Intramolecular proton transfer reaction dynamics using machine learned *ab initio* potential energy surfaces

Shampa Raghunathan[1, a)]
Center for Computational Natural Sciences and Bioinformatics, International Institute of Information Technology, Hyderabad 500 032, India.

(Dated: 29 June 2021)

Hydrogen bonding interactions central to various physicochemical processes are investigated in the present study using *ab initio*-based machine learning potential energy surfaces. Abnormally strong intramolecular O–H$\cdots$O hydrogen bonds occurring in β-diketone enols of malonaldehyde, and its derivatives with substituents ranging from various electron-withdrawing to electron-donating functional groups are studied. Machine learning force fields were constructed by using a kernel-based force learning model employing *ab initio* molecular dynamics reference data. These models were used for molecular dynamics simulations at finite temperature, and dynamical properties were determined by computing their proton transfer free energy surfaces. Chemical systems studied here show progression towards forming barrierless proton transfer events at an accuracy of highly correlated electronic structure methods. Markov state models of the conformational states indicate shorter intramolecular hydrogen bonds exhibiting higher proton transfer rates. We demonstrate how functional group substitution can modulate the strength of intramolecular hydrogen bonds by studying their thermodynamic and kinetic properties.

I. INTRODUCTION:

Hydrogen bonding interactions are essential in variety of physicochemical processes, such as, enzymatic catalysis,[1–5] protein-protein interactions,[6] nucleobase interactions in RNA and DNA,[7] solid-liquid interfaces,[8,9] polymerizations,[10] molecular recognition[11] etc.. Hydrogen bonding interactions which are considered as electrostatic in nature, can also have a certain extent of covalency in the bonding characteristics.[12,13] Consequently, the energy spectrum of hydrogen bonds (HBs) lies in the broad range $\sim$1–40 kcal/mol.[14,15] Among such candidates intramolecular HBs in malonaldehyde (MA, propanedial) are extensively studied experimentally[12,16–19] as well as theoretically.[20–29] Two equivalent minima of MA in their enolic forms (enol-MA) exist in a symmetric double well potential energy surface (PES), connected via a transition state (see Figure 1). MA and its derivatives come from the 1,4-diketones family, where some of the shortest intramolecular hydrogen bond distances are realized.[30–39] The low energy barrier proton transfer (PT) in such enolized systems is known to occur due to the π-delocalization over the six-membered cyclic transition state (TS) structure.[36,37] The introduction of electron-donating and/or electron-withdrawing functional groups modulates the barrier height.

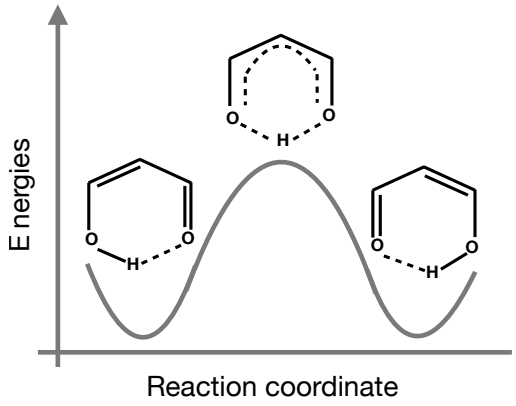


FIG. 1: Energy profile of proton transfer pathway: Two equivalent enol-MA equilibrium structures are interconverting via a transition state.

a)Electronic mail: shampa.santra@gmail.com

MA and its derivatives have been used as model systems in many studies where proton transfer was central to the investigation.[33] Specially for MA, tunneling splittings were characterized experimentally by infrared spectroscopy.[18,19] In fact, MA was studied extensively at various levels of theoretical methods starting from simple HF to high-level CCSD(T), and calculated O–H bond distance therein, lies in the range of 1.69–1.88 Å[23,40] compared to the experimentally determined value of 1.68 Å.[41] So far, the PT barrier height is estimated to be in the range $\sim$ 2.9–5.4 kcal/mol.[23,42,43] The most reliable barrier height known till now, was computed by Bowman and co-workers using the CCSD(T) level of theory at the basis set limit.[23]

However, computationally expensive nature of the high-level QM methods poses certain limitations of their usage in calculating accurate PES. Only by fully exploring the phase space using less expensive force field (FF) based approach, one may reach the $3N-6$ full-dimensionality of the PES. Unfortunately, general force fields are not suited for modelling such chemical processes where covalent interactions, such as, bond-formation and bond-breaking take place. However, there are few reactive force fields which were designed to study such PT in MA. One such FF developed by Meuwly and co-workers estimated a PT barrier height of 4.3 kcal/mol.[26]

For computationally intractable challenging problems, modern machine learning (ML) techniques have provided affordable yet accurate solutions.[44–57] Building of molecular force fields based on ML architecture that uses the properties from the *ab initio* and/or *ab initio* molecular dynamics (AIMD)[58] trajectories in order to reconstruct the PES without particularly employing any customized interatomic potentials fitted to experimental data is one of the current trends.[59–62] Both neural network-[53,63,64] and kernel-based[65,66] machine learning force fields (MLFFs) are nowadays being readily used in variety of applications for simulations of molecules and materials. Both approaches can be used to construct FFs via either energy and/or force learning. Force learning maybe advantageous over energy learning-based ML architecture;[62] after all atomic forces are key to MD simulations.[52]

In the present work we have chosen the symmetric gradient domain machine learning (sGDML) approach of Chmiela *et al.*.[65] The sGDML model was successfully employed in case of predicting the PT in MA at the CCSD(T) level of accuracy with only a few hundred reference conformations. Considering the accuracy and reasonably low computational cost, we have chosen the sGDML architecture to further construct molecular FFs for the current study.

A systematic study on various derivatives of malonaldehyde has been carried out in order to find suitable descriptors to characterize the strength of intramolecular HBs. In the present work, derivatives of MA consist of various electron-withdrawing groups, like, CN, NO_2, BH_2 and electron-donating groups, like, CH_3, NH_2, OCH_3. Density functional theory (DFT)-based AIMD energies and forces were computed for generating reference datasets. We have constructed MLFFs for all systems using the sGDML model. MLFF was then employed to explore the full-dimensional reactive PES, which was not feasible using a conventional molecular force field. Now, subtle quantum effects described by the sGDML's reconstructed PES enabled characterization of the HB strength through geometrical properties, π-delocalization indices, vibrational spectra, and rates of proton transfer processes of MA-systems. We have investigated how various functional groups modulate the PT barrier, and to what extent one can achieve a barrierless PT process.[67,68]

II. METHODOLOGY

The working pipeline includes the following steps: (1) Generation of reference datasets using AIMD simulations, (2) Training sGDML models and predicting force and energy labels, (3) Performing long timescale MD simulations using trained models. The entire workflow is depicted in Figure 2, which is discussed in detail in the following sections.

