for $P(v')$. These manifest themselves as deviations for small and high v' or extra undulations in Figure 4C. However, because state space is finite ($v'_{max} = 47$, $j'_{max} = 240$), differences for high v' are only partially relevant.

2.3. Gaussian Processes and Bayesian Inference

Gaussian process (GP) regression is a nonparametric, supervised learning technique and one of several kernel-based methods to generate ML models.[100] Previously, GP has been applied to regression and classification problems, but more recently it has also been used to represent intermolecular PESs[71,101] and, combined with Bayesian optimization, to address the inverse problem of reactive scattering.[102] The "inverse problem" of determining the interaction potential from scattering data is a long-standing problem in chemical physics[103−107] and has been handled within Tikhonov regularization and using minimization procedures for diatomic molecules.[108]

The problem has been formulated in a combined GP and Bayesian inference context.[109] The formal development starts from a global, reactive PES expressed as $V(\vec{r}) = \sum_{i=1}^{N} w_i(\vec{r}) E_i$, similar to MS-ARMD,[34] where E_i are energies from an *ab initio* calculation and the weights $w_i(\vec{r})$ are optimized through GP regression to yield the most accurate scattering cross sections. A GP is entirely specified by the conditional mean $\mu(\vec{r}_i)$ and the conditional variances $\sigma(\vec{r}_0)$ at an arbitrary configuration $\vec{r}_0$. The conditional mean and variances can be represented as an n-dimensional vector and a $n \times n$ square matrix of covariances, respectively.[102] The GP model is trained by determining the best covariance matrix given a set of reference data (e.g., energies from *ab initio* calculations or scattering cross sections). In order to progress, a model for the covariances needs to be assumed which can be Matern functions, Gaussians, rational quadratic kernels, or other simple functions.[100] The model is then optimized using a suitable likelihood function which is the log marginal likelihood as is customary for GP optimization. It is noted that up to this point such an approach is also reminiscent of the reproducing kernel Hilbert space technique which employs different classes of functions for representing the kernel matrix and which has been successfully used to represent PESs.[74,75,81,82,110−120]

To make an inference (prediction) of an unknown value y_* at a given value $\vec{r}_*$, one can use Bayes' theorem according to which

$$P(y_*|\vec{y}) = \int_\theta P(y_*|\theta) P(\theta|\vec{y}) d\theta \tag{5}$$

where θ is a vector containing the model parameters and $\vec{y}$ are the known values (e.g., energies from *ab initio* calculations or scattering cross sections).[102] If the model parameters θ are not fixed but random variables themselves, a Bayesian NN is obtained. More specifically, if Gaussian distributions are assumed for the parameters and in the limes of an infinite number of neurons, the Bayesian NN has been shown to map onto a GP.[102]

This formalism has been recently applied to reactive scattering for the H + H_2 → H_2 + H and the OH + H_2 → H_2O + H reactions. It was reported that for the H + H_2 → H_2 + H reaction with as few as 37 points (taken from a total of 8701 reference energies from which a global PES had been fitted previously) accurate reaction probabilities can be obtained.[109] Similarly, for the more challenging OH + H_2 → H_2O + H reaction 290 points from a total of ∼17000 *ab initio* energies were sufficient to cover the entire 6-dimensional PES to obtain accurate reaction probabilities as a function of the translational energy of the reactants.[109] While such an approach still relies on available rates determined from a separate calculation, recent progress in GP with Bayesian inference has become largely independent of such reference data. The approach is then to determine the "Bayesian Information Criterion" (BIC) and select the distribution of points that maximizes the BIC.[121]

It is worthwhile to mention that with increasing dimensionality GP-based representations become computationally expensive.[86] On the other hand, increasing the kernel complexity, e.g. by using composite kernels, the performance[122,123] and accuracy[121,124] of GP-based representations have been found to improve.

2.4. Comparison of ML-Based Methods with Established PES-Fitting Approaches

An alternative to ML-based methods to represent high-dimensional PESs is fitting to a predefined and parametrized functional form. An early example is the London−Eyring−Polanyi−Sato (LEPS) surface[125−127] that was also used to study the H + H_2 reaction.[26] With the advent of sufficiently efficient

https://doi.org/10.1021/acs.chemrev.1c00033
Chem. Rev. XXXX, XXX, XXX−XXX

and accurate electronic structure methods, the configurational space could be more broadly covered by computing energies for many geometries. Hence, the problem to be solved shifted to representing this potentially large number of reference calculations. Currently, nonreactive and reactive molecular PESs based on $\sim 10^5$ high-level reference calculations are available,[128,129] and the challenge is to represent those with an accuracy comparable to that of the underlying quantum chemical calculations.

In the following, fitting PESs for triatomic systems, in particular van der Waals complexes, is considered. Expansion of the total potential into products of Legendre polynomials $P_\lambda(\cos\theta)$ and corresponding radial strength functions $V_\lambda(R, r)$, i.e. $V(R, r, \theta) = \sum_{\lambda=0}^{\lambda_{max}} V_\lambda(R, r)P_\lambda(\cos\theta)$ has been found to be advantageous.[130−133] Such a formulation lends itself to both fitting the parameters to experimental data or to reference electronic structure calculations. Alternatively, the total interaction energy can be decomposed into long- and short-range parts and to represent them separately. This can be advantageous as the long-range part (multipolar electrostatic interactions and polarizabilities) of the interaction depends primarily on monomer properties which are amenable to or available from experiments.[132,134] Alternatively, more conventional parametrized fits can be carried out, e.g. using modified Shepard interpolation,[135−137] moving least-squares,[138−140] or permutation invariant polynomials.[141−143]

In most applications fitting the parameters is a cumbersome, at best semiautomated process even for the most sophisticated functional forms. The nonlinear least-squares fits do not easily converge, and often considerable human intervention is required. This is why more automated approaches such as representing data as a RKHS[75] or as a NN[7,79] have emerged as an attractive alternative. In all these fitting exercises, the final PESs need to be validated with respect to their extrapolation behaviors to short- and long-range and for potential artifacts, including spurious minima or undulations.

3. REACTION RATES AND PATHWAYS

Reaction rates can be determined from classical or quantum dynamics simulations if suitable PESs are available that allow bond formation and bond breaking. Ideally, such PESs are *global*; that is, they allow—starting from a reactant structure—formation of all chemically meaningful and energetically accessible product states. In practice, generating such global PESs is extremely challenging[39,144] or even impossible due to the large number of reaction pathways. The global nature of the PES is particularly important in high-energy processes such as hypersonics or in combustion. While for hypersonics[72,92,145−149] the relevant species are often atoms and diatomics, this is not the case for combustion[2] or for atmospheric and astrophysically relevant[1,3,150,151] processes for which the species involved can be larger and the number of possible product channels therefore increases considerably.

To illustrate the problem for following a chemical reaction of an atmospherically relevant molecule, isomerization and decomposition pathways for acetaldehyde (AA) are considered. These processes are relevant for atmospheric chemistry because it has been proposed that formic acid (FA) can be generated via oxidation by the hydroxyl radical[152,153] following photo-tautomerization of AA to its enol form VA.[154−156] Pathways in addition to the conventional route (photochemical oxidation of biogenic and anthropogenic volatile organic compounds (VOCs)) for formation of FA are required to account for the global budget of formic acid.[157]

To characterize the isomerization between AA and VA under conditions relevant to the atmosphere, a NN-based, reactive PES was constructed[129] based on PhysNet.[79] Such a computationally efficient PES is required for running statistically significant numbers of trajectories[34,158] because *ab initio* MD simulations are computationally too expensive. The excitation energy in the simulations was 93.6 kcal/mol, to be compared with energies of 86.6−95.3 kcal/mol for actinic photons. At this excitation energy no isomerization reaction from AA to VA was observed but decomposition into CH_4 + CO and H_2 + H_2C_2O occurred. It was found that for an accurate representation of all states involved (AA, VA, CH_4 + CO and H_2 + H_2C_2O) and for stable *NVE* simulations, more than 4×10^5 reference energies at the MP2 level of theory with an aug-cc-pVTZ basis set were required.[129] For energies up to 93.6 kcal/mol (isomerization barrier between AA and VA at 68 kcal/mol and excitation energy by actinic photons) the MAE and RMSE are 0.0071 and 0.0145 kcal/mol, respectively. In order to validate that the NN-PES does allow isomerization, higher excitation energies up to 127.6 kcal/mol were used. The global PES has then a MAE and an RMSE of 0.0132 and 0.0307 kcal/mol, respectively.

For excitation energies of ∼95 kcal/mol, i.e. the energy available in actinic photons, not a single isomerization between AA and VA occurred on the 500 ns time scale. Hence, it is unlikely that FA is generated following electronic excitation of AA with actinic photons and subsequent ground state relaxation and isomerization to VA (and/or further chemical processing) because collisional deexcitation is known to occur on the 100 ns time scale. This is in contrast with the interpretation of the experiments.[152] Rather, after photoexcitation of AA the system either decomposes or relaxes through internal vibrational energy distribution.

Another ML-based approach that was recently followed to construct reactive PESs and use them in dynamics simulations is based on permutationally invariant polynomials (PIP)[76,159] combined with a neural network (PIP-NN).[160] For conventional PIP, the expansion coefficients in the polynomials (usually in Morse-variables) are fitted using a linear least-squares algorithm whereas for PIP-NN the coefficients are trained by a NN. PIP-NN has been applied to both gas phase and surface reactions. For the photodissociation of acetaldehyde into CH_4 + CO, a PES based on PIP[128] was used to confirm that the dominant decomposition pathway involves roaming.[161] A subsequent study extended the number of accessible product states considerably and reported fragmentation for eight different pathways together with branching ratios.[162]

In the gas phase, PIP-NN has been applied[163] to reactions such as HO + CO → H + CO_2 which is relevant in the atmosphere and in combustion.[164] A total of ∼75000 was used to represent this channel. The PES was used in QCT and quantum dynamics simulations to determine total reaction probabilities, thermal rates, differential cross sections, and product state vibrational and rotational distributions[163] as well as tunneling probabilities and survival fractions.[165] Comparison between a pure NN, a pure PIP, and the PIP-NN approaches demonstrates that despite the rather small differences in the fitting quality certain observables, such as product state distributions or differential cross sections, can sensitively depend on the shape of the PES.

PIP-NN has also been used for reactive scattering involving metal surfaces. Systems investigated include H_2/Ag(111),[166]

https://doi.org/10.1021/acs.chemrev.1c00033
Chem. Rev. XXXX, XXX, XXX−XXX

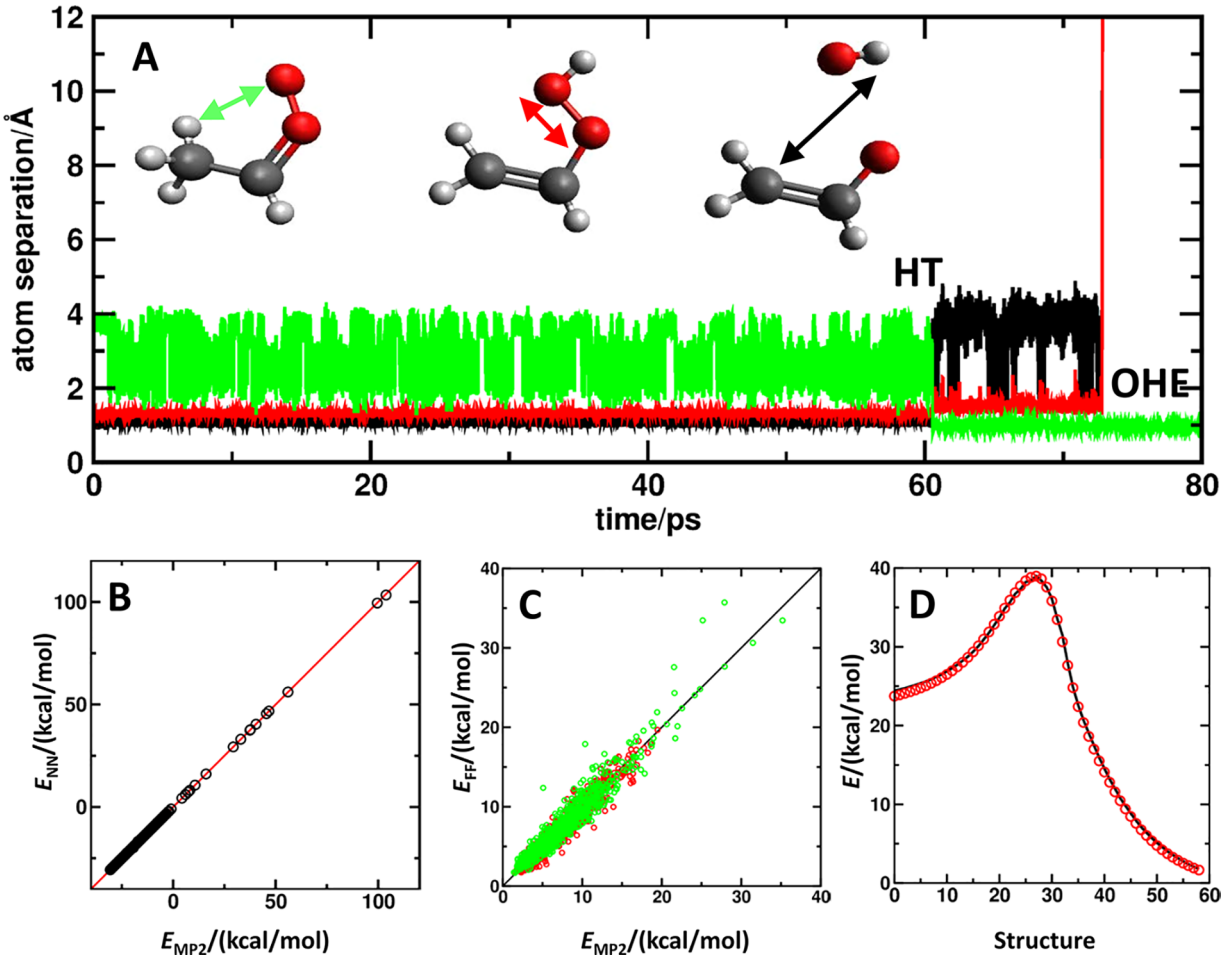


Figure 5. Stepwise reactions for the Criegee intermediate. Panel A: Time dependence of important atom separations, the CH (black), OO (red), and OH (green), for the H-transfer (HT) and OH-elimination (OHE) reaction from MS-ARMD simulations. Panel B: PhysNet[79] model for ~10^5 MP2 reference energies together with the correlation between them (red). Panel C: MS-ARMD[34] model for ~10^3 reference structures for reactant (red) and product (green) structures. The RMSD is ~1 kcal/mol, and the 1:1 correlation is the black line. Note the different energy scales in panels B and C. Panel D: minimum energy path for H-transfer between reference (red circles) and MS-ARMD fit (black).

H_2/Co(0001),[167] H_2O/Ni(111),[168] and CO_2/Ni(100).[169] The number of reference points in these applications ranged from several 1000 to ~25000. For the dissociative chemisorption of water on rigid Ni(111), QCT simulations using a nine-dimensional PIP-NN PES fitted to energies from density functional theory, it was found that the reactivity depends on the impact sites and the incident angle of the water molecule.[168] Furthermore, analysis of the simulations demonstrated that both the barrier height and the topography of the PES influence the reaction rate as does the translational energy both parallel and perpendicular to the surface.

More recently, PIP-NN has also been extended to larger systems, including the F+ $CH_3OH \rightarrow HF + CH_3O$ reaction[170] or the investigation of the Cl + $CH_3OH \rightarrow HCl + CH_3O$/$CH_2OH$ reaction.[171] These examples illustrate that it is possible to follow reactions with multiple product states in a realistic fashion. For instance, the HCl vibrational and rotational product distributions and the product translational energy distributions compare well with experiment.