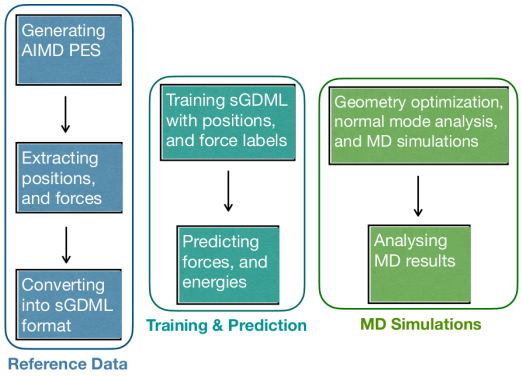


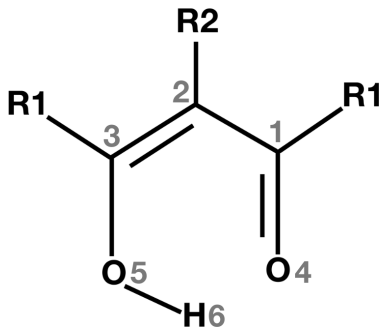
FIG. 2: Workflow from reference datasets generation → training models → performing MD simulations for computing trajectories at finite temperature.

A. Reference datasets

Terminology used throughout: Molecular systems (a total of sixteen molecules) studied currently are shown in **Scheme 1**. Here, parent malonaldehyde structure is shown with functional group substitutions at symmetrical carbons C1 and C3 by R1, and at central carbon C2 by R2. Structure-types **I**–**IV**, each consisting of a particular R1-substituent defined in **Scheme 2** are as follows: (1) Structure-type **I** has H atom, (2) Structure-type **II** has CH_3, (3) Structure-type **III** has NH_2, and (4) Structure-type **IV** has OCH_3. Therefore, "R1" term will be omitted hereafter in the usage of structure-types and/or structures in order to have compact notations. Each structure-type has four molecules with different R2-substituents. Subsequently, functional groups at R2 position will be mentioned explicitly without the term "R2". For example, NO_2-structure **I** refers a molecule where, R2 = NO_2, and R1 = H; likewise, BH_2-structure **III** refers a molecule where, R2 = BH_2, and R1 = NH_2.

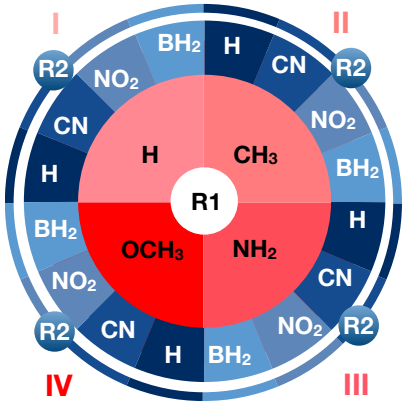
In this study, the smallest system has 9 atoms, and the largest has 19 atoms. Born-Oppenheimer molecular dynamics (BOMD) simulations were performed for generating reference data using the Car-Parrinello MD (CPMD)[69] package which integrates DFT and MD methods. Starting structures were optimized and then AIMD trajectories were generated. GTH pseudopotentials[70–72] have been used for all atoms and the valence electronic orbitals were expanded in plane waves with the maximum kinetic energy cutoff of 80 Rydberg. The PBE[73] gradient-corrected exchange-correlation functional was used. The total energy was converged to 1E-7 Hartree for a system in a cubic box of dimension 12 Å. A timestep of 21 a.u. ($\sim$ 0.5 fs) was used. The total simulation length was 5 ps for each system. This resulted in 9865 reference data points for each molecular system. The equilibrium temperature was kept at 300 K controlled with the Nóse-Hoover chain thermostats.

Scheme 1



R1 = H, CH_3, NH_2, OCH_3
R2 = H, CN, NO_2, BH_2

Scheme 2: Structure-types



B. The sGDML model

The implementation details of the sGDML model can be found in the original work by Chmiela and co-workers.[61,65] The sGDML model uses the kernel ridge regression (KKR) technique,

$$(\mathbf{K}_{\mathrm{Hess}(\kappa)} + \lambda \mathbb{I})\vec{\alpha} = \nabla V_{\mathrm{BO}} = -\mathbf{F} \tag{1}$$

trained on forces $\mathbf{F}$. Here, $\mathbf{K}_{\mathrm{Hess}(\kappa)}$ is the kernel matrix. The regularization is done by hyperparameter λ. $\mathbb{I}$ and $\vec{\alpha}$ are the identity matrix and the parameter-vectors, respectively. Hessian matrix of the kernel, $\mathrm{Hess}(\kappa)$ often termed as force field kernel is obtained as the covariance between examples $\vec{x}$ and $\vec{x'}$

$$\kappa(\vec{x},\vec{x'}) = \langle \phi(\vec{x}), \phi(\vec{x'}) \rangle_{\mathscr{H}} \tag{2}$$

in the reproducing kernel Hilbert space $\mathscr{H}$. Here, $\phi : \chi \mapsto \mathscr{H}$ is the mapping from input space $\mathbf{x} \in \chi$ to feature space which is defined implicitly through scalar-valued kernel function via so-called kernel trick. Parametric Matérn family kernel functions were used,

$$\kappa(d) = (1 + \frac{\sqrt{5}d}{\sigma} + \frac{5d^2}{3\sigma^2})\exp(-\frac{\sqrt{5}d}{\sigma}) \tag{3}$$

where, $d = ||\vec{x} - \vec{x'}||$ is the Euclidean distance between two inputs, and the length scale σ is a hyperparameter. With $\vec{x} = D(\vec{r})$, the kernel function is associated with the descriptor D; where, $\vec{r}$ is the Cartesian molecular geometry. Each element of the descriptor matrix is defined as the reciprocal of the Euclidean distance for a pair of atoms.

$$D_{ij} = \begin{cases} ||r_i - r_j||^{-1} & \text{for } i > j \\ 0 & \text{for } i \leq j \end{cases}$$

Prediction of forces are done using the force estimator

$$\hat{\mathbf{f}}_{\mathrm{F}}(\vec{x}) = \sum_{i}^{M}\sum_{l}^{3N}\sum_{q}^{S}(\mathbf{P}_q\vec{\alpha}_i)_l \frac{\partial}{\partial x_l}\nabla\kappa(\vec{x},\mathbf{P}_q\vec{x}_i) \tag{4}$$

which collects all contributions ($3N$ coordinates) from all M training samples, and S symmetry transformations realized by permutation matrix $\mathbf{P}$. Further, by integrating $\hat{\mathbf{f}}_F$ w.r.t. the Cartesian geometry, the corresponding energy predictor is obtained.

$$\hat{f}_{\mathrm{E}}(\vec{x}) = \int \hat{\mathbf{f}}_{\mathrm{F}} . d\mathbf{x} = \sum_{i}^{M}\sum_{l}^{3N}\sum_{q}^{S}(\mathbf{P}_q\vec{\alpha}_i)_l \frac{\partial}{\partial x_l}\kappa(\vec{x},\mathbf{P}_q\vec{x}_i) \tag{5}$$

C. MD simulations

In order to generate MD trajectories using the MLFF, we have used Atomistic Simulation Environment[74] with a 0.2 fs timestep. A temperature of 300 K was maintained via a Langevin thermostat. A thermostat friction value of 0.002 a.u. was used for all systems unless stated otherwise. Each trajectory of 500 ps (2500000 steps) long was used to evaluate the performance of the sGDML model. MD simulations were performed for all systems.

D. Kinetic analysis: PT rates

We have used the transition path theory (TPT)[75–77] to calculate the rates of the PT processes as implemented in the PyEMMA[78] package. To model dynamical events involved in transitions among various metastable states, a Markov state model (MSM) was constructed using the MD simulations data at a finite time interval (1 fs). MD conformational space was discretized into three-state model using the O$\cdots$H distance feature. State **2** was defined as the transition state with a O$\cdots$H separation of 1.2 Å, and the rest of the conformations are of state **1** and **3** types–corresponding to a pair of minima. To estimate the MSM model, a suitable lag time of 1 fs was chosen, and also Chapman-Kolmogorov (CK) test was performed to validate (Markovianity) the MSM models at higher lag times.

III. RESULTS AND DISCUSSION

Each of the sGDML models is trained on a set of reference datasets of consistent size with about 9.8k samples with a resolution of 0.5 fs. Energy values (relative to the structure at the first timestep)

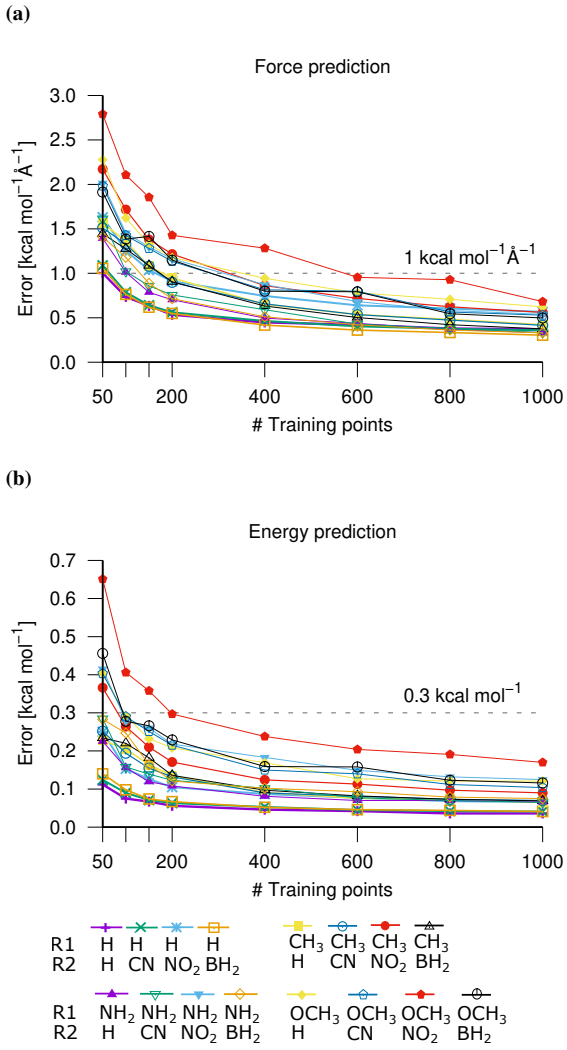


FIG. 3: Energy and force prediction accuracies measured by evaluation metric, mean absolute error (MAE) as a function of training set size. All sixteen models were trained on AIMD forces.

across the dataset span over a range of about 0–25 kcal/mol, and corresponding force magnitudes lie in the range ∼50–250 kcal mol^{-1}Å^{-1}.

A. Accuracy of the model

Prediction accuracies of sGDML models in terms of forces and energies are given in Figures 3a and 3b, respectively, as mean absolute error (MAE) basis of performance metrics. From Figure 3a, it can be seen that all sGDML models achieve a force accuracy of 1 kcal mol^{-1}Å^{-1}, using just about 200 training examples, except for the structure-type **IV** which needed about 600 training examples to reach the same accuracy. However, energy-based ML models typically need 2–3 orders of magnitude more data to gain comparable accuracy.[60] Similarly, the energy prediction accuracy is achieved below 0.3 kcal mol^{-1} for all models using just 200 training examples (see Figure 3b). Figure S1 in Supplementary Material shows the convergence of MAE with increasing training set size for six independent runs per training set size for the molecule BH_2-structure **III**. This displays how errors are distributed: As the training set size increases MAE is reduced, indicating a low variance of the errors. Additionally, energy and force predictions for all structures for a consecutive 500 steps are given in Figures S2–S5 in Supplementary Material.

B. Ground state properties of ab initio accuracy

Firstly, MLFFs are used to optimize the minimum energy structure of the enol-forms of all systems, and vibrational frequencies were computed numerically (see Table S1 in Supplementary Material for the OH stretch which is responsible for the hydrogen transfer). O···O distances of all optimized geometries are given in Table I where each row is augmented with a 2^{nd} row containing the values from the work of Schaefer and co-workers.[33] It is interesting to see that among malonaldehyde derivatives along the R1-series, *i.e.*, substituents at the symmetrical carbon, the shortest O···O distance was found for the NH_2 substituent, irrespective of the substituent at the R2-carbon (central carbon). This trend is similar to what was reported earlier by Schaefer and co-workers.[33] Considering a fixed R2-substituent, along the R1-series, O···O distance decreases as we move towards the electron-rich substituents, however it increases in case of the OCH_3.

TABLE I: O···O distances (Å) between two oxygen atoms (O4 and O5, for numbering, see **Scheme 1**) in MA and its derivatives. π-delocalization indices $|Q|$ are given in parentheses. O···O distances in lower rows were calculated at DFT-B3LYP/DZP++ level of theory, taken from Ref.[33]

R2 \ R1	**H**	**CH_3**	**NH_2**	**OCH_3**
H	2.502 (0.108) 2.546	2.471 (0.099) 2.511	2.466 (0.083) 2.474	2.479 (0.093) 2.498
CN	2.483 (0.104) 2.526	2.469 (0.105) 2.471	2.457 (0.070) 2.448	2.455 (0.078) 2.464
NO_2	2.485 (0.097) 2.521	2.436 (0.080) 2.423	2.415 (0.048) 2.380	2.449 (0.078) 2.442
BH_2	2.468 (0.086) 2.499	2.433 (0.072) 2.419	2.424 (0.050) 2.398	2.437 (0.068) 2.421

The structure-type **IV** is showing almost always a similar O···O distance as in the case of structure-type **II**. Now, for a fixed R1-substituent, along the R2-series there is a decrease in the O···O distance as one goes from the H atom to BH_2 functional group, except for the NO_2-structure **III** that has the shortest O···O distance among all systems studied here. All these findings are very similar to the previously found results.[33] This is quite convincing that sGDML models are able to predict the ground state geometries quite well.