Complementary to reactions with multiple reaction products, multiple step reactions with one or several intermediates between reactant and product pose another challenge. A recent application concerned the thermal activation of methane by MgO^+ for which experimental rates were determined between 300 and 600 K.[172] Another example is the unimolecular decomposition of the CH_3COOH Criegee intermediate.[173] Recent simulations based on an MS-ARMD and NN-trained full-dimensional energy surface involving the reactant, H-transferred intermediate, and OH-elimination product (see Figure 5) demonstrate that stepwise reactions can also be followed by such techniques. Figure 5 demonstrates that empirical FFs can be fit with an accuracy of ~1 kcal/mol ("chemical accuracy") whereas using PhysNet to train the same data reaches an accuracy of 0.02 kcal/mol.

It should be emphasized that despite the undoubted accuracy reactive PESs can reach and their utility for interpreting experiments and further elucidating the reaction dynamics of the systems, these are not in general truly "global" PESs. Typically, the research decides what channels are considered or of interest and the corresponding breakup channels are included in the ML-based construction of the interpolant. Thus, for systems with more degrees of freedom and a larger number of product channels, it is possible that important states are omitted despite the powerful methods available for (re)constructing the underlying PESs. If omitted, such channels will fundamentally affect the resulting reaction network and the realism with which one is able to map it to a computational model and provide a meaningful complement to experiment.

https://doi.org/10.1021/acs.chemrev.1c00033
Chem. Rev. XXXX, XXX, XXX−XXX

One particularly attractive possibility ML-based methods offer is to conceive models at one level of electronic structure theory and retrain them with considerably fewer data from a higher level of theory. Such approaches capitalize on the fact that the global shape of a PES for a given system remains largely unchanged if sufficiently accurate calculations have been carried out. These methods are often referred to as "transfer learning" or "Δ-learning", both of which also bear striking similarities with the "morphing" approach[133] which has also been discussed in the context of global improvements of full-dimensional PESs.[174]

Δ-learning can be formalized as

$$O_t(\vec{R}_t) \approx O_r(\vec{R}_r) + \sum_{i=1}^{N} \alpha_i k(\vec{R}_r, \vec{R}_i) \tag{6}$$

where O_r and O_t are a reference and target observable (e.g., "energy" and "enthalpy"), $\vec{R}_r$ and $\vec{R}_t$ are two different geometries at which O_r is known and O_t is sought, and the sum is an ML-learned model based on a distance kernel $k(\vec{R}_r,\vec{R}_i)$.[175] Such an ansatz has been used to learn total energies E_t from the (cheap) PM7 level of theory to the (expensive) G4MP2 model. Interestingly, it was found that all Δ-ML models outperform an ML model trained on the absolute value of E_t directly, without a baseline. Δ-ML has also been used to generate CCSD(T)-quality PESs for small (H_3O^+, CH_4) and intermediate sized (*N*-methyl-acetamide) molecules from PIP-representations based on DFT-reference data.[176] Finally, Δ-ML has also been applied to include many-body effects into molecular PESs.[177]

In transfer learning a full model (based e.g. on a NN, using PIP, or a kernel-based method) is first trained on a reference data set conceived at a computationally inexpensive level of theory, such as DFT. If this data is represented as a NN, the parameters of the trained model are used to initialize training with a much smaller (e.g., 100 times less) number of reference data from much higher-level calculations (e.g., CCSD(T)). This usually leads to better models with a smaller number of high-level data. Such an approach is advantageous in situations where high-level reference data is scarce or expensive to generate, whereas data for training the base model is readily available.[178] Hence, TL is a valuable tool to avoid the high computational cost of modern high-level electronic structure calculations for the full data set.[179] In a study of malonaldehyde and its methylated variants,[80] it was found that CCSD(T)-quality PESs can be obtained for all three systems from a base model at the MP2 level of theory by transfer learning together with energies from a higher level of theory. However, when querying the higher-level PESs for *specific* information, e.g. the barrier height for hydrogen transfer, it was found that depending on the position of the higher-level reference points improvement is not always substantial. Hence, strategic positioning of the new reference points can lead to much improved higher-level models when developing PESs for specific purposes. As for Δ-ML it was also found that the learning curves for the transfer-learned models converged more rapidly than directly training independent models on the high-level data.

4. ML APPLICATIONS TO REACTIVE BIOLOGICAL SYSTEMS

4.1. Reactive Molecular Dynamics for Ligand Binding to Proteins

Proteins are too large and the time scales of bond-breaking and bond-forming reactions too long to be amenable to full *ab initio* MD simulations. Hence, mixed QM/MM simulations which decompose the system into a (small) reactive subsystem treated with quantum mechanical methods and an environment described at the level of an empirical force field are one of the methods of choice.[10,12] Alternatively, methods such as the empirical valence bond (EVB) theory[27,42] or multistate adiabatic reactive MD[34] have been developed and applied. More recently, ML-based energy functions such as reproducing kernels (RKHSs) have been used to follow bond-breaking and bond-formation in biological systems. One example is nitric oxide binding to myoglobin (Mb).[117] For this, a 3-dimensional RKHS PES was fitted to reference density functional theory calculations for radial and angular degrees of freedom of the NO ligand with respect to the heme unit and the iron out-of-plane motion with respect to the heme plane. All remaining degrees of freedom of the solvated protein−ligand system were treated with an empirical energy function.

Extensive reactive MD simulations with such a mixed ML/empirical energy function provided the first structural interpretation of the metastable states in Mb-NO.[180] Consistent with recent optical and X-ray absorption experiments, which are unable to directly relate spectroscopic response with the underlying structure, these simulations found two processes: one on the 10 ps and another one on the 100 ps time scale. They correspond to rebinding of the ligand to Histidine64 in an "open" and a "closed" conformation. A mixed and reactive ML/empirical energy function paired with the accuracy of RKHS to represent the reference points is required to carry out the necessary statistical sampling and reach the time scale of the processes which is not possible with QM/MM techniques.

4.2. Computer-Guided Enzyme Design

Within the "theozyme" approach (i.e., enzymes or enzyme active sites generated from computation), computer-based methods have also been used to modify or design amino acid sequences that catalyze organic reactions.[181,182] Directed evolution has been used experimentally to (re)design enzyme function.[183] Starting from an originally designed Kemp eliminase,[184] the efficiency of the protein was assessed after 7 and 17 rounds of evolution and found to have increased by more than a 9 orders of magnitude. Analysis of nuclear magnetic resonance (NMR) measurements indicated that the key difference between the original and the evolved enzyme is the dynamic fluctuations of the catalytic amino acids that increase the probability to occupy catalytically proficient conformation and reduce the number of overall possible conformations.[183] Given the inherent similarities of evolutionary strategies and neural networks, it is expected that ML-based technologies will provide further scope to apply computer-based techniques for optimization and even reshaping protein sequences for particular reactions. Solving such problems is akin to the quest followed in materials design which aims at using ML to develop materials with given properties (such as electrical conductivity, melting point, or hardness).

Starting from a computationally designed protein (1A53-2) that catalyzes the base-promoted E2 elimination of 5-nitrobenzisoxazole, which is a Kemp elimination reaction, it

https://doi.org/10.1021/acs.chemrev.1c00033
Chem. Rev. XXXX, XXX, XXX−XXX

was shown that the dynamics of the evolved protein differed from the original structure.[185] Experimentally, this showed up in the negative activation heat capacity that indicates pronounced adaptation to temperature. One of the molecular foundations that accompanies changes in amino acid sequence that determine the reaction free energy landscape is communication[186] between spatially separated sites, i.e. "allostery". However, directly relating changes in reactivity with modifications of the amino acid sequence and the ensuing molecular dynamics from computations is extremely challenging, as the "signals" are often small and superimposed on a large background of equilibrium fluctuations. Recent computational work on ketol-acid reductoisomerase reported that in protein−ligand complexes "reactivity promoting regions" within the conformational space sampled along the reaction path should exist which distinguish reactive from almost-reactive trajectories.[187] This analysis was based on 68 geometrical features that were regularized using LASSO and subsequently clustered. This lead to the suggestion to first identify such special regions in the protein and use this knowledge for subsequent modification to obtain significant rate enhancements.

5. MACHINE LEARNING IN THE CONTEXT OF EXPERIMENTS

Relevant experimental observables in the context of chemical reactivity can include characteristics as diverse as the prediction of reaction probabilities,[55,56,102] differential cross sections or product state distributions,[58,72] the prediction of reaction outcomes (given specific input compounds),[188−190] finding optimal reaction conditions,[191,192] or predicting and identifying fragmentation patterns from unimolecular decomposition using mass spectroscopy.[193,194] For prediction of reaction outcomes it is worthwhile to mention that some of the efforts go back at least 50 years with initial efforts to use computer-aided strategies for organic synthesis (CAOS).[195−201] Similarly, using ML-based techniques (referred to as "AI" at the time) for analysis of mass spectrometric data started also in the mid-1960s with DENDRAL.[202−204]

5.1. Organic Reactions

The field of "retrosynthesis" started around 1967.[196] Since then, much progress has been made. Initially, rule-based expert systems such as CAMEO[201,205] or EROS[206] have attempted to predict reaction outcomes. It was found that such approaches do not scale well and are not easily generalizable. Later attempts were based on machine learning approaches from training of labeled reactions.[207] The ReactionPredictor uses a feature vector consisting of physicochemical and topological features, including molecular weight, formal and partial charges of the atoms, information about atom sizes, together with information similar to molecular fingerprints. For training the network, 1516 features were retained. The learning was done within an artificial NN with sigmoidal activation functions and one hidden layer, i.e. a shallow NN. For organic textbook reactions as the training and the validation set, ∼96% accuracy was reported.[207] Using a fingerprint-based NN, an accuracy of 80% of selected textbook reactions was found.[188] The main limitation in this prediction exercise was due to the limitations of the SMARTS transformation to describe the mechanism of the reaction type completely. Due to the flexibility in the descriptor, it is possible to further expand this algorithm to account for the reaction conditions.[188] Similarly, deep reinforcement learning has been applied to optimize chemical reactions.[191] For four different reactions it was shown that with the product yield as the objective to maximize the deep reaction optimizer (DRO) found the optimal conditions within 40 steps, with the total time of 30 min required to optimize a reaction in a microdroplet. Also, optimizing reaction conditions for one type of reaction and testing on a different reaction (here the Pomeranz−Fritsch synthesis of isoquinoline and the Friedländer synthesis of substituted quinoline, respectively) reached a higher yield with fewer optimization cycles.

More recently, a combination of Monte Carlo tree search (MCTS) with three NNs (one for proposing a restricted number of automatically extracted transformations, a second one to predict reaction feasibility, and a third one to estimate the position value for each transformation) to yield 3N-MCTS has been proposed.[189] To assess the quality of the machine-learned predictions, double blind tests with 45 graduate-level organic chemists from two world-leading chemistry institutes had to choose one of two routes leading to the same molecule on the basis of personal preference and synthetic plausibility. In one test the participants had to choose between a route reported by expert chemists in the literature and a route generated by the 3N-MCTS algorithm. For the nine routes to assess, three proposed by expert chemists were chosen whereas for the remaining six the one suggested by 3N-MCTS was preferred. Despite this success, important challenges remain, including synthesis routes for natural products, prediction of stereochemical outcomes, tautomerization equilibria, or prediction of reaction conditions.

Complementary to 3N-MCTS, Chematica relies on ∼50000 hand-annotated and curated reactions to predict reaction routes and outcomes and has been recently tested on reaction planning for eight medicinally relevant targets.[208] The approach combines expert chemical knowledge with network search and artificial intelligence algorithms. A direct comparison of the two approaches is as of now not available but would certainly be of considerable interest.

Despite these achievements, there are still limitations in using ML methods to predict the outcome of diverse organic reactions.[209] As an example of the intrinsic difficulties faced, one can consider the issue of "learning" in a conventional ML context. Every ML model learns from a finite number of training data, is tested on further independent data, and then can be used to predict unknown data.[7] Typically, the larger the size of the training data, the better the model. There are on the order of 10^7 reactions with more than $\sim 10^4$ different reaction types.[210] This leaves only 10^3 samples for every reaction which typically does not include different solvents, reaction conditions, substitutions, and other determinants that drive a chemical reaction. Hence, the statistics for "learning" still needs to be improved.

It will be of interest to see whether computer-assisted techniques can contribute to alleviate such problems. Given the unparalleled increase in computer efficiency, larger numbers of reactant, product, and transition state structures can be quantitatively evaluated routinely. One example is the very large data sets employed to train NN for molecular energy functions. The ANI-1 data set contains 2×10^7 structures of organic molecules.[211] For a summary of existing databases; see ref 212. With such approaches it may be possible to more broadly assess structural, substitutional, and electronic effects on chemical reactivity that undoubtedly are relevant for reaction outcomes. In addition, the effects of solvent need to be included. There has been recent progress in developing ML-based models for hydration and solvation free energies on compounds.[213−216]

https://doi.org/10.1021/acs.chemrev.1c00033
Chem. Rev. XXXX, XXX, XXX−XXX

This, together with the advances in electronic structure theory, may provide an avenue for further improvements of the models.

Recently, a modular robotic system for organic synthesis consisting of a Chemputer, a Chempiler, and a scripting language (ChASM) was combined to drive four modules consisting of a reaction flask, a filtration station, a liquid−liquid separation module, and a solvent evaporation module.[217] This system was used to automate the synthesis of compounds such as diphenhydramine hydrochloride, rufinamide, or sildenafil without human intervention.[217] Besides the attractive prospect to automate standard chemical procedures with the opportunity to discover new synthetic routes, such procedures also enhance the reproducibility of synthetic procedures. An alternative approach of a robotics-based platform driven by software uses 12.5 million published single-step reactions which were translated into a total of 163,723 rules.[218] With this input a forward NN was trained to predict what rules are most likely applicable for the synthesis of a particular target molecule. The testing of the platform was done on 15 reactions of different complexity. It was concluded that such robotic platforms coupled with curated data and powerful ML algorithms can relieve scientists from routine tasks so that they can rather focus on the more creative steps that lead to new ideas.[218]

Similar advances have been reported in the area of materials sciences by introducing the concept of autonomous experimentation[219,220] based on Phoenics[221] and its successor Gryffin.[222] Phoenics uses Bayesian neural networks to create a kernel-based surrogate model for the efficient optimization of chemical and material properties. Like other Bayesian optimization approaches, Phoenics balances the exploration and exploitation of parameter space to achieve sample-efficient experimental campaigns. Hence, such an approach also falls within the broader class of "model optimization problems" that are also used for reactive networks.[223]

For the design of novel, functional materials with specific, predefined properties, which—by definition—involve chemical reactions and transformations, the problem at hand is further exacerbated by the fact that the parameter space contains continuous (e.g., T or flow rates) and discrete (categorical) variables, such as the type of solvent, chemical substitutions, or the catalysts used. Gryffin,[222] an extension of the Phoenics framework, allows us to tackle such optimization problems by taking advantage of recent ML advances that enabled the continuous relaxation of discrete variables.[224,225] This approach has been applied to the autonomous optimization of a stereoselective Suzuki−Miyaura coupling between a vinyl sulfonate and an arylboronic acid to selectively generate the E-product isomer in high yield.[226] Along similar lines, reaction yield predictions were recently learned based on natural language architectures using an encoder model and a regression layer. As an example, the average R^2 for the Suzuki−Miyaura reactions was ∼0.8, similar to a model only trained on the Buchwald−Hartwig reactions. It was concluded that such models can perform equally well on different types of reactions and are robust with respect to the parameters and hyperparameters of the model.[60]

5.2. Fragmentation Reactions and Mass Spectrometry

Chemical structure determination using data from mass spectrometry was one of the early applications of expert systems ("AI") to problems involving decomposition reactions.[203] The earliest ML-based program to do this was "DENDRAL".[204] In a later effort,[227] the insights gained from reaction predictions using EROS[206] were used to develop a computational framework (MAss Spectra SIMulatOr - MASSIMO) to predict the mass spectrum given the structure of the compound. To put a criticism of DENDRAL into context ("...it is sad to say that, in the end, the DENDRAL project failed in its major objective of automatic structure elucidation by mass spectral data..."),[227] it should be noted that DENDRAL was aimed at the reverse task: structure determination for given mass spectrometric data.