Gilli and co-workers characterized the HB strength by the π-delocalization index, $|Q| = d(C3-O5) - d(C1-O4) + d(C1-C2) - d(C2-C3)$ of the short conjugated chain connecting the hydrogen bond donor and acceptor atoms.[37] Following the same method, we have calculated $|Q|$ values, and tabulated them in Table I. The lower the $|Q|$ value is, the stronger the π-delocalization a system shows. From Table I, we can see that as the electron withdrawing inductive effect is increasing along the R2-series, the $|Q|$ value is decreasing. Hence, stronger the delocalization of the π-conjugated system is, shorter the O···O distance becomes. To this end, π-delocalization index can be a reasonable descriptor for predicting the strength of a symmetrical intramolecular hydrogen bond.

C. MD results of ab initio quality

As, the performance of the models reached chemical accuracy, we further used them to perform MD simulations for longer timescales in order to have sufficient sampling of the conformation space. So that, we can have insightful analyses of thermodynamic and kinetic properties of molecular systems. In fact, well known sampling problem often limits our ability to compute rare events using MD simulations data. Followed by the seminal work on the sGDML model of MA, it was shown that proton transfer barrier in the symmetric double well potential of the MA molecule was about 4.0 kcal mol^{-1}, and the O···O bond distance was about 2.38 Å.[62] This is overwhelming as conventional force fields unable to do it. However, Cooper et al. have used a different approach where local part of the PES (instead of global) associated to a transition state was used for training a neural network model by taking into account energies, and their first and second derivatives for calculating accurate rate constant using instanton theory.[79]

We have analyzed the free energy surface from the MD trajectories obtained for each of the sGDML models. Figure 4 shows the two-dimensional free-energy surfaces as a function of a pair of O···H distances. It is evident that all sGDML models are able to qualitatively describe the symmetric double well potential (two minima)

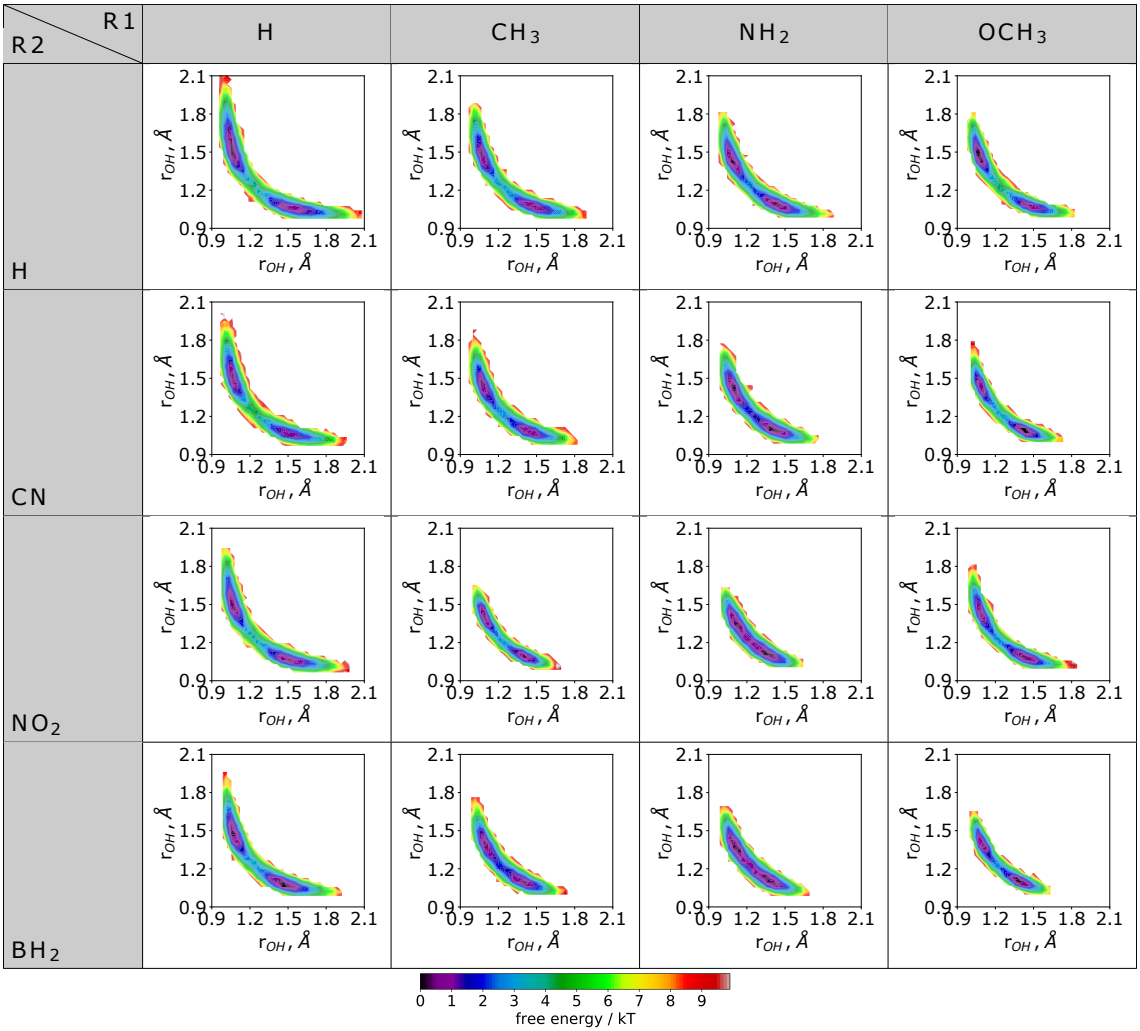


FIG. 4: Two-dimensional free energy surface along the two O···H distances of the O···H···O PT reaction coordinate for a malonaldehyde derivative consisting of a pair of R1 and R2 substituents. A total of sixteen plots are shown.

connected via a transition state. We found that as we go along either the R1-series or the R2-series the energy barrier is decreasing. In case of the structure-type **I**, PT barrier is lowered from ~ 4 to ~ 3 kcal/mol. In case of the structure-type **II**, PT barrier is lowered even more, from ~ 4 to ~ 2 kcal/mol. In a very recent study, Qu et al. have calculated a barrier of 3.5 kcal/mol for the H-structure **II** using a Δ-machine learning approach.[80] Clearly, major substituent effects are seen in case of the structure-type **III**. Here, PT barrier is seen to go down from ~ 2 to <1 kcal/mol. We are realizing almost barrierless transitions. Especially, NO_2-structure **III** is exhibiting the lowest PT barrier; earlier this value was calculated at the MP2 level as 0.6 kcal/mol.[31] Interestingly, BH_2-structure **III** has an equivalent PT energy barrier as the former. Besides, the area of the PES is decreasing, i.e., O···O distance is

becoming shorter. Structure-type **IV** seems to behave more like Structure-type **II** even though OCH_3 has stronger electron donating capability than CH_3 group.