More recently, NN-based techniques were developed to address the problem of competitive fragmentation modeling for electron ionization (CFM-EI).[193,228] Given a chemical structure, the model predicts an electron ionization (EI) mass spectrum (MS). Contrary to another approach available, CSI:FingerID,[229] competitive fragment modeling is applicable to both ions generated from electron ionization and electrospray ionization (ESI). CFM-ESI uses a probabilistic model based on systematic removal of all bond connections, every pair of bonds in rings, and considering all hydrogen rearrangements within the resulting fragments. The chemical features required for training the NN include properties such as broken bond types (single, double, and others), neighboring bond types, functional group features,[230] and others.[193] The data set for training, testing, and validation contained ∼20000 molecules. The performance of this model was 77% when querying against the measured reference spectra and 43% against the NIST database.

Compound structure identification (CSI) in predicting fingerprints and identifying metabolites (FingerID), referred to as CSI:FingerID,[229] uses molecular fragmentation trees with molecular fingerprint prediction based on multiple kernel learning.[231,232] Here, training was carried out on ∼6200 compounds. With the full training set the correct identification rate is ∼30%.[229] In a comparative assay based on PubChem, the identification rate for CSI:FingerID was ∼32% compared with ∼12% for CFM-ID.[229]

Most recently, analysis of high-resolution fragmentation mass spectra was carried out based on "class assignment and ontology prediction using mass spectrometry" (CANOPUS).[194] This workflow employs a number of support vector machines (SVMs) to predict fingerprints of the query compound which is the input to a deep neural network to predict all possible compound classes consistent with the query compound simultaneously. The SVMs are trained on experimental reference mass spectrometric data. Conversely, the DNNs are trained on millions of compound structures with molecular formulas as the feature vectors together with the number of atoms of a given type, the mass, and additional atom-based features as input to the DNN.[194] The binary molecular fingerprint, determined from CSI:FingerID,[229] and the molecular formula features from a fragmentation tree[233] are used as input to the DNN. The DNN is optimized using Adam.[99] With respect to performance as measured by the Matthews correlation coefficient (MCC = +1 for perfect classification and MCC = −1 for a completely wrong classification),[234] the ranges are from 0.875 for steroids to 0.972 for phosphocholines from the training set. For the test set, the MCCs ranged from 0.60 to 0.74.[194]

6. MACHINE LEARNING FOR ENTIRE REACTION NETWORKS

Reaction networks are relevant in various branches of chemistry, including but not limited to atmospheric reactions, combustion, and astrophysical and biological networks. Often such networks

https://doi.org/10.1021/acs.chemrev.1c00033
Chem. Rev. XXXX, XXX, XXX−XXX

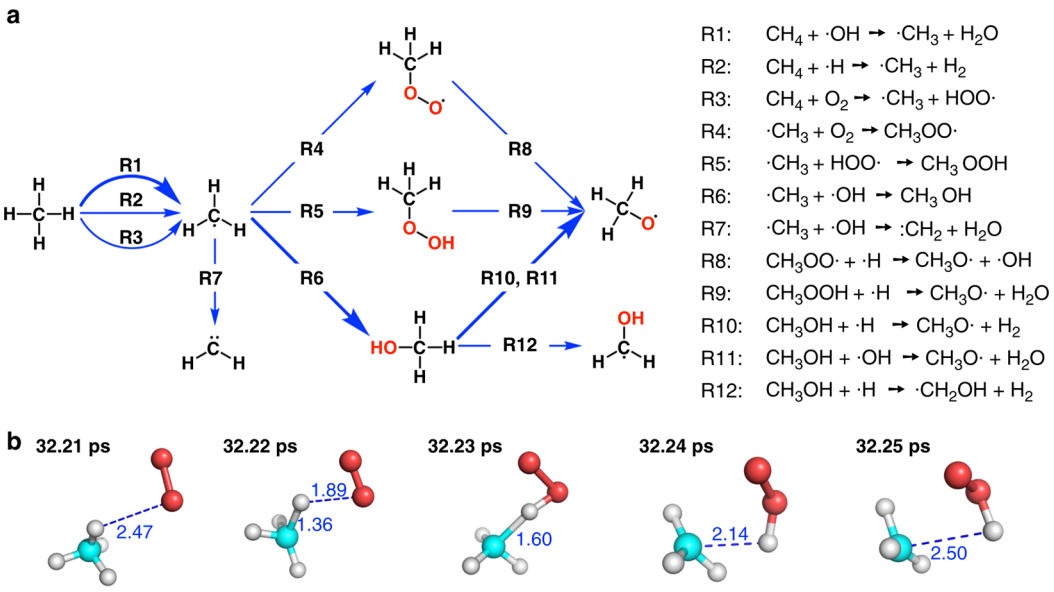


Figure 6. Initial stage of methane combustion. Panel a: Primary reaction pathways (left) and reactions R1 to R1 (right) during the initial stage of the combustion. Panel b: A real-time trajectory showing the reaction progress over 40 fs for hydrogen abstraction from methane by O_2. Carbon, oxygen, and hydrogen atoms are in cyan, red, and gray, respectively, with atom separations reported in angstroms. Figure adapted with permission from ref 237. Copyright 2020 Springer under Creative Commons CC-BY license, https://creativecommons.org/licenses/by/2.0/.

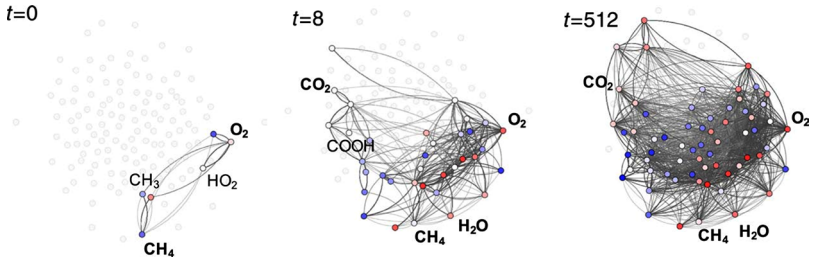


Figure 7. Methane combustion: each frame shows the reduced reaction network extracted from a microkinetic simulation of methane combustion depending on abstract, arbitrary simulation time. Reactants and products (bold) and intermediates (regular font) are indicated next to the nodes which are colored according to their absolute atomization energies from low (red) to high (blue). Figure adapted with permission from ref 239. Copyright 2020 Springer under Creative Commons CC-BY license, https://creativecommons.org/licenses/by/2.0/.

are sampled at the level of a stochastic network[235] by solving a large number of coupled ordinary differential equations.

More recently, it was attempted to directly propagate the nuclear dynamics within an *ab initio* nanoreactor.[236] Such an approach is still rooted in conventional *ab initio* molecular dynamics simulations and limited to the level of theory employed and the time scales accessible to such simulations. Very recently, a NN-based model was presented to follow combustion reactions in space and time, see Figure 6.[237] These simulations used the DeepMD NN architecture[238] to compute energies and forces for methane combustion (starting 100 CH_4 and 200 O_2 molecules) at 3000 K and found 798 different chemical reactions, some of which were as of now unknown.[237] The total simulation time covered was in the nanoseconds, and the accuracy of these simulations is only limited by the electronic structure data the NN was trained to.

In another recent attempt, methane combustion was simulated using an ML-trained model on atomization energies[239] using kernel ridge regression with a Smooth Overlap of Atomic Positions (SOAP) representation.[240] A mean-field, qualitative microkinetic simulation of a 50:50 mixture of CH_4 and O_2 using only the reaction energies (trained to an accuracy of ~0.1 eV) and the law of mass action was carried out. The resulting reduced reaction networks as a function of abstract simulation time are reported in Figure 7. Several notable species are formed in this simulation, including methanol, formic acid, and Criegee intermediates.

Machine-learning investigations of entire chemical reaction networks were recently undertaken.[223] Using "automated learning of algebraic models for optimization" (ALAMO),[241] the "Reaction Identification and Parameter Estimation" (RIPE) tool was developed and used to estimate and identify kinetic rate parameters from a postulated superset of reactions. RIPE was applied to combustion reactors to model catalyst conversion or alternative reaction mechanisms and stoichiometric relationships. Chemical looping combustion has been developed to isolate fuel from air in combustion reactions using an oxygen carrier shuttled between two flow reactors. In this application the task solved by the ML approach is to select the model to best describe the input data from experiment. For the application to alternative reaction mechanisms the technique is used to discern between a large number of alternative mechanistic pathways. Such approaches fall under the general heading of "optimal model selection" given a concrete, preconceived reaction network. The RIPE tool is available from www.idaes.org and can handle between 10^2 and 10^4 reactions.

https://doi.org/10.1021/acs.chemrev.1c00033
Chem. Rev. XXXX, XXX, XXX−XXX

7. FUTURE DEVELOPMENTS

In the following, possible developments in the field of ML-applications to chemical reactions are illustrated.

For small systems (containing few atoms) one question is whether in general the number of reference points for constructing global, reactive PESs can be dramatically reduced for accurate representations of intermolecular PESs when resorting to ML techniques. Previously, it was assumed that typically of the order of 10 points per degree of freedom is required for a good coverage of conformational space. Hence, for a diatom + diatom system of the order of 10^6 reference *ab initio* energies would be required. For global, reactive PESs this number is likely to be even larger. In the context of GP regression it has been argued that for an N-dimensional system only $\sim 10N$ well-selected points are required.[242] The work on combined ML and Bayesian optimization techniques[102] indicates that this is indeed possible for molecular systems, too. On the other hand, recent work on the SH + H → $S(^3P)$ + H_2 reaction[71] estimated that rather 500 points are required for faithful representation of the global, reactive PES, contrary to the ∼30 points that were found to be sufficient for the H + H_2 → H_2 + H system[102] despite the same dimensionality. Hence, the number of points required may depend on the presence of (permutational) symmetry and the chemical species involved and hence the overall topology of such a PES. An interesting future application of Bayesian optimization techniques is to combine it with Δ-ML, transfer learning or morphing on the basis of experimentally determined spectroscopic or reactive scattering data. In fact, this is how morphing[133] originally was conceived.

One of the challenges ahead in the field is to learn high-quality PESs while minimizing the number of reference points required. In other words, the task is to learn from "little data" as opposed to "big data". It has been recently demonstrated that for high-dimensional, nonreactive systems a few hundred points are sufficient to accurately represent the near-equilibrium PES using RKHS representation on energies and forces.[120] For the two largest molecules (CH_3CONH_2 and CH_3COCH_3) 2500 reference energies were found to be sufficient to obtain a mean averaged error of 0.01 and 0.07 kcal/mol on 1000 test points. The harmonic frequencies determined from such PESs are typically within 1 cm^{-1} of a normal-mode calculation using conventional normal-mode analysis from quantum chemical calculations at the same level of theory with a maximum deviation smaller than 10 cm^{-1}. One step toward minimizing the number of required reference energies was recently taken by using concurrent learning.[243] For this, a data set containing ∼36,000 structures of molecular clusters involved in alkane pyrolysis with different composition was generated to minimize redundancy. Based on this surprisingly small data set, reactive MD simulations were carried out using an NN-based model (DeepPot-SE) trained at the MN15/6-31G** level of theory. It was found that the fragmentation products from pyrolysis of *n*-dodecane are consistent with earlier reports from modeling experiments.

Another relevant question concerns the probing and specific improvement of high quality PESs in view of experimental observables. The question that arises in this context is which parts of the PES are "reliable" and which parts can be further improved. Chemical reactions by their very nature are sensitive to the global shape of a PES whereas other observables such as harmonic frequencies only probe the local shape and couplings between degrees of freedom. The PES regions sampled for specific observables has, e.g., been reported for the $N(^4S) + O_2(X^3\Sigma_g^-) \leftrightarrow O(^3P) + NO(X^2\Pi)$ reaction.[89] Another study developed a Bayesian ML approach to quantify uncertainties on PESs for the reactive O_2 + O system considering two different electronic states.[86] This effort started from Bayesian-based sensitivity analysis of computer models using GP.[244] Sensitivity analysis of PESs dates back at least 30 years[245] where the problem of inversion of rovibrational spectra for diatomic molecules has already been formulated within a Tikhonov regularization framework.

When using experimental observables to refine *ab initio* calculated PESs for Ne-HF, it has already been found that specific observables are only sensitive to particular regions of the PES.[133,246] Such "PES morphing approaches" have been extensively applied to small molecular systems.[247−263] There is also scope for extending this more broadly[36] and in the context of machine-learned PESs, as has been recently indicated for acetylacetone.[174] Such approaches also have the potential to more tightly integrate computation with experiment and to develop computational models that learn from experimental data.

Besides the actual number of points, it is also relevant to consider the question what configurations of a system to use for the reference calculations. The points should be placed in the most informative regions, i.e. the regions the observables of interest are actually sensitive to. This question has already been discussed in another contribution to this special issue.[8]

For highly accurate, small molecule reaction dynamics based on represented *ab initio* calculated PESs, one remaining challenge is the absence of a uniformly accurate and valid method to solve the electronic Schrödinger equation. While for single-reference problems coupled cluster (CC) techniques, such as CCSD(T), provide a valid "gold standard", such a generally applicable and robust technique is missing in regions of the PES for which a single-reference electronic wave function is not a good approximation. Multireference configuration interaction (MRCI) methods are a viable alternative, but they are not yet of the same overall quality as CC-based techniques, and they can be challenging to apply. Also, computing entire PESs with these methods can be nontrivial. It may be possible to apply mapping, scaling, and compound techniques to blend different methods across a precomputed grid of interaction energies as has been done for example for N_3^-.[264] One possibility is multireference coupled cluster (MRCC) methods, although they remain computationally challenging,[265] in particular when global, full-dimensional PESs are required. On the other hand, full configuration interaction PESs have been computed for small, few-electron systems,[132] and with increasing computer speed such calculations and systems become tractable.

Another area of future development concerns reactivity involving electronically excited states. For small systems, *ab initio* calculations, ML-based representations of the PES and QCT dynamics simulations similar to ground state problems, have been used for rate calculations for quite some time.[73,88,89,118,266] However, for larger systems, the dynamics in the excited state is a challenging problem in itself and ML-based techniques applied to this (nonreactive) problem only start to appear.[267] One particular point of concern is the nonadiabatic dynamics and the coupling matrix elements involved in the transition between neighboring electronic states.[268,269]

https://doi.org/10.1021/acs.chemrev.1c00033
Chem. Rev. XXXX, XXX, XXX−XXX

It is expected that combining currently available ML techniques for fragmentation using mass spectrometry with advanced *ab initio* calculations and data from existing databases (see ref 212) will further boost the quantitative side of chemical structure determination from MS experiments. It is now possible to determine optimized structures for entire chemical libraries containing millions of compounds[211,270] at the density functional theory level of theory. Such information is potentially useful for better describing the reaction energetics of such decomposition and fragmentation processes and the relative importance of each of the channels.

Going beyond the "theozyme" approach, it can be anticipated that ML methods applied to enzyme design will also bring the field of enzymology and ligand design forward. Although a comprehensive and generally accepted ML-based technology to do this is not available as of now, the recent success of ML[271] in winning the CASP14[272] clearly foreshadows that machine learning will play an eminent role in design of enzymatically active amino acid sequences. Compared with more established approaches, one of the deciding factors was the deeper search capabilities of a CNN together with the choice of objective function to be optimized which included distances between C_β atoms, backbone torsion angles, and the prevention of steric clashes.[271] Thus, a combination of "trainability" through availability of curated and sufficiently complete reference data, judicious choice of target information with respect to which optimization can be carried out, and architecture of the underlying NN is a driver for successful ML missions.

Successes in application of ML-based methods to pharmacological tasks include toxicity[273] or ligand search for serotonin receptors such as 5-HT_6.[274] However, one of the main obstacles in broader application of ML-based methods, specifically for deep learning approaches, is the weak link between the findings of an ML-based treatment of the problem and the chemical reasoning, i.e. the "cause and effect" problem. Further progress[275] in "interpretable ML" will be essential to move ML-based approaches away from "black box" methods. Another challenge in drug design is the database quality and structural diversity of pharmaceutically active compounds. Contrary to the problem of protein structure prediction from sequence,[272] ligand design and development of pharmaceutically active substances requires an understanding of "why" particular modifications on a ligand are beneficial for their physiological effect. Similarly, for the question of actual "protein folding pathways" and rates, it is anticipated that a deeper understanding of the actual process is required which differs from the "end point problem" that is solved by alphafold.[271]

As pointed out at various places in this review, ML-based methods are inherently connected with the analysis of large data sets. Such data sets often first need to be generated. Because this in itself is a time-consuming and laborious process, publication and open access of these data sets is important for the community at large and needs to be mandatory. To be most valuable to researchers, it is important to develop and adhere to standards how such data sets should be generated, stored, searched, and accessed. For quantum chemical databases this also includes specification of information such as the codes used, convergence criteria for SCF calculations, and structure minimization. With such standards in place it will be easier to benchmark different ML techniques on the same data sets and to further extend existing sets. In experimental physical chemistry, specifically in spectroscopy, the "NIST chemistry webbook" is an important source of information for validated experimental results for a large number of molecules.[276] The role of "National Institutes" in coordinating such efforts has been recently discussed for experimental benchmark data sets.[277] One of the challenges for "ML based on quantum chemical data" is the sheer size of some of these data sets. For example, the ANI-I data set contains energies for 20 million molecular structures.[278] Given the fact that with modern computer architectures and using density functional theory-based methods such data sets can be generated in a few weeks, it is anticipated that managing, curating, and making available such data sets constitutes a major challenge.

Visualization and virtual reality techniques for exploration of chemical reactivity[279−282] are other areas in which ML-based techniques may become relevant.[283] Such approaches are likely to also become increasingly prevalent in classroom chemistry.[283−286] A personal note is related to my first viewing of an MD simulation of myoglobin (using VMD[280]) on a desktop screen as a postdoc working with Prof. M. Karplus. It was fascinating to follow the dance of atoms when nitric oxide was binding to and unbinding from the heme-iron.[287] Watching this reminded me immediately of the notion "...that everything that living things do can be understood in terms of the jigglings and wigglings of atoms."[288] The impression of this shaped my personal view of chemical and biological systems as inherently dynamical. Without capturing the dynamics it was evident that a comprehensive understanding of chemical and biological function,[289] including enzymatic activity,[183] would not be possible. Developments for ML-based techniques for reaction-path-following methods[290] and generation of reaction networks[291] are currently under way but as of now do not explicitly make use of machine learning. The combination of electronic structure theory, molecular dynamics, machine learning, and virtual reality brings this one step closer and will also have potentially far reaching, transformative effects on the way we will teach chemistry in the future.[292]

Machine learning techniques also have the potential to transform the way in which the community regards the relationship between experiment, simulation, and theory. Together with automatization (robotics), the combination of ML, experiment, and simulation bears the potential to develop integrated systems which optimize chemical reaction systems with respect to a particular task ("loss function"), such as maximizing yield, turnover, and rate or minimizing use of problematic solvent. One of the fields which has seen many advances is reaction planning for organic synthesis[189,190] although examples for such system optimization have been presented more than 20 years ago for gas phase reaction dynamics.[293−295] in the context of "controlled chemistry".

In summary, ML-based approaches applied to chemical reactivity is a rapidly expanding field. The challenges ahead concern the accurate, quantitative, and exhaustive determination of reaction outcomes, rates, and (internal) state distributions. Coupled with robotic platforms, reaction yields and reaction conditions can be optimized using ML and Bayesian techniques. In the field of enzyme design, appreciable improvements of turnover rates can be expected from coupling experiments with ML-based approaches, and for protein−ligand interaction and recognition the recent advances made for protein structure prediction will provide important insights. Finally, exploring entire reaction networks has been possible very recently for specific processes (e.g., methane oxidation). Together with improved, high-quality reference data, exploration of chemical

https://doi.org/10.1021/acs.chemrev.1c00033
Chem. Rev. XXXX, XXX, XXX−XXX

space for reactions relevant to combustion, atmospheric sciences, and astrophysical chemistry will become viable.

AUTHOR INFORMATION

Corresponding Author

Markus Meuwly − *Department of Chemistry, University of Basel, 4056 Basel, Switzerland; Department of Chemistry, Brown University, Providence, Rhode Island 02912, United States;* orcid.org/0000-0001-7930-8806; Email: m.meuwly@unibas.ch

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.chemrev.1c00033

Notes

The author declares no competing financial interest.

Biography

Markus Meuwly studied Physics at the University of Basel and completed his Ph.D. in Physical Chemistry working with Prof. J. P. Maier. After postdocs with Prof. J. Hutson (Durham) and Prof. M. Karplus (Strasbourg and Harvard) as a Swiss National Science Foundation Postdoctoral Scholar, he started as a Förderprofessor at the University of Basel in 2002, where he is Full Professor of Physical and Computational Chemistry. He also holds a visiting professorship at Brown University, Providence, RI. His scientific interests range from accurate intermolecular interactions based on multipolar and kernel- and neural network-based representations to applications of quantitative molecular simulations for cold (interstellar) and hot (hypersonics) environments and the investigation of the reactive dynamics and spectroscopy in proteins and in the condensed phase. Several of the tools are available in the CHARMM molecular simulation program.

ACKNOWLEDGMENTS

I am pleased to acknowledge the contributions to this research effort from many of my students and co-workers over the past decade. Special thanks go to Ms. Upadhyay for assistance with Figure 5. I also thank Profs. J. M. Bowman, H. Guo, R. Krems, D. Wishart, N. Sahinidis, A. Aspuru-Guzik, and A. von Lilienfeld and Dr. M. Aldeghi for correspondence. This work has been financially supported by the Swiss National Science Foundation (NCCR-MUST and Grant No. 200021-7117810), the AFOSR, and the University of Basel.

REFERENCES

(1) Vereecken, L.; Aumont, B.; Barnes, J.; Bozzelli, I.; Goldman, M.; Green, S.; Madronich, W. H.; McGillen, M.; Mellouki, A. Perspective on Mechanism Development and Structure-Activity Relationships for Gas-Phase Atmospheric Chemistry. *Int. J. Chem. Kinet.* **2018**, *50*, 435−469.

(2) Klippenstein, S. J. From Theoretical Reaction Dynamics to ChemicalModeling of Combustion. *Proc. Combust. Inst.* **2017**, *36*, 77−111.

(3) Wakelam, V.; Herbst, E.; Loison, J. C.; Smith, I. W. M.; Chandrasekaran, V.; Pavone, B.; Adams, N. G.; Bacchus-Montabonel, M. C.; Bergeat, A.; Beroff, K.; et al. A Kinetic Database for Astrochemistry (KIDA). *Astrophys. J., Suppl. Ser.* **2012**, *199*, 21.

(4) Van Der Kamp, M. W.; Mulholland, A. J. Combined Quantum Mechanics/Molecular Mechanics (QM/MM) Methods in Computational Enzymology. *Biochemistry* **2013**, *52*, 2708−2728.

(5) Cheong, P. H.-Y.; Legault, C. Y.; Um, J. M.; Celebi-Ölcüm, N.; Houk, K. N. Quantum Mechanical Investigations of Organocatalysis: Mechanisms, Reactivities, and Selectivities. *Chem. Rev.* **2011**, *111*, 5042−5137.

(6) Koner, D.; Salehi, S. M.; Mondal, P.; Meuwly, M. Non-Conventional Force Fields for Applications in Spectroscopy and Chemical Reaction Dynamics. *J. Chem. Phys.* **2020**, *153*, 010901.

(7) Unke, O. T.; Chmiela, S.; Sauceda, H. E.; Gastegger, M.; Poltavsky, I.; Schütt, K. T.; Tkatchenko, A.; Müller, K.-R. Machine Learning Force Fields. *Chem. Rev.* **2021**, DOI: 10.1021/acs.chemrev.0c01111.

(8) Manzhos, S.; Carrington, T., Jr Neural Network Potential Energy Surfaces for Small Molecules and Reactions. *Chem. Rev.* **2020**, DOI: 10.1021/acs.chemrev.0c00665.

(9) Jiang, B.; Li, J.; Guo, H. High-Fidelity Potential Energy Surfaces for Gas-Phase and Gas- Surface Scattering Processes From Machine Learning. *J. Phys. Chem. Lett.* **2020**, *11*, 5120−5131.

(10) Cui, Q. Perspective: Quantum Mechanical Methods in Biochemistry and Biophysics. *J. Chem. Phys.* **2016**, *145*, 140901.

(11) Herbert, J. M.; Head-Gordon, M. Accelerated, energy-conserving Born-Oppenheimer molecular dynamics via Fock matrix extrapolation. *Phys. Chem. Chem. Phys.* **2005**, *7*, 3269−3275.

(12) Senn, H. M.; Thiel, W. QM/MM Methods for Biomolecular Systems. *Angew. Chem., Int. Ed.* **2009**, *48*, 1198−1229.

(13) Roston, D.; Demapan, D.; Cui, Q. Leaving Group Ability Observably Affects Transition State Structure in a Single Enzyme Active Site. *J. Am. Chem. Soc.* **2016**, *138*, 7386−7394.

(14) Kulik, H. J.; Zhang, J.; Klinman, J. P.; Martinez, T. J. How Large Should the QM Region be in QM/MM Calculations? The Case of Catechol O-Methyltransferase. *J. Phys. Chem. B* **2016**, *120*, 11381−11394.

(15) Jindal, G.; Warshel, A. Exploring the Dependence of QM/MM Calculations of Enzyme Catalysis on the Size of the QM Region. *J. Phys. Chem. B* **2016**, *120*, 9913−9921.

(16) Kitaura, K.; Ikeo, E.; Asada, T.; Nakano, T.; Uebayasi, M. Fragment Molecular Orbital Method: An Approximate Computational Method for Large Molecules. *Chem. Phys. Lett.* **1999**, *313*, 701−706.

(17) Nakata, H.; Fedorov, D. G.; Nagata, T.; Kitaura, K.; Nakamura, S. Simulations of Chemical Reactions With the Frozen Domain Formulation of the Fragment Molecular Orbital Method. *J. Chem. Theory Comput.* **2015**, *11*, 3053−3064.

(18) Gordon, M. S.; Fedorov, D. G.; Pruitt, S. R.; Slipchenko, L. V. Fragmentation Methods: A Route to Accurate Calculations on Large Systems. *Chem. Rev.* **2012**, *112*, 632−672.

(19) Babu, K.; Gadre, S. R. Ab Initio Quality One-Electron Properties of Large Molecules: Development and Testing of Molecular Tailoring Approach. *J. Comput. Chem.* **2003**, *24*, 484−495.

(20) Deev, V.; Collins, M. A. Approximate Ab Initio Energies by Systematic Molecular Fragmentation. *J. Chem. Phys.* **2005**, *122*, 154102.

(21) Huang, L.; Massa, L.; Karle, J. Kernel Energy Method Illustrated With Peptides. *Int. J. Quantum Chem.* **2005**, *103*, 808−817.

(22) Zhang, D. W.; Zhang, J. Molecular Fractionation with Conjugate Caps for Full Quantum Mechanical Calculation of Protein−Molecule Interaction Energy. *J. Chem. Phys.* **2003**, *119*, 3599−3605.

(23) Xu, M.; Zhu, T.; Zhang, J. Z. A Force Balanced Fragmentation Method for Ab Initio Molecular Dynamic Simulation of Protein. *Front. Chem.* **2018**, *6*, 189.

(24) Ellison, F. O. A Method of Diatomics in Molecules. I. General Theory and Application to H_2O. *J. Am. Chem. Soc.* **1963**, *85*, 3540−3544.

(25) Ellison, F. O.; Huff, N. T.; Patel, J. C. A Method of Diatomics in Molecules. II. H and H_3^+. *J. Am. Chem. Soc.* **1963**, *85*, 3544−3547.

(26) Karplus, M.; Porter, R. N.; Sharma, R. D. Exchange Reactions With Activation Energy. I. Simple Barrier Potential for (H, H_2). *J. Chem. Phys.* **1965**, *43*, 3259−3287.

(27) Warshel, A.; Weiss, R. an Empirical Valence Bond Approach for Comparing Reactions in Solutions and in Enzymes. *J. Am. Chem. Soc.* **1980**, *102*, 6218−6226.

(28) Schmitt, U.; Voth, G. Multistate Empirical Valence Bond Model for Proton Transport in Water. *J. Phys. Chem. B* **1998**, *102*, 5547−5551.