Presumably, larger OCH_3 group in Structure-type **IV** reduces the efficacy of the R2-substituents in spite of their increasing electron withdrawing power along the R2-series. Earlier too, it was found that stronger electron withdrawing group at R2 and electron donating group at R1 tend to produce stronger intramolecular hydrogen bonds. However, bulky groups at R1 may have altered influence on the hydrogen transfer barrier.[33] They have calculated high-level coupled cluster PT energy barriers at the CBS limit. Presently, we are able to achieve energetics of chemical processes using sGDML models efficiently at the cost of MD with coupled cluster accuracy. Substituent effect was often used as basis for qualitative interpretation of HB strength.[81] Here too, we show the same from the free energy landscape computed using the MLFFs.

D. Kinetic analyses

Feature selection for constructing kinetic models: We have computed the PT rates using the MD trajectories which were simulated using the sGDML FFs for all studied MA-systems. One of the important aspects of analyzing kinetics of a chemical process is the selection of the feature which is able to describe its nature well. Despite the large number of dedicated experimental and theoretical studies, our understanding on determinants of HB strength is surprisingly fragmentary.[82] Nevertheless, O$\cdots$H distances are often used.[81,83] The population of the two O$\cdots$H distances is shown in Figure 5. We have chosen four structures: H-structure **I**, H-structure **III**, NO_2-structure **I**, and NO_2-structure **III** to show the extent of substituent effect on the O$\cdots$H population. We can clearly see that NO_2 substituent has significant effect on O$\cdots$H population in structure-types **I** and **III** (shown in Figures 5c and d, respectively) than H atom (Figures 5a and b); which is expected as the former is known to have strong electron withdrawing power. Besides, NO_2-structure **III** has an electron-rich NH_2 group at the R2 center. As a result, O$\cdots$H populations are affected here the most. From this, we can easily infer that two O$\cdots$H distances are good choices as features for kinetic modelling.

Discrete state kinetic models, like, Markov state models (MSMs) are shown to be useful for understanding conformational states involved in bio(molecular) transitions.[84,85] Pipeline of an MSM model is given in Figure S6 in Supplementary Material. MSM analyses results are shown in Figure 6 which shows the rate matrix obtained from the three-state MSM, represented by a PyEMMA network plot.[78] After analyzing the fluxes between the two metastable sets, it can easily be found that major transfer of fluxes happens between states **1** and **3** via state **2**. The individual transition rate is given in the unit of fs^{-1}. We can see that the two minima (states **1** and **3**) have larger populations than the state **2** (TS) in all systems except NO_2-structure **III** and BH_2-structure **III** molecules; here state **2** has populations as high as states **1** and **3**. Comparing the rates along the R1-series, we can see that PT rate is becoming higher from structure-type **I** to **II** to **III** (row-wise). Also, along the R2-series, PT rate is elevated, for example, from H-structure **I** to CN-structure **I** to NO_2-structure **I** to BH_2-structure **I** (column-wise). Similar trend was observed in case of free energy barrier also. Using transition state theory, the rate of PT was calculated to be 0.0076 ps^{-1} (applying a frequency factor of $10^{12} s^{-1}$) for the parent MA[86] with the activation barrier of 2.91 kcal mol^{-1}. In case of the structure-type **I** we can see that rate is increased from 0.002–0.004 fs^{-1} with R2-substituents varying from H atom to functional group BH_2. Whereas, structure-type **II** changes its rate from 0.004–0.01 fs^{-1} as the electron withdrawing effect of the substituents increases. Structure-type **III** shows rate from 0.006–0.014 fs^{-1} among various electron withdrawing substituents. Structure-type **IV** showed similar PT rates as structure-type **II**, likewise the trend, we saw in case of free energy barriers. Rates of proton transfer processes are found to be very consistent with free energy barriers, as expected. In practice, we know that as the barrier height decreases rate of a chemical process should become faster. The proton transfer rate in Structure-type **IV** is only moderately greater than the parent MA. This can be attributed to the larger size of the OCH_3 group. Mod-

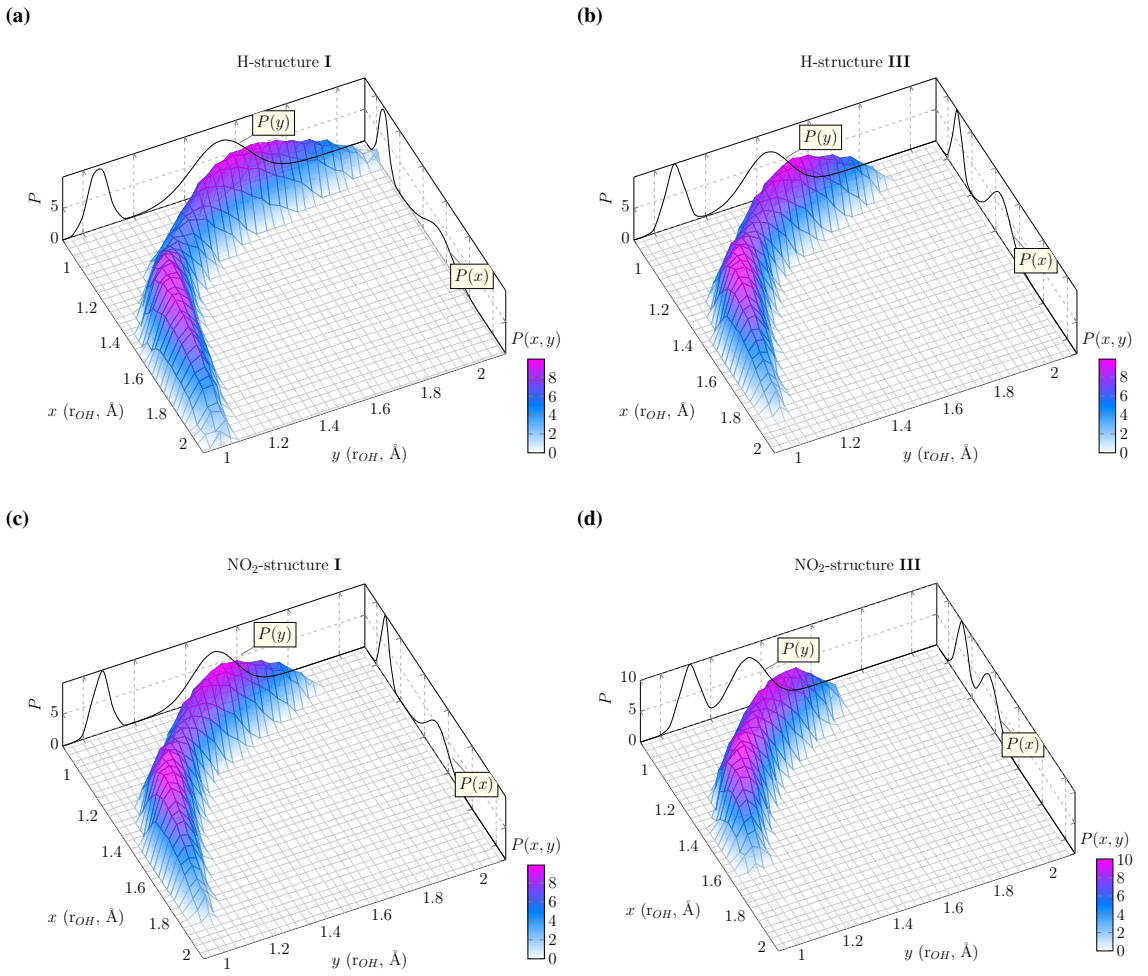


FIG. 5: Joint probabilities, $P(x,y)$ of the two O$\cdots$H distances of the O$\cdots$H$\cdots$O PT reaction coordinate along with their individual probabilities. (a) H-structure **I**, (b) H-structure **III**, (c) NO_2-structure **I**, and (d) NO_2-structure **III**. Probability values are scaled by a factor of 10 for clarity.

erate increase in reaction rates in case of the bulky functional groups was suggested earlier as well.[87] Our ML models are able to predict PT rates where fundamental physical phenomenon of molecular kinetics are reflected. Therefore, PT rate could be a reliable descriptor of intramolecular HB strength.