(29) Van Duin, A.; Dasgupta, S.; Lorant, F.; Goddard, W. ReaxFF: A Reactive Force Field for Hydrocarbons. *J. Phys. Chem. A* **2001**, *105*, 9396−9409.
(30) Stoliarov, S. I.; Westmoreland, P. R.; Nyden, M. R.; Forney, G. P. A Reactive Molecular Dynamics Model of Thermal Decomposition in Polymers: I. Poly (Methyl Methacrylate). *Polymer* **2003**, *44*, 883−894.
(31) Smith, K.; Stoliarov, S.; Nyden, M.; Westmoreland, P. RMDff: A Smoothly Transitioning, Forcefield-Based Representation of Kinetics for Reactive Molecular Dynamics Simulations. *Mol. Simul.* **2007**, *33*, 361−368.
(32) Nutt, D. R.; Meuwly, M. Studying Reactive Processes With Classical Dynamics: Rebinding Dynamics in MbNO. *Biophys. J.* **2006**, *90*, 1191−1201.
(33) Danielsson, J.; Meuwly, M. Atomistic Simulation of Adiabatic Reactive Processes Based on Multi-State Potential Energy Surfaces. *J. Chem. Theory Comput.* **2008**, *4*, 1083−1093.
(34) Nagy, T.; Yosa Reyes, J.; Meuwly, M. Multisurface Adiabatic Reactive Molecular Dynamics. *J. Chem. Theory Comput.* **2014**, *10*, 1366−1375.
(35) Schmid, M. H.; Das, A. K.; Landis, C. R.; Meuwly, M. Multi-State VALBOND for Atomistic Simulations of Hypervalent Molecules, Metal Complexes, and Reactions. *J. Chem. Theory Comput.* **2018**, *14*, 3565−3578.
(36) Xu, Z.-H.; Meuwly, M. Multistate Reactive Molecular Dynamics Simulations of Proton Diffusion in Water Clusters and in the Bulk. *J. Phys. Chem. B* **2019**, *123*, 9846−9861.
(37) Farah, K.; Müller-Plathe, F.; Böhm, M. C. Classical Reactive Molecular Dynamics Implementations: State of the Art. *ChemPhysChem* **2012**, *13*, 1127−1151.
(38) Collins, M. A. Molecular Potential-Energy Surfaces for Chemical Reaction Dynamics. *Theor. Chem. Acc.* **2002**, *108*, 313−324.
(39) Bowman, J. M.; Czako, G.; Fu, B. High-Dimensional Ab Initio Potential Energy Surfaces for Reaction Dynamics Calculations. *Phys. Chem. Chem. Phys.* **2011**, *13*, 8094−8111.
(40) Hynes, J. T. Chemical Reaction Dynamics in Solution. *Annu. Rev. Phys. Chem.* **1985**, *36*, 573−597.
(41) Hynes, J. T. Molecules in Motion: Chemical Reaction and Allied Dynamics in Solution and Elsewhere. *Annu. Rev. Phys. Chem.* **2015**, *66*, 1−20.
(42) Warshel, A. Computer Simulations of Enzyme Catalysis: Methods, Progress, and Insights. *Annu. Rev. Biophys. Biomol. Struct.* **2003**, *32*, 425−443.
(43) Gao, J.; Ma, S.; Major, D. T.; Nam, K.; Pu, J.; Truhlar, D. G. Mechanisms and Free Energies of Enzymatic Reactions. *Chem. Rev.* **2006**, *106*, 3188−3209.
(44) Himo, F. Recent Trends in Quantum Chemical Modeling of Enzymatic Reactions. *J. Am. Chem. Soc.* **2017**, *139*, 6780−6786.
(45) Amaro, R. E.; Mulholland, A. J. Multiscale Methods in Drug Design Bridge Chemical and Biological Complexity in the Search for Cures. *Nature Reviews Chemistry* **2018**, *2*, 1−12.
(46) Arcus, V. L.; Mulholland, A. J. Temperature, Dynamics, and Enzyme-Catalyzed Reaction Rates. *Annu. Rev. Biophys.* **2020**, *49*, 163−180.
(47) Farah, K.; Leroy, F.; Müller-Plathe, F.; Böhm, M. C. Interphase Formation During Curing: Reactive Coarse Grained Molecular Dynamics Simulations. *J. Phys. Chem. C* **2011**, *115*, 16451−16460.
(48) Vapnik, V. N. *Statistical Learning Theory*; Wiley-Interscience, 1998.
(49) Minsky, M.; Papert, S. *Perceptrons: An Introduction to Computational Geometry*; MIT Press, 1969.
(50) Fukushima, K. Neocognitron - A Self-Organizing Neural Network Model for a Mechanism of Pattern-Recognition Unaffected by Shift in Position. *Biol. Cybern.* **1980**, *36*, 193−202.
(51) Hopfield, J. Neural Networks and Physical Systems With Emergent Collective Computational Abilities. *Proc. Natl. Acad. Sci. U. S. A.* **1982**, *79*, 2554−2558.
(52) Linnainmaa, S. Taylor Expansion of the Accumulated Rounding Error. *BIT Numerical Mathematics* **1976**, *16*, 146−160.
(53) Rumelhart, D.; Hinton, G.; Williams, R. Learning Representations by Back-Propagating Errors. *Nature* **1986**, *323*, 533−536.
(54) Werbos, P. *Beyond Regression: New Tools for Prediction and Analysis in the Behavioral Sciences*. Ph.D. Dissertation, Harvard University, 1974.
(55) Houston, P. L.; Nandi, A.; Bowman, J. M. A Machine Learning Approach for Prediction of Rate Constants. *J. Phys. Chem. Lett.* **2019**, *10*, 5250−5258.
(56) Nandi, A.; Bowman, J. M.; Houston, P. A Machine Learning Approach for Rate Constants. II. Clustering, Training, and Predictions for the $O(^3P)$+ HCl ⇒OH+ Cl Reaction. *J. Phys. Chem. A* **2020**, *124*, 5746−5755.
(57) Komp, E.; Valleau, S. Machine Learning Quantum Reaction Rate Constants. *J. Phys. Chem. A* **2020**, *124*, 8607−8613.
(58) Koner, D.; Unke, O. T.; Boe, K.; Bemish, R. J.; Meuwly, M. Exhaustive State-to-State Cross Sections for Reactive Molecular Collisions From Importance Sampling Simulation and a Neural Network Representation. *J. Chem. Phys.* **2019**, *150*, 211101.
(59) Arnold, J.; Koner, D.; Käser, S.; Singh, N.; Bemish, R. J.; Meuwly, M. Machine Learning for Observables: Reactant to Product State Distributions for Atom-Diatom Collisions. *J. Phys. Chem. A* **2020**, *124*, 7177−7190.
(60) Schwaller, P.; Vaucher, A. C.; Laino, T.; Reymond, J.-L. Prediction of Chemical Reaction Yields Using Deep Learning. *Mach. Learn. Sci. Technol.* **2021**, *2*, 015016.
(61) Jorgensen, W. L.; Tirado-Rives. J. the OPLS Potential Functions for Proteins - Energy Minimizations for Crystals of Cyclic-Peptides and Crambin. *J. Am. Chem. Soc.* **1988**, *110*, 1657−1666.
(62) MacKerell, A.; Bashford, D.; Bellott, M.; Dunbrack, R.; Evanseck, J.; Field, M.; Fischer, S.; Gao, J.; Guo, H.; Ha, S.; et al. All-Atom Empirical Potential for Molecular Modeling and Dynamics Studies of Proteins. *J. Phys. Chem. B* **1998**, *102*, 3586−3616.
(63) Wang, J.; Wolf, R. M.; Caldwell, J. W.; Kollman, P. a.; Case, D. A. Development and Testing of a General Amber Force Field. *J. Comput. Chem.* **2004**, *25*, 1157−1174.
(64) Oostenbrink, C.; Villa, A.; Mark, A. E.; Van Gunsteren, W. F. a Biomolecular Force Field Based on the Free Enthalpy of Hydration and Solvation: The GROMOS Force-Field Parameter Sets 53A5 and 53A6. *J. Comput. Chem.* **2004**, *25*, 1656−1676.
(65) Ponder, J. W.; Wu, C.; Ren, P.; Pande, V. S.; Chodera, J. D.; Schnieders, M. J.; Haque, I.; Mobley, D. L.; Lambrecht, D. S.; DiStasio, R. a., Jr; et al. Current Status of the AMOEBA Polarizable Force Field. *J. Phys. Chem. B* **2010**, *114*, 2549−2564.
(66) Kramer, C.; Gedeck, P.; Meuwly, M. Atomic Multipoles: Electrostatic Potential Fit, Local Reference Axis Systems and Conformational Dependence. *J. Comput. Chem.* **2012**, *33*, 1673−1688.
(67) Shi, Y.; Xia, Z.; Zhang, J.; Best, R.; Wu, C.; Ponder, J. W.; Ren, P. Polarizable Atomic Multipole-Based AMOEBA Force Field for Proteins. *J. Chem. Theory Comput.* **2013**, *9*, 4046−4063.
(68) Bereau, T.; Kramer, C.; Meuwly, M. Leveraging Symmetries of Static Atomic Multipole Electrostatics in Molecular Dynamics Simulations. *J. Chem. Theory Comput.* **2013**, *9*, 5450−5459.
(69) Leven, I.; Hao, H.; Das, A. K.; Head-Gordon, T. a Reactive Force Field With Coarse-Grained Electrons for Liquid Water. *J. Phys. Chem. Lett.* **2020**, *11*, 9240−9247.
(70) McDaniel, J. G.; Schmidt, J. First-Principles Many-Body Force Fields From the Gas Phase to Liquid: A "Universal" Approach. *J. Phys. Chem. B* **2014**, *118*, 8042−8053.
(71) Kolb, B.; Marshall, P.; Zhao, B.; Jiang, B.; Guo, H. Representing Global Reactive Potential Energy Surfaces Using Gaussian Processes. *J. Phys. Chem. A* **2017**, *121*, 2552−2557.
(72) Koner, D.; Bemish, R. J.; Meuwly, M. Dynamics on Multiple Potential Energy Surfaces: Quantitative Studies of Elementary Processes Relevant to Hypersonics. *J. Phys. Chem. A* **2020**, *124*, 6255−6269.
(73) Koner, D.; Bemish, R. J.; Meuwly, M. The $C(^3P) + NO(X^2\Pi) \rightarrow O(^3P) + CN(X^2\Sigma^+)$, $N(^2D)/N(^4S) + CO(X^1\Sigma^+)$ Reaction: Rates, Branching Ratios, and Final States From 15 to 20000 K. *J. Chem. Phys.* **2018**, *149*, 094305.

https://doi.org/10.1021/acs.chemrev.1c00033
Chem. Rev. XXXX, XXX, XXX−XXX

(74) Hollebeek, T.; Ho, T.-S.; Rabitz, H. Constructing Multidimensional Molecular Potential Energy Surfaces From Ab Initio Data. *Annu. Rev. Phys. Chem.* **1999**, *50*, 537−570.
(75) Unke, O. T.; Meuwly, M. Toolkit for the Construction of Reproducing Kernel-Based Representations of Data: Application to Multidimensional Potential Energy Surfaces. *J. Chem. Inf. Model.* **2017**, *57*, 1923−1931.
(76) Qu, C.; Yu, Q.; Bowman, J. M. Permutationally Invariant Potential Energy Surfaces. *Annu. Rev. Phys. Chem.* **2018**, *69*, 151−175.
(77) Behler, J.; Parrinello, M. Generalized Neural-Network Representation of High-Dimensional Potential-Energy Surfaces. *Phys. Rev. Lett.* **2007**, *98*, 146401.
(78) Schütt, K. T.; Sauceda, H. E.; Kindermans, P.-J.; Tkatchenko, A.; Müller, K.-R. SchNet−a Deep Learning Architecture for Molecules and Materials. *J. Chem. Phys.* **2018**, *148*, 241722.
(79) Unke, O. T.; Meuwly, M. PhysNet: A Neural Network for Predicting Energies, Forces, Dipole Moments, and Partial Charges. *J. Chem. Theory Comput.* **2019**, *15*, 3678−3693.
(80) Käser, S.; Unke, O. T.; Meuwly, M. Reactive Dynamics and Spectroscopy of Hydrogen Transfer From Neural Network-Based Reactive Potential Energy Surfaces. *New J. Phys.* **2020**, *22*, 055002.
(81) Ho, T.-S.; Rabitz, H. A General Method for Constructing Multidimensional Molecular Potential Energy Surfaces From Ab Initio Calculations. *J. Chem. Phys.* **1996**, *104*, 2584−2597.
(82) Hollebeek, T.; Ho, T.-S.; Rabitz, H. A Fast Algorithm for Evaluating Multidimensional Potential Energy Surfaces. *J. Chem. Phys.* **1997**, *106*, 7223−7227.
(83) Truhlar, D. G.; Muckerman, J. T. In *Atom - Molecule Collision Theory*; Bernstein, R. B., Ed.; Springer US, 1979; pp 505−566.
(84) Yurchenko, S. N.; Thiel, W.; Jensen, P. Theoretical ROVibrational Energies (TROVE): A Robust Numerical Approach to the Calculation of Rovibrational Energies for Polyatomic Molecules. *J. Mol. Spectrosc.* **2007**, *245*, 126−140.
(85) Csaszar, A. G.; Fabri, C.; Szidarovszky, T.; Matyus, E.; Furtenbacher, T.; Czako, G. The Fourth Age of Quantum Chemistry: Molecules in Motion. *Phys. Chem. Chem. Phys.* **2012**, *14*, 1085−1106.
(86) Venturi, S.; Jaffe, R. L.; Panesi, M. Bayesian Machine Learning Approach to the Quantification of Uncertainties on Ab Initio Potential Energy Surfaces. *J. Phys. Chem. A* **2020**, *124*, 5129−5146.
(87) Bose, D.; Candler, G. V. Thermal Rate Constants of the O_2 + N → NO+ O Reaction Based on the $^2A'$ and $^4A'$ Potential-Energy Surfaces. *J. Chem. Phys.* **1997**, *107*, 6136−6145.
(88) Castro-Palacio, J. C.; Nagy, T.; Bemish, R. J.; Meuwly, M. Computational Study of Collisions Between $O(^3P)$ and $NO(^2\Pi)$ at Temperatures Relevant to the Hypersonic Flight Regime. *J. Chem. Phys.* **2014**, *141*, 164319.
(89) San Vicente Veliz, J. C.; Koner, D.; Schwilk, M.; Bemish, R. J.; Meuwly, M. The $N(^4S)+O_2(X^3\Sigma_g^-)\leftrightarrow O(^3P)+NO(X^2\Pi)$Reaction: Thermal and Vibrational Relaxation Rates for the $^2A'$, $^4A'$ and $^2A''$ States. *Phys. Chem. Chem. Phys.* **2020**, *22*, 3927−3939.
(90) Andersson, S.; Marković, N.; Nyman, G. Computational Studies of the Kinetics of the C+ NO and O+ CN Reactions. *J. Phys. Chem. A* **2003**, *107*, 5439−5447.
(91) Veliz, J. C. S. V.; Koner, D.; Schwilk, M.; Bemish, R. J.; Meuwly, M. The $C(^3P)$ + O_2 $(^3\Sigma_g^-)$ ↔ CO_2 ↔ CO $(^1\Sigma^+)$ + O $(^1D)/O(^3P)$ Reaction: Thermal and Vibrational Relaxation Rates from 15 to 20000 K. *Phys. Chem. Chem. Phys.* **2021**.
(92) Boyd, I. D.; Schwartzentruber, T. E. *Nonequilibrium Gas Dynamics and Molecular Simulation*; Cambridge University Press, 2017; Vol. *42*.
(93) Grover, M. S.; Torres, E.; Schwartzentruber, T. E. Direct Molecular Simulation of Internal Energy Relaxation and Dissociation in Oxygen. *Phys. Fluids* **2019**, *31*, 076107.
(94) He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016; pp 770−778.
(95) Glorot, X.; Bengio, Y. Understanding the Difficulty of Training Deep Feedforward Neural Networks. *Proceedings of the 13th International Conference on Artificial Intelligence and Statistics*, 2010; pp 249−256.
(96) Nair, V.; Hinton, G. E. Rectified Linear Units Improve Restricted Boltzmann Machines. *Proceedings of the 27th International Conference on Machine Learning* **2010**, 807−814.
(97) Klambauer, G.; Unterthiner, T.; Mayr, A.; Hochreiter, S. Self-Normalizing Neural Networks. *Advances in Neural Information Processing Systems* **2017**, *30*, 971−980.
(98) Unke, O. T.; Meuwly, M. A Reactive, Scalable, and Transferable Model for Molecular Energies From a Neural Network Approach Based on Local Information. *J. Chem. Phys.* **2018**, *148*, 241708.
(99) Kingma, D.; Ba, J. Adam: A Method for Stochastic Optimization. *arXiv Preprint arXiv:1412.6980* **2014**.
(100) Rasmussen, C. E.; Williams, C. K. I. *Gaussian Processes for Machine Learning*, www.GaussianProcess.org; MIT Press: Cambridge, 2006; Editor: Dietterich, T.
(101) Cui, J.; Krems, R. V. Efficient Non-Parametric Fitting of Potential Energy Surfaces for Polyatomic Molecules With Gaussian Processes. *J. Phys. B: At., Mol. Opt. Phys.* **2016**, *49*, 224001.
(102) Krems, R. V. Bayesian Machine Learning for Quantum Molecular Dynamics. *Phys. Chem. Chem. Phys.* **2019**, *21*, 13392−13410.
(103) Firsov, O. Determination of Forces Acting Between Atoms With the Use of the Differential Cross Section of Elastic Scattering. *Zhur. Eksptl'. I Teoret Fiz.* **1953**, *24*, 279−283.
(104) Buck, U.; Pauly, H. Determination of Intermolecular Potentials by Inversion of Molecular Beam Scattering Data. *J. Chem. Phys.* **1969**, *51*, 1662−1664.
(105) Buck, U. Inversion of Molecular Scattering Data. *Rev. Mod. Phys.* **1974**, *46*, 369−389.
(106) Child, M.; Gerber, R. Inversion of Inelastic Atom-Atom Scattering Data: Recovery of the Interaction Function. *Mol. Phys.* **1979**, *38*, 421−432.
(107) Buck, U. Inversion of Molecular-Scattering Data. *Comput. Phys. Rep.* **1986**, *5*, 1−58.
(108) Boyd, R.; Ho, T.-S.; Rabitz, H. Determination of Multiple Diabatic Potentials by the Inversion of Atom−Atom Scattering Data. *J. Chem. Phys.* **1995**, *103*, 4052−4060.
(109) Vargas-Hernandez, R. A.; Guan, Y.; Zhang, D. H.; Krems, R. V. Bayesian Optimization for the Inverse Scattering Problem in Quantum Reaction Dynamics. *New J. Phys.* **2019**, *21*, 022001.
(110) Aronszajn, N. Theory of Reproducing Kernels. *Trans. Am. Math. Soc.* **1950**, *68*, 337−404.
(111) Soldán, P.; Hutson, J. M. On the Long-Range and Short-Range Behavior of Potentials From Reproducing Kernel Hilbert Space Interpolation. *J. Chem. Phys.* **2000**, *112*, 4415−4416.
(112) Ho, T.; Rabitz, H.; Scoles, G. Reproducing Kernel Technique for Extracting Accurate Potentials From Spectral Data: Potential Curves of the Two Lowest States $X\ ^1\Sigma_g^+$and $a^3\Sigma_u^+$ of the Sodium Dimer. *J. Chem. Phys.* **2000**, *112*, 6218−6227.
(113) Hollebeek, T.; Ho, T.-S.; Rabitz, H. Efficient Potential Energy Surfaces From Partially Filled Ab Initio Data Over Arbitrarily Shaped Regions. *J. Chem. Phys.* **2001**, *114*, 3940−3944.
(114) Hollebeek, T.; Ho, T.-S.; Rabitz, H.; Harding, L. B. Construction of Reproducing Kernel Hilbert Space Potential Energy Surfaces for the $^1A''$ and $^1A'$ States of the Reaction $N(^2D)+H_2$. *J. Chem. Phys.* **2001**, *114*, 3945−3948.
(115) Ho, T.-S.; Rabitz, H. Reproducing Kernel Hilbert Space Interpolation Methods as a Paradigm of High Dimensional Model Representations: Application to Multidimensional Potential Energy Surface Construction. *J. Chem. Phys.* **2003**, *119*, 6433−6442.
(116) Luo, X.; Lu, Z.; Xu, X. Reproducing Kernel Technique for High Dimensional Model Representations (HDMR). *Comput. Phys. Commun.* **2014**, *185*, 3099−3108.
(117) Soloviov, M.; Meuwly, M. Reproducing Kernel Potential Energy Surfaces in Biomolecular Simulations: Nitric Oxide Binding to Myoglobin. *J. Chem. Phys.* **2015**, *143*, 105103.
(118) Koner, D.; Veliz, J. C. S. V.; Bemish, R. J.; Meuwly, M. Accurate Reproducing Kernel-Based Potential Energy Surfaces for the Triplet Ground States of N_2O and Dynamics for the N + NO ⇔ O + N_2 and N_2