Finally, determined rates can be used in the Hammett equation, $\log k = \log k^0 + \rho\sigma$. Where, k, and k^0 are rates of the substituted and unsubstituted compounds, respectively. The substituent constant σ measures inductive effect relative to hydrogen of a substituent in the *meta*- or *para*-center. Here, we use the mesomeric constant (σ_R^0)[88] which is more relevant considering the resonance-assisted nature of the HBs presented in the current study. ρ can be thought as the susceptibility of the reaction to the inductive effects. Figure 7a and 7b shows $\log(k/k^0)$ for a given substituent at R2 (while R1 varies), and

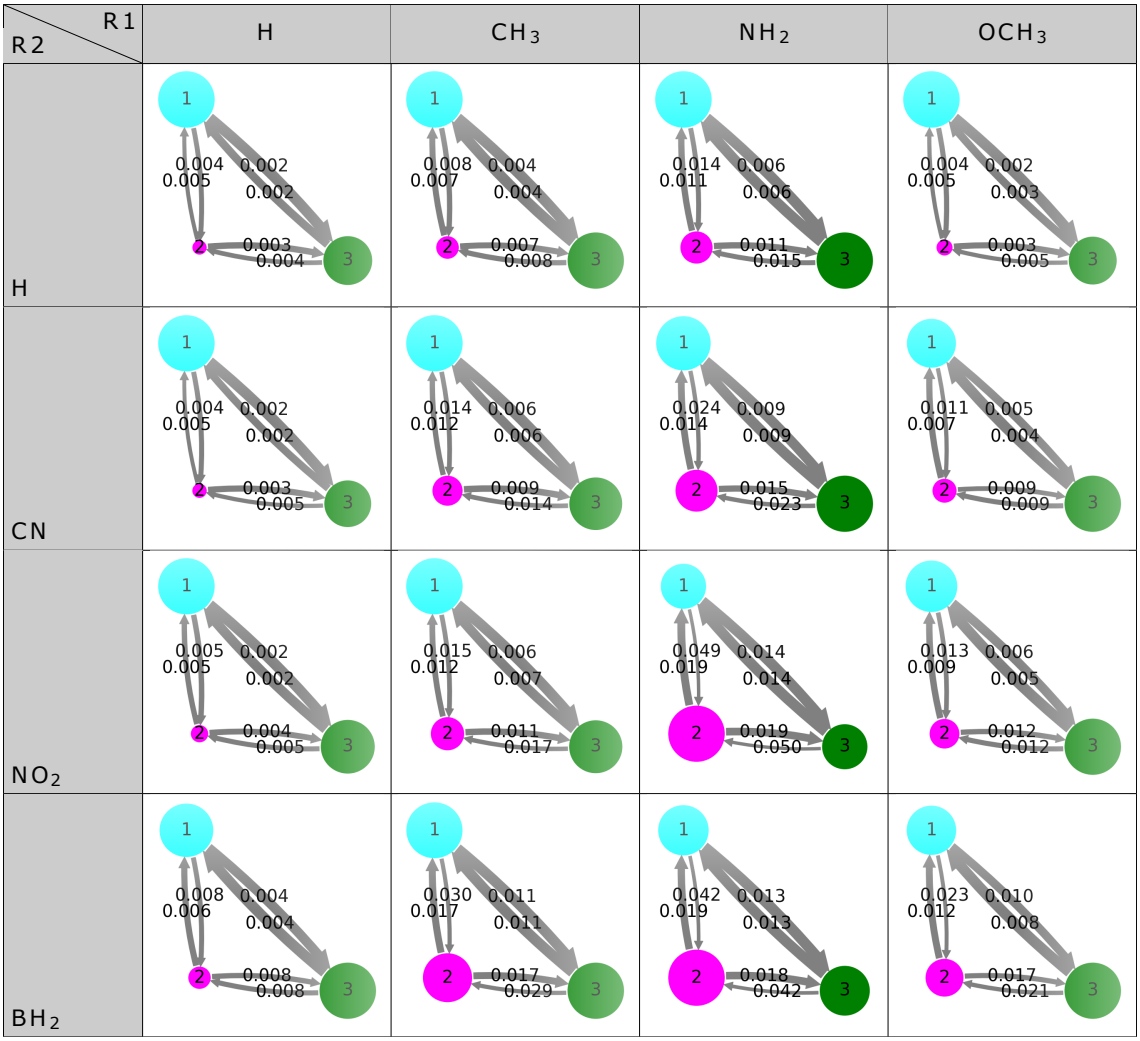


FIG. 6: Kinetic rate network between metastable states (a pair of enol-MAs **1** and **3**, and a TS **2**) with lifetimes of the order of femtoseconds. The arrows represent transitions between states, with their thickness proportional to the transition probability. In every intermediate state, the flux that enters the state is equal to the flux leaving the state. Similarly, disc area is proportional to the state population. Values above the arrows quantify the transition rates between pairs of metastable sets. Each of the sixteen plots refers a MA derivative consisting of a pair of R1 and R2 substituents.

at R1 (while R2 varies), respectively. Figure 7a shows that PT rate increases as we move to the electron-rich substituents, however drops in case of the OCH_3. Figure 7b shows that PT rate increases till the electron-poor substituent NO_2. Further, rates are elevated for BH_2 even though it functions as a weak electron-withdrawing group. All these findings suggest that both electronic and steric effects are important factors for seemingly barrier less intramolecular proton transfer in substituted malon-