https://doi.org/10.1021/acs.chemrev.1c00033
Chem. Rev. XXXX, XXX, XXX−XXX

+ O → 2N + O Reactions. *Phys. Chem. Chem. Phys.* **2020**, *22*, 18488−18498.
(119) Käser, S.; Koner, D.; Christensen, A. S.; von Lilienfeld, O. A.; Meuwly, M. Machine Learning Models of Vibrating H_2CO: Comparing Reproducing Kernels, FCHL, and PhysNet. *J. Phys. Chem. A* **2020**, *124*, 8853−8865.
(120) Koner, D.; Meuwly, M. Permutationally Invariant, Machine-Learned and Kernel-Based Potential Energy Surfaces for Polyatomic Molecules: From Formaldehyde to Acetone. *J. Chem. Theory Comput.* **2020**, *16*, 5474−5484.
(121) Dai, J.; Krems, R. V. Interpolation and Extrapolation of Global Potential Energy Surfaces for Polyatomic Systems by Gaussian Processes With Composite Kernels. *J. Chem. Theory Comput.* **2020**, *16*, 1386−1395.
(122) Vargas-Hernandez, R. A.; Sous, J.; Berciu, M.; Krems, R. V. Extrapolating Quantum Observables With Machine Learning: Inferring Multiple Phase Transitions From Properties of a Single Phase. *Phys. Rev. Lett.* **2018**, *121*, 255702.
(123) Deng, Z.; Tutunnikov, I.; Averbukh, I. S.; Thachuk, M.; Krems, R. Bayesian optimization for inverse problems in time-dependent quantum dynamics. *J. Chem. Phys.* **2020**, *153*, 164111.
(124) Sugisawa, H.; Ida, T.; Krems, R. Gaussian Process Model of 51-Dimensional Potential Energy Surface for Protonated Imidazole Dimer. *J. Chem. Phys.* **2020**, *153*, 114101.
(125) London, F. Quantum Mechanical Interpretation of the Process of Activation. *Z. Elektrochem.* **1929**, *35*, 552−555.
(126) Eyring, H.; Polanyi, M. Concerning Simple Gas Reactions. *Z. Phys. Chem. Abt. B* **1931**, *12*, 279−311.
(127) Sato, S. Potential Energy Surface of the System of Three Atoms. *J. Chem. Phys.* **1955**, *23*, 2465−2466.
(128) Shepler, B. C.; Braams, B. J.; Bowman, J. M. Quasiclassical Trajectory Calculations of Acetaldehyde Dissociation on a Global Potential Energy Surface Indicate Significant Non-Transition State Dynamics. *J. Phys. Chem. A* **2007**, *111*, 8282−8285.
(129) Käser, S.; Unke, O. T.; Meuwly, M. Isomerization and Decomposition Reactions of Acetaldehyde Relevant to Atmospheric Processes From Dynamics Simulations on Neural Network-Based Potential Energy Surfaces. *J. Chem. Phys.* **2020**, *152*, 214304.
(130) Hutson, J. M. Intermolecular Forces From the Spectroscopy of Van Der Waals Molecules. *Annu. Rev. Phys. Chem.* **1990**, *41*, 123−154.
(131) Hutson, J. M. An Introduction to the Dynamics of Van Der Waals Molecules. *Adv. Mol. Vibrat. Coll. Dyn.* **1991**, *1*, 1−45.
(132) Koner, D.; San Vicente Veliz, J. C.; van der Avoird, A.; Meuwly, M. Near Dissociation States for H_2^+−He on MRCI and FCI Potential Energy Surfaces. *Phys. Chem. Chem. Phys.* **2019**, *21*, 24976−24983.
(133) Meuwly, M.; Hutson, J. Morphing Ab Initio Potentials: A Systematic Study of Ne-HF. *J. Chem. Phys.* **1999**, *110*, 8338−8347.
(134) Karman, T.; Van Der Avoird, A.; Groenenboom, G. C. Potential Energy and Dipole Moment Surfaces of the Triplet States of the O_2 ($X^3\Sigma_g^-$) - O_2 ($X^3\Sigma_g^-$, $a^1\Delta_g$, $B^1\Sigma_g^+$) Complex. *J. Chem. Phys.* **2017**, *147*, 084306.
(135) Franke, R.; Nielson, G. Smooth Interpolation of Large Sets of Scattered Data. *Int. J. Numer. Meth. Eng.* **1980**, *15*, 1691−1704.
(136) Nguyen, K. A.; Rossi, I.; Truhlar, D. G. A Dual-Level Shepard Interpolation Method for Generating Potential Energy Surfaces for Dynamics Calculations. *J. Chem. Phys.* **1995**, *103*, 5522−5530.
(137) Bettens, R. P.; Collins, M. A. Learning to Interpolate Molecular Potential Energy Surfaces With Confidence: A Bayesian Approach. *J. Chem. Phys.* **1999**, *111*, 816−826.
(138) Lancaster, P.; Salkauskas, K. Surfaces Generated by Moving Least Squares Methods. *Math. Comp.* **1981**, *37*, 141−158.
(139) Ischtwan, J.; Collins, M. A. Molecular Potential Energy Surfaces by Interpolation. *J. Chem. Phys.* **1994**, *100*, 8080−8088.
(140) Dawes, R.; Thompson, D. L.; Wagner, A. F.; Minkoff, M. Interpolating Moving Least-Squares Methods for Fitting Potential Energy Surfaces: A Strategy for Efficient Automatic Data Point Placement in High Dimensions. *J. Chem. Phys.* **2008**, *128*, 084107.
(141) Cassam-Chenaï, P.; Patras, F. Symmetry-Adapted Polynomial Basis for Global Potential Energy Surfaces-Applications to XY_4 Molecules. *J. Math. Chem.* **2008**, *44*, 938−966.
(142) Braams, B. J.; Bowman, J. M. Permutationally Invariant Potential Energy Surfaces in High Dimensionality. *Int. Rev. Phys. Chem.* **2009**, *28*, 577−606.
(143) Paukku, Y.; Yang, K. R.; Varga, Z.; Truhlar, D. G. Global Ab Initio Ground-State Potential Energy Surface of N_4. *J. Chem. Phys.* **2013**, *139*, 044309.
(144) Li, J.; Zhao, B.; Xie, D.; Guo, H. Advances and New Challenges to Bimolecular Reaction Dynamics Theory. *J. Phys. Chem. Lett.* **2020**, *11*, 8844−8860.
(145) Olejnicak, J.; Candler, G. Vibrational-Energy Conservation With Vibration-Dissociation Coupling - General Theory and Numberical Studies. *Phys. Fluids* **1995**, *7*, 1764−1774.
(146) Sarma, G. Physico-Chemical Modelling in Hypersonic Flow Simulation. *Progr. Aerospace Sci.* **2000**, *36*, 281−349.
(147) Bertin, J.; Cummings, R. Fifty Years of Hypersonics: Where We've Been, Where We're Going. *Prog. Aerospace Sci.* **2003**, *39*, 511−536.
(148) Knight, D.; Longo, J.; Drikakis, D.; Gaitonde, D.; Lani, A.; Nompelis, I.; Reimann, B.; Walpot, L. Assessment of CFD Capability for Prediction of Hypersonic Shock Interactions. *Progr. Aerospace Sci.* **2012**, *48*−*49*, 8−26.
(149) Leyva, I. the Relentless Pursuit of Hypersonic Flight. *Phys. Today* **2017**, *70*, 30−36.
(150) Wakelam, V.; Smith, I.; Herbst, E.; Troe, J.; Geppert, W.; Linnartz, H.; Öberg, K.; Roueff, E.; Agúndez, M.; Pernot, P.; et al. Reaction Networks for Interstellar Chemical Modelling: Improvements and Challenges. *Space Sci. Rev.* **2010**, *156*, 13−72.
(151) Balucani, N. Elementary Reactions and Their Role in Gas-Phase Prebiotic Chemistry. *Int. J. Mol. Sci.* **2009**, *10*, 2304−2335.
(152) Shaw, M. F.; Sztáray, B.; Whalley, L. K.; Heard, D. E.; Millet, D. B.; Jordan, M. J.; Osborn, D. L.; Kable, S. H. Photo-Tautomerization of Acetaldehyde as a Photochemical Source of Formic Acid in the Troposphere. *Nat. Commun.* **2018**, *9*, 1−7.
(153) So, S.; Wille, U.; Da Silva, G. Atmospheric Chemistry of Enols: A Theoretical Study of the Vinyl Alcohol + OH + O_2 Reaction Mechanism. *Environ. Sci. Technol.* **2014**, *48*, 6694−6701.
(154) Archibald, A. T.; McGillen, M. R.; Taatjes, C. A.; Percival, C. J.; Shallcross, D. E. Atmospheric Transformation of Enols: A Potential Secondary Source of Carboxylic Acids in the Urban Troposphere. *Geophys. Res. Lett.* **2007**, *34*, L21801.
(155) Andrews, D. U.; Heazlewood, B. R.; Maccarone, A. T.; Conroy, T.; Payne, R. J.; Jordan, M. J. T.; Kable, S. H. Photo-Tautomerization of Acetaldehyde to Vinyl Alcohol: A Potential Route to Tropospheric Acids. *Science* **2012**, *337*, 1203−1206.
(156) Clubb, A. E.; Jordan, M. J. T.; Kable, S. H.; Osborn, D. L. Phototautomerization of Acetaldehyde to Vinyl Alcohol: A Primary Process in UV-Irradiated Acetaldehyde From 295 to 335 Nm. *J. Phys. Chem. Lett.* **2012**, *3*, 3522−3526.
(157) Millet, D. B.; Baasandorj, M.; Farmer, D. K.; Thornton, J. A.; Baumann, K.; Brophy, P.; Chaliyakunnel, S.; De Gouw, J. A.; Graus, M.; Hu, L.; et al. A Large and Ubiquitous Source of Atmospheric Formic Acid. *Atmos. Chem. Phys.* **2015**, *15*, 6283−6304.
(158) Reyes, J. Y.; Nagy, T.; Meuwly, M. Competitive Reaction Pathways in Vibrationally Induced Photodissociation of H_2SO_4. *Phys. Chem. Chem. Phys.* **2014**, *16*, 18533−18544.
(159) Braams, B. J.; Bowman, J. M. Permutationally Invariant Potential Energy Surfaces in High Dimensionality. *Int. Rev. Phys. Chem.* **2009**, *28*, 577−606.
(160) Jiang, B.; Li, J.; Guo, H. Potential Energy Surfaces From High Fidelity Fitting of Ab Initio Points: The Permutation Invariant Polynomial-Neural Network Approach. *Int. Rev. Phys. Chem.* **2016**, *35*, 479−506.
(161) Heazlewood, B. R.; Jordan, M. J.; Kable, S. H.; Selby, T. M.; Osborn, D. L.; Shepler, B. C.; Braams, B. J.; Bowman, J. M. Roaming Is the Dominant Mechanism for Molecular Products in Acetaldehyde