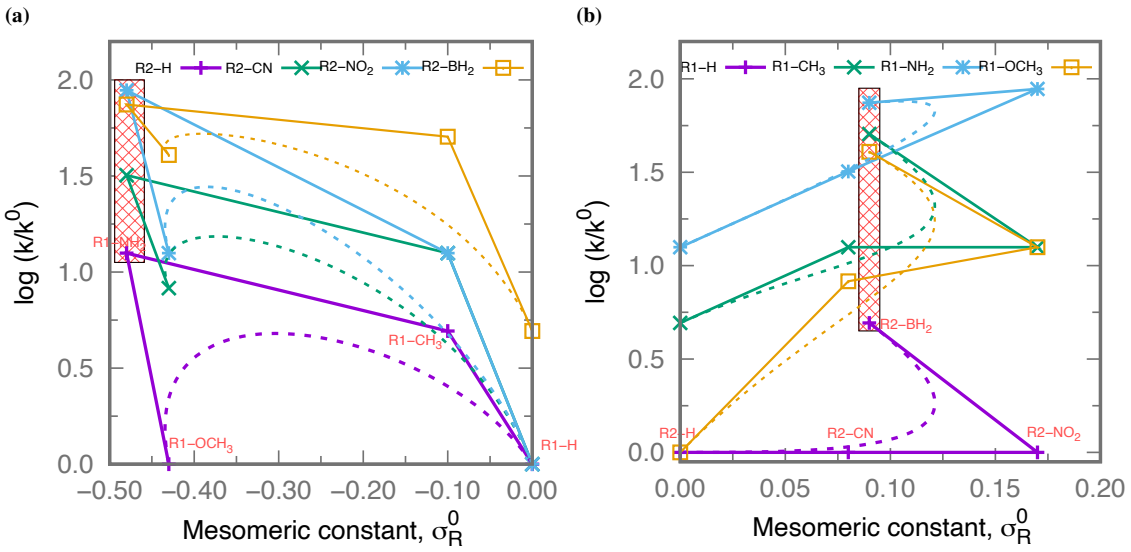


FIG. 7: $\log(k/k^0)$ as a function of the mesomeric constant σ_R^0 of (a) R2 substituents; shaded rectangle showing the highest rate when R1 is NH_2, and (b) R1 substituents; shaded rectangle showing the highest rate when R2 is BH_2. Dashed curves are the smooth Bézier interpolation of the solid lines, indicating how $\log(k/k^0)$ drops for OCH_3 in (a) and for NO_2 in (b) in spite of being the best electron-donating and electron-withdrawing groups, respectively.

aldehydes.

IV. CONCLUSIONS

Occurrence of intramolecular hydrogen bonding interactions are widespread in chemical and biological systems. Characterization of such HBs help to determine their functional roles on enzymatic reactions, mechanism of action of drugs, base pairing interactions in DNA, RNA etc.. Malonaldehyde exhibiting intramolecular HBs has long been serving as prototype system for testing and validating computational approaches, like, developing reactive models for studying chemical reaction dynamics. Here, we have extended our model systems including various substituted MAs. Presently, with the availability of the ML approaches, full-dimensional long timescale studies are feasible too. In this study, we showed that MLFFs are able to describe intramolecular low energy barrier HBs at the high-level of QM-accuracy. Here, free energy barrier of the proton transfer process in symmetric OHO β-diketone tautomers (O-H···O=C) was shown to modulate with the nature of the substituents including their electronic and steric factors. MLFFs are indeed able to capture the characteristic dynamical behavior of the intramolecular HBs responsible for exhibiting correct trend of the ground state structures as well as the proton transfer kinetics studied by Markov state modelling. We have shown that functional group inductive effect can serve as a factual basis for a smooth and consistent interpretation of HB strength. It will be interesting to see how asymmetric NHO resonance assisted hydrogen bonding systems, like, arylazophenol (N=N···H-O) and its arylhydrazo-quinone tautomer (N-NH···O=C) will be influenced by the nature of the functional groups apart from those considered in the present study, like, various halogen functionalities.

SUPPLEMENTARY MATERIAL

See supplementary material for additional information on prediction accuracies, OH stretching frequencies, and pipeline for a Markov state model.

ACKNOWLEDGMENTS

We thank Dr. S. Chmiela for making the sGDML code available. The DST/WOS-A (grant, no. SR/WOS-A/CS-19/2018 (G)) agency is acknowledged for the financial support.

DATA AVAILABILITY

The reference and training data used in this study are available from the authors upon reasonable request.