https://doi.org/10.1021/acs.chemrev.1c00033
Chem. Rev. XXXX, XXX, XXX−XXX

Photodissociation. *Proc. Natl. Acad. Sci. U. S. A.* **2008**, *105*, 12719−12724.
(162) Han, Y.-C.; Shepler, B. C.; Bowman, J. M. Quasiclassical Trajectory Calculations of the Dissociation Dynamics of CH_3CHO at High Energy Yield Many Products. *J. Phys. Chem. Lett.* **2011**, *2*, 1715−1719.
(163) Li, J.; Chen, J.; Zhang, D. H.; Guo, H. Quantum and Quasi-Classical Dynamics of the OH+ CO ⇒H+ CO_2 Reaction on a New Permutationally Invariant Neural Network Potential Energy Surface. *J. Chem. Phys.* **2014**, *140*, 044327.
(164) Francisco, J. S.; Muckerman, J. T.; Yu, H.-G. HOCO Radical Chemistry. *Acc. Chem. Res.* **2010**, *43*, 1519−1526.
(165) Wagner, A. F.; Dawes, R.; Continetti, R. E.; Guo, H. Theoretical/Experimental Comparison of Deep Tunneling Decay of Quasi-Bound H (D) OCO to H (D)+ CO_2. *J. Chem. Phys.* **2014**, *141*, 054304.
(166) Jiang, B.; Guo, H. Six-Dimensional Quantum Dynamics for Dissociative Chemisorption of H_2 and D_2 on Ag (111) on a Permutation Invariant Potential Energy Surface. *Phys. Chem. Chem. Phys.* **2014**, *16*, 24704−24715.
(167) Jiang, B.; Hu, X.; Lin, S.; Xie, D.; Guo, H. Six-Dimensional Quantum Dynamics of Dissociative Chemisorption of H_2 on Co (0001) on an Accurate Global Potential Energy Surface. *Phys. Chem. Chem. Phys.* **2015**, *17*, 23346−23355.
(168) Jiang, B.; Guo, H. Dynamics of Water Dissociative Chemisorption on Ni (111): Effects of Impact Sites and Incident Angles. *Phys. Rev. Lett.* **2015**, *114*, 166101.
(169) Jiang, B.; Guo, H. Communication: Enhanced Dissociative Chemisorption of CO_2 via Vibrational Excitation. *J. Chem. Phys.***2016**, *144*, 091101.
(170) Weichman, M. L.; DeVine, J. A.; Babin, M. C.; Li, J.; Guo, L.; Ma, J.; Guo, H.; Neumark, D. M. Feshbach Resonances in the Exit Channel of the F+ CH_3OH ⇒ HF + CH_3O Reaction Observed Using Transition-State Spectroscopy. *Nat. Chem.* **2017**, *9*, 950.
(171) Lu, D.; Li, J.; Guo, H. Comprehensive Investigations of the Cl + CH_3OH ⇒HCl + CH_3O/CH_2OH Reaction: Validation of Experiment and Dynamic Insights. *CCS Chemistry* **2020**, *2*, 882−894.
(172) Sweeny, B. C.; Pan, H.; Kassem, A.; Sawyer, J. C.; Ard, S. G.; Shuman, N. S.; Viggiano, A. A.; Brickel, S.; Unke, O. T.; Upadhyay, M.; et al. Thermal Activation of Methane by MgO+: Temperature Dependent Kinetics, Reactive Molecular Dynamics Simulations and Statistical Modeling. *Phys. Chem. Chem. Phys.* **2020**, *22*, 8913−8923.
(173) Kidwell, N. M.; Li, H.; Wang, X.; Bowman, J. M.; Lester, M. I. Unimolecular Dissociation Dynamics of Vibrationally Activated CH_3CHOO Criegee Intermediates to OH Radical Products. *Nat. Chem.* **2016**, *8*, 509−514.
(174) Qu, C.; Conte, R.; Houston, P. L.; Bowman, J. M. Full-Dimensional Potential Energy Surface for Acetylacetone and Tunneling Splittings. *Phys. Chem. Chem. Phys.* **2021**, *23*, 7758
(175) Ramakrishnan, R.; Dral, P. O.; Rupp, M.; Von Lilienfeld, O. A. Big Data Meets Quantum Chemistry Approximations: The Δ-Machine Learning Approach. *J. Chem. Theory Comput.* **2015**, *11*, 2087−2096.
(176) Nandi, A.; Qu, C.; Houston, P. L.; Conte, R.; Bowman, J. M. Δ-Machine Learning for Potential Energy Surfaces: A PIP Approach to Bring a DFT-Based PES to CCSD (T) Level of Theory. *J. Chem. Phys.* **2021**, *154*, 051102.
(177) Stöhr, M.; Medrano Sandonas, L.; Tkatchenko, A. Accurate Many-Body Repulsive Potentials for Density-Functional Tight Binding From Deep Tensor Neural Networks. *J. Phys. Chem. Lett.* **2020**, *11*, 6835−6843.
(178) Pan, S. J.; Yang, Q. A Survey on Transfer Learning. *IEEE Transactions on Knowledge and Data Engineering* **2010**, *22*, 1345−1359.
(179) Smith, J. S.; Nebgen, B. T.; Zubatyuk, R.; Lubbers, N.; Devereux, C.; Barros, K.; Tretiak, S.; Isayev, O.; Roitberg, A. E. Approaching Coupled Cluster Accuracy With a General-Purpose Neural Network Potential Through Transfer Learning. *Nat. Commun.* **2019**, *10*, 1−8.
(180) Soloviov, M.; Das, A. K.; Meuwly, M. Structural Interpretation of Metastable States in Myoglobin-NO. *Angew. Chem., Int. Ed.* **2016**, *55*, 10126−10130.
(181) Tantillo, D. J.; Jiangang, C.; Houk, K. N. Theozymes and Compuzymes: Theoretical Models for Biological Catalysis. *Curr. Opin. Chem. Biol.* **1998**, *2*, 743−750.
(182) Kiss, G.; Çelebi-Ölçüm, N.; Moretti, R.; Baker, D.; Houk, K. Computational Enzyme Design. *Angew. Chem., Int. Ed.* **2013**, *52*, 5700−5725.
(183) Otten, R.; Pádua, R. A.; Bunzel, H. A.; Nguyen, V.; Pitsawong, W.; Patterson, M.; Sui, S.; Perry, S. L.; Cohen, A. E.; Hilvert, D.; et al. How Directed Evolution Reshapes the Energy Landscape in an Enzyme to Boost Catalysis. *Science* **2020**, *370*, 1442−1446.
(184) Privett, H. K.; Kiss, G.; Lee, T. M.; Blomberg, R.; Chica, R. a.; Thomas, L. M.; Hilvert, D.; Houk, K. N.; Mayo, S. L. Iterative Approach to Computational Enzyme Design. *Proc. Natl. Acad. Sci. U. S. A.* **2012**, *109*, 3790−3795.
(185) Bunzel, H. A.; Kries, H.; Marchetti, L.; Zeymer, C.; Mittl, P. R. E.; Mulholland, A. J.; Hilvert, D. Emergence of a Negative Activation Heat Capacity During Evolution of a Designed Enzyme. *J. Am. Chem. Soc.* **2019**, *141*, 11745−11748.
(186) Bunzel, H. A.; Anderson, J. R.; Mulholland, A. J. Designing Better Enzymes: Insights From Directed Evolution. *Curr. Opin. Struct. Biol.* **2021**, *67*, 212−218.
(187) Bonk, B. M.; Weis, J. W.; Tidor, B. Machine Learning Identifies Chemical Characteristics That Promote Enzyme Catalysis. *J. Am. Chem. Soc.* **2019**, *141*, 4108−4118.
(188) Wei, J. N.; Duvenaud, D.; Aspuru-Guzik, A. Neural Networks for the Prediction of Organic Chemistry Reactions. *ACS Cent. Sci.* **2016**, *2*, 725−732.
(189) Segler, M. H.; Preuss, M.; Waller, M. P. Planning Chemical Syntheses With Deep Neural Networks and Symbolic AI. *Nature* **2018**, *555*, 604−610.
(190) Filipa De Almeida, A.; Moreira, R.; Rodrigues, T. Synthetic Organic Chemistry Driven by Artificial Intelligence. *Nat. Rev. Chem.* **2019**, *3*, 589−604.
(191) Zhou, Z.; Li, X.; Zare, R. N. Optimizing Chemical Reactions With Deep Reinforcement Learning. *ACS Cent. Sci.* **2017**, *3*, 1337−1344.
(192) Gao, H.; Struble, T. J.; Coley, C. W.; Wang, Y.; Green, W. H.; Jensen, K. F. Using Machine Learning to Predict Suitable Conditions for Organic Reactions. *ACS Cent. Sci.* **2018**, *4*, 1465−1476.
(193) Allen, F.; Pon, A.; Greiner, R.; Wishart, D. Computational Prediction of Electron Ionization Mass Spectra to Assist in GC/MS Compound Identification. *Anal. Chem.* **2016**, *88*, 7689−7697.
(194) Dührkop, K.; Nothias, L.-F.; Fleischauer, M.; Reher, R.; Ludwig, M.; Hoffmann, M. A.; Petras, D.; Gerwick, W. H.; Rousu, J.; Dorrestein, P. C.; et al. Systematic Classification of Unknown Metabolites Using High-Resolution Fragmentation Mass Spectra. *Nat. Biotechnol.* **2021**, *39*, 1−10.
(195) Vléduts, G.; Finn, V. Creating a Machine Language for Organic Chemistry. *Inf. Storage Retr.* **1963**, *1*, 101−116.
(196) Corey, E. J. General Methods for the Construction of Complex Molecules. *Pure Appl. Chem.* **1967**, *14*, 19−38.
(197) Corey, E.; Wipke, W. Computer-Assisted Design of Complex Organic Syntheses. *Science* **1969**, *166*, 178−192.
(198) Hendrickson, J. B. Systematic Characterization of Structures and Reactions for Use in Organic Synthesis. *J. Am. Chem. Soc.* **1971**, *93*, 6847−6854.
(199) Gelernter, H.; Sridharan, N. S.; Hart, A. J.; Yen, S.-C.; Fowler, F. W.; Shue, H.-J. *New Concepts I*; Springer, 1973; pp 113−150.
(200) Gelernter, H.; Sanders, A.; Larsen, D.; Agarwal, K.; Boivie, R.; Spritzer, G.; Searleman, J. Empirical Explorations of SYNCHEM. *Science* **1977**, *197*, 1041−1049.
(201) Salatin, T. D.; Jorgensen, W. L. Computer-Assisted Mechanistic Evaluation of Organic Reactions. 1. Overview. *J. Org. Chem.* **1980**, *45*, 2043−2051.
(202) Lederberg, J. Topological Mapping of Organic Molecules. *Proc. Natl. Acad. Sci. U. S. A.* **1965**, *53*, 134−139.

https://doi.org/10.1021/acs.chemrev.1c00033
Chem. Rev. XXXX, XXX, XXX−XXX

(203) Lindsay, R. K.; Buchanan, B. G.; Feigenbaum, E. a.; Lederberg, J. DENDRAL: A Case Study of the First Expert System for Scientific Hypothesis Formation. *Artif. Intell.* **1993**, *61*, 209−261.
(204) Lederberg, J.; Sutherland, G. L.; Buchanan, B. G.; Feigenbaum, E. A.; Robertson, A. v.; Duffield, A. M.; Djerassi, C. Applications of Artificial Intelligence for Chemical Inference. I. Number of Possible Organic Compounds. Acyclic Structures Containing Carbon, Hydrogen, Oxygen, and Nitrogen. *J. Am. Chem. Soc.* **1969**, *91*, 2973−2976.
(205) Jorgensen, W. L.; Laird, E. R.; Gushurst, A. J.; Fleischer, J. M.; Gothe, S. A.; Helson, H. E.; Paderes, G. D.; Sinclair, S. CAMEO: A Program for the Logical Prediction of the Products of Organic Reactions. *Pure Appl. Chem.* **1990**, *62*, 1921−1932.
(206) Gasteiger, J.; Hutchings, M. G.; Christoph, B.; Gann, L.; Hiller, C.; Löw, P.; Marsili, M.; Saller, H.; Yuki, K. *Organic Synthesis, Reactions and Mechanisms*; Springer, 1987; pp 19−73.
(207) Kayala, M. A.; Baldi, P. ReactionPredictor: Prediction of Complex Chemical Reactions at the Mechanistic Level Using Machine Learning. *J. Chem. Inf. Model.* **2012**, *52*, 2526−2540.
(208) Klucznik, T.; Mikulak-Klucznik, B.; McCormack, M. P.; Lima, H.; Szymkuć, S.; Bhowmick, M.; Molga, K.; Zhou, Y.; Rickershauser, L.; Gajewska, E. P.; et al. Efficient Syntheses of Diverse, Medicinally Relevant Targets Planned by Computer and Executed in the Laboratory. *Chem.* **2018**, *4*, 522−532.
(209) Skoraczyński, G.; Dittwald, P.; Miasojedow, B.; Szymkuć, S.; Gajewska, E.; Grzybowski, B. A.; Gambin, A. Predicting the Outcomes of Organic Reactions via Machine Learning: Are Current Descriptors Sufficient? *Sci. Rep.* **2017**, *7*, 1−9.
(210) Szymkuć, S.; Gajewska, E. P.; Klucznik, T.; Molga, K.; Dittwald, P.; Startek, M.; Bajczyk, M.; Grzybowski, B. A. Computer-Assisted Synthetic Planning: The End of the Beginning. *Angew. Chem., Int. Ed.* **2016**, *55*, 5904−5937.
(211) 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.
(212) Huang, B.; Von Lilienfeld, O. A. Quantum Machine Learning in Chemical Compound Space. *arXiv Preprint arXiv:2012.07502* **2020**.
(213) Lim, H.; Jung, Y. Delfos: Deep Learning Model for Prediction of Solvation Free Energies in Generic Organic Solvents. *Chem. Sci.* **2019**, *10*, 8306−8315.
(214) Basdogan, Y.; Groenenboom, M. C.; Henderson, E.; De, S.; Rempe, S. B.; Keith, J. A. Machine Learning-Guided Approach for Studying Solvation Environments. *J. Chem. Theory Comput.* **2020**, *16*, 633−642.
(215) Rauer, C.; Bereau, T. Hydration Free Energies From Kernel-Based Machine Learning: Compound-Database Bias. *J. Chem. Phys.* **2020**, *153*, 014101.
(216) Scheen, J.; Wu, W.; Mey, A. S.; Tosco, P.; Mackey, M.; Michel, J. Hybrid Alchemical Free Energy/Machine-Learning Methodology for the Computation of Hydration Free Energies. *J. Chem. Inf. Model.* **2020**, *60*, 5331−5339.
(217) Steiner, S.; Wolf, J.; Glatzel, S.; Andreou, A.; Granda, J. M.; Keenan, G.; Hinkley, T.; Aragon-Camarasa, G.; Kitson, P. J.; Angelone, D. et al. Organic Synthesis in a Modular Robotic System Driven by a Chemical Programming Language. *Science* **2019**, *363*, eaav2211
(218) Coley, C. W.; Thomas, D. a.; Lummiss, J. A.; Jaworski, J. N.; Breen, C. P.; Schultz, V.; Hart, T.; Fishman, J. S.; Rogers, L.; Gao, H.; et al. A Robotic Platform for Flow Synthesis of Organic Compounds Informed by AI Planning. *Science* **2019**, *365*, No. eaax1566.
(219) Roch, L. M.; Häse, F.; Kreisbeck, C.; Tamayo-Mendoza, T.; Yunker, L. P.; Hein, J. E.; Aspuru-Guzik, A. ChemOS: Orchestrating Autonomous Experimentation. *Science Robotics* **2018**, *3*, eaat5559
(220) Flores-Leonar, M. M.; Mejía-Mendoza, L. M.; Aguilar-Granda, A.; Sanchez-Lengeling, B.; Tribukait, H.; Amador-Bedolla, C.; Aspuru-Guzik, A. Materials Acceleration Platforms: On the Way to Autonomous Experimentation. *Curr. Opin. Green Sust. Chem.* **2020**, *25*, 100370.
(221) Hase, F.; Roch, L. M.; Kreisbeck, C.; Aspuru-Guzik, A. Phoenics: A Bayesian Optimizer for Chemistry. *ACS Cent. Sci.* **2018**, *4*, 1134−1145.
(222) Häse, F.; Roch, L. M.; Aspuru-Guzik, A. Gryffin: An Algorithm for Bayesian Optimization for Categorical Variables Informed by Physical Intuition With Applications to Chemistry. *arXiv Preprint arXiv:2003.12127* **2020**.
(223) Wilson, Z. T.; Sahinidis, N. V. Automated Learning of Chemical Reaction Networks. *Comput. Chem. Eng.* **2019**, *127*, 88−98.
(224) Maddison, C. J.; Mnih, A.; Teh, Y. W. the Concrete Distribution: A Continuous Relaxation of Discrete Random Variables. *arXiv Preprint arXiv:1611.00712* **2016**.
(225) Jang, E.; Gu, S.; Poole, B. Categorical Reparameterization With Gumbel-Softmax. *arXiv Preprint arXiv:1611.01144* **2016**.
(226) Christensen, M.; Yunker, L.; Adedeji, F.; Häse, F.; Roch, L.; Gensch, T.; dos Passos Gomes, G.; Zepel, T.; Sigman, M.; Aspuru-Guzik, A. et al. *Data-science driven autonomous process optimization*; 2020; DOI: 10.26434/chemrxiv.13146404.v2.
(227) Gasteiger, J.; Hanebeck, W.; Schulz, K.-P. Prediction of Mass Spectra From Structural Information. *J. Chem. Inf. Com. Sci.* **1992**, *32*, 264−271.
(228) Allen, F.; Pon, A.; Wilson, M.; Greiner, R.; Wishart, D. CFM-ID: A Web Server for Annotation, Spectrum Prediction and Metabolite Identification From Tandem Mass Spectra. *Nucleic Acids Res.* **2014**, *42*, W94−w99.
(229) Dührkop, K.; Shen, H.; Meusel, M.; Rousu, J.; Böcker, S. Searching Molecular Structure Databases With Tandem Mass Spectra Using CSI: FingerID. *Proc. Natl. Acad. Sci. U. S. A.* **2015**, *112*, 12580−12585.
(230) Feunang, Y. D.; Eisner, R.; Knox, C.; Chepelev, L.; Hastings, J.; Owen, G.; Fahy, E.; Steinbeck, C.; Subramanian, S.; Bolton, E.; et al. ClassyFire: Automated Chemical Classification With a Comprehensive, Computable Taxonomy. *J. Cheminf.* **2016**, *8*, 61.
(231) Heinonen, M.; Shen, H.; Zamboni, N.; Rousu, J. Metabolite Identification and Molecular Fingerprint Prediction Through Machine Learning. *Bioinformatics* **2012**, *28*, 2333−2341.
(232) Shen, H.; Dührkop, K.; Böcker, S.; Rousu, J. Metabolite Identification Through Multiple Kernel Learning on Fragmentation Trees. *Bioinformatics* **2014**, *30*, i157−i164.
(233) Böcker, S.; Dührkop, K. Fragmentation Trees Reloaded. *J. Cheminf.* **2016**, *8*, 5.
(234) Matthews, B. W. Comparison of the Predicted and Observed Secondary Structure of T4 Phage Lysozyme. *Biochim. Biophys. Acta, Protein Struct.* **1975**, *405*, 442−451.
(235) Gillespie, D. T. Stochastic Simulation of Chemical Kinetics. *Annu. Rev. Phys. Chem.* **2007**, *58*, 35−55.
(236) Wang, L.-P.; Titov, A.; McGibbon, R.; Liu, F.; Pande, V. S.; Martínez, T. J. Discovering Chemistry With an Ab Initio Nanoreactor. *Nat. Chem.* **2014**, *6*, 1044−1048.
(237) Zeng, J.; Cao, L.; Xu, M.; Zhu, T.; Zhang, J. Z. Complex Reaction Processes in Combustion Unraveled by Neural Network-Based Molecular Dynamics Simulation. *Nat. Commun.* **2020**, *11*, 1−9.
(238) Zhang, L.; Han, J.; Wang, H.; Car, R.; Weinan, E. Deep Potential Molecular Dynamics: A Scalable Model With the Accuracy of Quantum Mechanics. *Phys. Rev. Lett.* **2018**, *120*, 143001.
(239) Stocker, S.; Csanyi, G.; Reuter, K.; Margraf, J. T. Machine Learning in Chemical Reaction Space. *Nat. Commun.* **2020**, *11*, 5505.
(240) Bartók, A. P.; Kondor, R.; Csányi, G. On Representing Chemical Environments. *Phys. Rev. B: Condens. Matter Mater. Phys.* **2013**, *87*, 184115.
(241) Cozad, A.; Sahinidis, N. v.; Miller, D. C. Learning Surrogate Models for Simulation-Based Optimization. *AIChE J.* **2014**, *60*, 2211−2227.
(242) Loeppky, J. L.; Sacks, J.; Welch, W. J. Choosing the Sample Size of a Computer Experiment: A Practical Guide. *Technometrics* **2009**, *51*, 366−376.
(243) Zeng, J.; Zhang, L.; Wang, H.; Zhu, T. Exploring the Chemical Space of Linear Alkane Pyrolysis via Deep Potential GENerator. *Energy Fuels* **2021**, *35*, 762−769.
(244) Kennedy, M. C.; O'Hagan, A. Bayesian Calibration of Computer Models. *J. R. Stat. Soc.: Ser. B* **2001**, *63*, 425−464.

https://doi.org/10.1021/acs.chemrev.1c00033
Chem. Rev. XXXX, XXX, XXX−XXX

(245) Heo, H.; Ho, T.-S.; Lehmann, K. K.; Rabitz, H. Regularized Inversion of Diatomic Vibration−Rotation Spectral Data: A Functional Sensitivity Analysis Approach. *J. Chem. Phys.* **1992**, *97*, 852−861.
(246) Gazdy, B.; Bowman, J. An Adjusted Global Potential Surface for HCN Based on Rigorous Vibrational Calculations. *J. Chem. Phys.* **1991**, *95*, 6309−6316.
(247) McIntosh, a.; Wang, Z.; Castillo-Chara, J.; Lucchese, R.; Bevan, J.; Suenram, R.; Legon, A. The Structure and Ground State Dynamics of Ar-IH. *J. Chem. Phys.* **1999**, *111*, 5764−5770.
(248) Lorenz, K.; Westley, M.; Chandler, D. Rotational State-to-State Differential Cross Sections for the HCl-Ar Collision System Using Velocity-Mapped Ion Imaging. *Phys. Chem. Chem. Phys.* **2000**, *2*, 481−494.
(249) Xu, Y.; Jager, W. The Dynamics of the CO-N_2 Interaction: Strong Coriolis Coupling in CO-paraN_2. *J. Chem. Phys.* **2000**, *113*, 514−524.
(250) Castillo-Chara, J.; Lucchese, R.; Bevan, J. Differentiation of the Ground Vibrational and Global Minimum Structures in the Ar: HBr Intermolecular Complex. *J. Chem. Phys.* **2001**, *115*, 899−911.
(251) Kerenskaya, G.; Kaledin, A.; Heaven, M. Potential Energy Surfaces for CH($A\,^2\Delta$)-Ar and Analysis of the $a\,^2\Delta$ $X^2\Pi$ Band System. *J. Chem. Phys.* **2001**, *115*, 2123−2133.
(252) Van Mourik, T.; Harris, G.; Polyansky, O.; Tennyson, J.; Csaszar, A.; Knowles, P. Ab Initio Global Potential, Dipole, Adiabatic, and Relativistic Correction Surfaces for the HCN-HNC System. *J. Chem. Phys.* **2001**, *115*, 3706−3718.
(253) Shroll, R.; Lohr, L.; Barker, J. Empirical Potentials for Rovibrational Energy Transfer of Hydrogen Fluoride in Collisions With Argon. *J. Chem. Phys.* **2001**, *115*, 4573−4585.
(254) Howson, J.; Hutson, J. Morphing the He-OCS Intermolecular Potential. *J. Chem. Phys.* **2001**, *115*, 5059−5065.
(255) Shirin, S.; Polyansky, O.; Zobov, N.; Barletta, P.; Tennyson, J. Spectroscopically Determined Potential Energy Surface of (H_2O)-O^{16} Up to 25 000 cm^{-1}. *J. Chem. Phys.* **2003**, *118*, 2124−2129.
(256) Xu, D.; Guo, H.; Zou, S.; Bowman, J. A Scaled Ab Initio Potential Energy Surface for Acetylene and Vinylidene. *Chem. Phys. Lett.* **2003**, *377*, 582−588.
(257) Bowman, J.; Xantheas, S. Morphing" of Ab Initio-Based Interaction Potentials to Spectroscopic Accuracy: Application to Cl-(H_2O). *Pure Appl. Chem.* **2004**, *76*, 29−35.
(258) McElmurry, B.; Lucchese, R.; Bevan, J.; Belov, S. Analysis of the Submillimetre Ar: HI Sigma Bending Transition as a Test of a Morphed Potential. *Phys. Chem. Chem. Phys.* **2004**, *6*, 5318−5323.
(259) Wang, Z.; Lucchese, R.; Bevan, J. A Kr-BrH Global Minimum Structure Determined on the Basis of Potential Morphing. *J. Phys. Chem. A* **2004**, *108*, 2884−2892.
(260) Rivera-Rivera, L. A.; Lucchese, R. R.; Bevan, J. W. A Parameterized Compound-Model Chemistry for Morphing the Intermolecular Potential of OC-HCl. *Chem. Phys. Lett.* **2008**, *460*, 352−358.
(261) Spirko, V. Morphing Ab Initio Potential Energy Curve of Beryllium Monohydride. *J. Mol. Spectrosc.* **2016**, *330*, 89−95.
(262) Yurchenko, S. N.; Lodi, L.; Tennyson, J.; Stolyarov, A. V. Duo: A General Program for Calculating Spectra of Diatomic Molecules. *Comput. Phys. Commun.* **2016**, *202*, 262−275.
(263) Augustovicova, L. D.; Spirko, V. Morphing Radial Molecular Property Functions of Hydroxyl. *J. Quant. Spectrosc. Radiat. Transfer* **2020**, *254*, 107211.
(264) Sebald, P.; Stein, C.; Oswald, R.; Botschwina, P. Rovibrational States of N_3^--and CO_2 Up to High *J*: A Theoretical Study Beyond Fc-CCSD(T). *J. Phys. Chem. A* **2013**, *117*, 13806−13814.
(265) Evangelista, F. A. Perspective: Multireference Coupled Cluster Theories of Dynamical Electron Correlation. *J. Chem. Phys.* **2018**, *149*, 030901.
(266) Denis-Alpizar, O.; Bemish, R. J.; Meuwly, M. Reactive Collisions for NO($^2\Pi$) + N(^{4}S) at Temperatures Relevant to the Hypersonic Flight Regime. *Phys. Chem. Chem. Phys.* **2017**, *19*, 2392−2401.
(267) Westermayr, J.; Gastegger, M.; Marquetand, P. Combining SchNet and SHARC: The SchNarc Machine Learning Approach for Excited-State Dynamics. *J. Phys. Chem. Lett.* **2020**, *11*, 3828−3834.
(268) Tully, J. C. Perspective: Nonadiabatic Dynamics Theory. *J. Chem. Phys.* **2012**, *137*, 22a301.
(269) Guo, H.; Yarkony, D. R. Accurate Nonadiabatic Dynamics. *Phys. Chem. Chem. Phys.* **2016**, *18*, 26335−26352.
(270) Fink, T.; Reymond, J.-L. Virtual Exploration of the Chemical Universe Up to 11 Atoms of C, N, O, F: Assembly of 26.4 Million Structures (110.9 Million Stereoisomers) and Analysis for New Ring Systems, Stereochemistry, Physicochemical Properties, Compound Classes, and Drug Discovery. *J. Chem. Inf. Model.* **2007**, *47*, 342−353.
(271) Senior, A. W.; Evans, R.; Jumper, J.; Kirkpatrick, J.; Sifre, L.; Green, T.; Qin, C.; Zidek, A.; Nelson, A. W. R.; Bridgland, A.; et al. Improved Protein Structure Prediction Using Potentials From Deep Learning. *Nature* **2020**, *577*, 706−710.
(272) *AlphaFold: A Solution to a 50-Year-Old Grand Challenge in Biology*. https://deepmind.com/blog/article/Alphafold-a-Solution-to-a-50-Year-Old-Grand-Challenge-in-Biology/, Accessed: 2020-12-29.
(273) Karim, A.; Mishra, A.; Newton, M. H.; Sattar, A. Efficient Toxicity Prediction via Simple Features using Shallow Neural Networks and Decision Trees. *ACS Omega* **2019**, *4*, 1874−1888.
(274) Smusz, S.; Kurczab, R.; Satała, G.; Bojarski, A. J. Fingerprint-based Consensus Virtual Screening towards Structurally new 5-HT6R Ligands. *Bioorg. Med. Chem. Lett.* **2015**, *25*, 1827−1830.
(275) Doshi-Velez, F.; Kim, B. Towards a Rigorous Science of Interpretable Machine Learning. *arXiv preprint arXiv:1702.08608* **2017**.
(276) Linstrom, P. J. *NIST Chemistry Webbook*. http://webbook.Nist.Gov 2005.
(277) Mata, R. A.; Suhm, M. A. Benchmarking Quantum Chemical Methods: Are We Heading in the Right Direction? *Angew. Chem., Int. Ed.* **2017**, *56*, 11011−11018.
(278) Smith, J. S.; Isayev, O.; Roitberg, A. E. ANI-1, a Data Set of 20 Million Calculated Off-Equilibrium Conformations for Organic Molecules. *Sci. Dat.* **2017**, *4*, 1−8.
(279) Atkinson, W. D.; Bond, K. E.; Tribble, G. L., III; Wilson, K. R. Computing With Feeling. *Computers & Graphics* **1977**, *2*, 97−103.
(280) Stone, J. E.; Gullingsrud, J.; Schulten, K. a System for Interactive Molecular Dynamics Simulation. *Proceedings of the 2001 Symposium on Interactive 3D Graphics*. 2001; pp 191−194.
(281) Martínez, T. J. Ab Initio Reactive Computer Aided Molecular Design. *Acc. Chem. Res.* **2017**, *50*, 652−656.
(282) Haag, M. P.; Vaucher, A. C.; Bosson, M.; Redon, S.; Reiher, M. Interactive Chemical Reactivity Exploration. *ChemPhysChem* **2014**, *15*, 3301−3319.
(283) O'Connor, M. B.; Bennie, S. J.; Deeks, H. M.; Jamieson-Binnie, A.; Jones, A. J.; Shannon, R. J.; Walters, R.; Mitchell, T. J.; Mulholland, A. J.; Glowacki, D. R. Interactive Molecular Dynamics in Virtual Reality From Quantum Chemistry to Drug Binding: An Open-Source Multi-Person Framework. *J. Chem. Phys.* **2019**, *150*, 220901.
(284) Aspuru-Guzik, A.; Lindh, R.; Reiher, M. The Matter Simulation (R) Evolution. *ACS Cent. Sci.* **2018**, *4*, 144−152.
(285) Martino, M.; Salvadori, A.; Lazzari, F.; Paoloni, L.; Nandi, S.; Mancini, G.; Barone, V.; Rampino, S. Chemical Promenades: Exploring Potential-Energy Surfaces With Immersive Virtual Reality. *J. Comput. Chem.* **2020**, *41*, 1310−1323.
(286) Juul, J. Virtual Reality: Fictional All the Way Down (And That's OK). *Disputatio-Int. J. Philos.* **2019**, *11*, 333−343.
(287) Meuwly, M.; Becker, O.; Stote, R.; Karplus, M. NO Rebinding to Myoglobin: A Reactive Molecular Dynamics Study. *Biophys. Chem.* **2002**, *98*, 183−207.
(288) Feynman, R. P.; Leighton, R. B.; Sands, M. The Feynman Lectures on Physics; Vol. I. *Am. J. Phys.* **1965**, *33*, 750−752.
(289) Van Der Kamp, M. W.; Prentice, E. J.; Kraakman, K. L.; Connolly, M.; Mulholland, A. J.; Arcus, V. L. Dynamical Origins of Heat Capacity Changes in Enzyme-Catalysed Reactions. *Nat. Commun.* **2018**, *9*, 1−7.

https://doi.org/10.1021/acs.chemrev.1c00033
Chem. Rev. XXXX, XXX, XXX−XXX

(290) Wang, L.-P.; McGibbon, R. T.; Pande, V. S.; Martinez, T. J. Automated Discovery and Refinement of Reactive Molecular Dynamics Pathways. *J. Chem. Theory Comput.* **2016**, *12*, 638−649.
(291) Zeng, J.; Cao, L.; Chin, C.-H.; Ren, H.; Zhang, J. Z.; Zhu, T. ReacNetGenerator: An Automatic Reaction Network Generator for Reactive Molecular Dynamics Simulations. *Phys. Chem. Chem. Phys.* **2020**, *22*, 683−691.
(292) Amabilino, S.; Bratholm, L. A.; Bennie, S. J.; VaucherM, A. C.; Reiher, M.; Glowacki, D. R. Training Neural Nets to Learn Reactive Potential Energy Surfaces Using Interactive Quantum Chemistry in Virtual Reality. *J. Phys. Chem. A* **2019**, *123*, 4486−4499.
(293) Judson, R.; Rabitz, H. Teaching Lasers to Control Molecules. *Phys. Rev. Lett.* **1992**, *68*, 1500−1503.
(294) Roslund, J.; Shir, O. M.; Dogariu, A.; Miles, R.; Rabitz, H. Control of Nitromethane Photoionization Efficiency With Shaped Femtosecond Pulses. *J. Chem. Phys.* **2011**, *134*, 154301.
(295) Dong, D.; Xing, X.; Ma, H.; Chen, C.; Liu, Z.; Rabitz, H. Learning-Based Quantum Robust Control: Algorithm, Applications, and Experiments. *IEEE Trans. Cybern.* **2020**, *50*, 3581−3593.

https://doi.org/10.1021/acs.chemrev.1c00033
Chem. Rev. XXXX, XXX, XXX−XXX