[1] W. Cleland and M. M. Kreevoy, Science **264**, 1887 (1994).
[2] A. Warshel, A. Papazyan, P. A. Kollman, W. Cleland, M. M. Kreevoy, and P. A. Frey, Science **269**, 102 (1995).
[3] J. A. Gerlt, M. M. Kreevoy, W. Cleland, and P. A. Frey, Chemistry & Biology **4**, 259 (1997).
[4] C. L. Perrin and J. B. Nielson, Annu. Rev. Phys. Chem. **48**, 511 (1997).
[5] V. Vennelakanti, H. W. Qi, R. Mehmood, and H. J. Kulik, Chem. Sci. **12**, 1147 (2021).
[6] S. Raghunathan, K. El Hage, J. L. Desmond, L. Zhang, and M. Meuwly, J. Phys. Chem. B **122**, 7038 (2018).
[7] S. Chandorkar, S. Raghunathan, T. Jaganade, and U. D. Priyakumar, Int. J. Mol. Sci. **22** (2021).
[8] O. Björneholm, M. H. Hansen, A. Hodgson, L.-M. Liu, D. T. Limmer, A. Michaelides, P. Pedevilla, J. Rossmeisl, H. Shen, G. Tocci, E. Tyrode, M.-M. Walz, J. Werner, and H. Bluhm, Chem. Rev. **116**, 7698 (2016).
[9] M. Hellström, V. Quaranta, and J. Behler, Chem. Sci. **10**, 1232 (2019).
[10] H.-J. Schneider, Angew. Chem. Int. Ed. **48**, 3924 (2009).
[11] D. Kathuria, A. A. Bankar, and P. V. Bharatam, J. Mol. Struc. **1152**, 61 (2018).
[12] P. Gilli, V. Bertolasi, V. Ferretti, and G. Gilli, J. Am. Chem. Soc. **116**, 909 (1994).
[13] S. J. Grabowski, Chem. Rev. **111**, 2597 (2011).
[14] J. Ireta, J. Neugebauer, and M. Scheffler, J. Phys. Chem. A **108**, 5692 (2004).
[15] G. R. Desiraju, Acc. Chem. Res. **29**, 441 (1996).
[16] A. Bothner-By and R. Harris, J. Org. Chem. **30**, 254 (1965).
[17] T. Baba, T. Tanaka, I. Morino, K. M. Yamada, and K. Tanaka, J. Chem. Phys. **110**, 4131 (1999).
[18] N. O. Lüttschwager, T. N. Wassermann, S. Coussan, and M. A. Suhm, Mol. Phys. **111**, 2211 (2013).
[19] N. O. Lüttschwager, T. N. Wassermann, S. Coussan, and M. A. Suhm, Phys. Chem. Chem. Phys. **12**, 8201 (2010).
[20] T. Carrington Jr and W. H. Miller, J. Chem. Phys. **84**, 4364 (1986).
[21] D. P. Tew, N. C. Handy, and S. Carter, J. Chem. Phys. **125**, 084313 (2006).
[22] M. Ben-Nun and T. J. Martínez, J. Phys. Chem. A **103**, 6055 (1999).
[23] Y. Wang, B. J. Braams, J. M. Bowman, S. Carter, and D. P. Tew, J. Chem. Phys. **128**, 224314 (2008).
[24] J. Huang, M. Buchowiecki, T. Nagy, J. Vaníček, and M. Meuwly, Phys. Chem. Chem. Phys. **16**, 204 (2014).
[25] T. Hammer and U. Manthe, J. Chem. Phys. **134**, 224305 (2011).
[26] Y. Yang and M. Meuwly, J. Chem. Phys. **133**, 064503 (2010).
[27] M. Schröder, F. Gatti, and H.-D. Meyer, J. Chem. Phys. **134**, 234307 (2011).
[28] A. Hazra, J. H. Skone, and S. Hammes-Schiffer, J. Chem. Phys. **130**, 054108 (2009).
[29] A. Viel, M. D. Coutinho-Neto, and U. Manthe, J. Chem. Phys. **126**, 024308 (2007).
[30] G. Buemi and C. Gandolfo, J. Chem. Soc., Faraday Trans. 2 **85**, 215 (1989).
[31] G. K. Madsen, C. Wilson, T. M. Nymand, G. J. McIntyre, and F. K. Larsen, J. Phys. Chem. A **103**, 8684 (1999).
[32] A. Basheer, H. Yamataka, S. C. Ammal, and Z. Rappoport, J. Org. Chem. **72**, 5297 (2007).
[33] J. C. Hargis, F. A. Evangelista, J. B. Ingels, and H. F. Schaefer III, J. Am. Chem. Soc. **130**, 17471 (2008).
[34] P. Durlak and Z. Latajka, J. Chem. Theory Comput. **9**, 65 (2013).
[35] P. Durlak, K. Mierzwicki, and Z. Latajka, J. Phys. Chem. B **117**, 5430 (2013).
[36] P. Gilli, V. Bertolasi, L. Pretto, A. Lyčka, and G. Gilli, J. Am. Chem. Soc. **124**, 13554 (2002).
[37] P. Gilli, V. Bertolasi, L. Pretto, L. Antonov, and G. Gilli, J. Am. Chem. Soc. **127**, 4943 (2005).
[38] P. E. Hansen and J. Spanget-Larsen, Molecules **22**, 552 (2017).
[39] S. Käser, O. Unke, and M. Meuwly, New J. Phys. (2020).
[40] M. J. Frisch, A. C. Scheiner, H. F. Schaefer III, and J. S. Binkley, J. Chem. Phys. **82**, 4194 (1985).
[41] S. L. Baughcum, R. W. Duerst, W. F. Rowe, Z. Smith, and E. B. Wilson, J. Am. Chem. Soc. **103**, 6296 (1981).
[42] G. V. Mil'nikov, K. Yagi, T. Taketsugu, H. Nakamura, and K. Hirao, J. Chem. Phys. **119**, 10 (2003).
[43] G. V. Mil'nikov, K. Yagi, T. Taketsugu, H. Nakamura, and K. Hirao, J. Chem. Phys. **120**, 5036 (2004).
[44] K. Hansen, F. Biegler, R. Ramakrishnan, W. Pronobis, O. A. von Lilienfeld, K.-R. Müller, and A. Tkatchenko, J. Phys. Chem. Lett. **6**, 2326 (2015).
[45] Z. Li, J. R. Kermode, and A. De Vita, Phys. Rev. Lett. **114**, 096405 (2015).
[46] R. Ramakrishnan and O. A. von Lilienfeld, Rev. Comput. Chem. **30**, 225 (2017).

[47] O. T. Unke and M. Meuwly, J. Chem. Theory Comput. **15**, 3678 (2019).

[48] S. Laghuvarapu, Y. Pathak, and U. D. Priyakumar, J. Comput. Chem. (2019).

[49] Y. Pathak, S. Laghuvarapu, S. Mehta, and U. D. Priyakumar, in *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34 (2020) pp. 873–880.

[50] P. Pattnaik, S. Raghunathan, T. Kalluri, P. Bhimalapuram, C. V. Jawahar, and U. D. Priyakumar, J. Phys. Chem. A **124**, 6954 (2020).

[51] Y. Pathak, K. S. Juneja, G. Varma, M. Ehara, and U. D. Priyakumar, Phys. Chem. Chem. Phys. **22**, 26935 (2020).

[52] O. A. von Lilienfeld and K. Burke, Nat. Commun. **11**, 1 (2020).

[53] J. S. Smith, O. Isayev, and A. E. Roitberg, Chem. Sci. **8**, 3192 (2017).

[54] J. Behler, J. Chem. Phys. **145**, 170901 (2016).

[55] K. T. Schütt, F. Arbabzadah, S. Chmiela, K. R. Müller, and A. Tkatchenko, Nat. Commun. **8**, 1 (2017).

[56] A. P. Bartók, S. De, C. Poelking, N. Bernstein, J. R. Kermode, G. Csányi, and M. Ceriotti, Sci. Adv. **3**, e1701816 (2017).

[57] T. D. Huan, R. Batra, J. Chapman, S. Krishnan, L. Chen, and R. Ramprasad, NPJ Comput. Mater. **3**, 1 (2017).

[58] R. Car and M. Parrinello, Phys. Rev. Lett. **55**, 2471 (1985).

[59] F. Brockherde, L. Vogt, L. Li, M. E. Tuckerman, K. Burke, and K.-R. Müller, Nat. Commun. **8**, 1 (2017).

[60] S. Chmiela, A. Tkatchenko, H. E. Sauceda, I. Poltavsky, K. T. Schütt, and K.-R. Müller, Sci. Adv. **3**, e1603015 (2017).

[61] S. Chmiela, H. E. Sauceda, K.-R. Müller, and A. Tkatchenko, Nat. Commun. **9**, 1 (2018).

[62] H. E. Sauceda, S. Chmiela, I. Poltavsky, K.-R. Müller, and A. Tkatchenko, J. Chem. Phys. **150**, 114102 (2019).

[63] S. Manzhos and T. Carrington, Chem. Rev. **0**, null (0), pMID: 33021368, https://doi.org/10.1021/acs.chemrev.0c00665.

[64] J. Behler and M. Parrinello, Phys. Rev. Lett. **98**, 146401 (2007).

[65] S. Chmiela, H. E. Sauceda, I. Poltavsky, K.-R. Müller, and A. Tkatchenko, Comput. Phys. Commun. **240**, 38 (2019), GitHub repository: https://github.com/stefanch/sGDML.

[66] T. Bereau, D. Andrienko, and O. A. Von Lilienfeld, J. Chem. Theory Comput. **11**, 3225 (2015).

[67] T. Hollebeek, T.-S. Ho, and H. Rabitz, Annu. Rev. Phys. Chem. **50**, 537 (1999).

[68] P. Gilli, L. Pretto, V. Bertolasi, and G. Gilli, Acc. Chem. Res. **42**, 33 (2009).

[69] J. Hutter, Copyright IBM Corp 1990-2019, Copyright MPI für Festkörperforschung Stuttgart 1997-2001 .

[70] S. Goedecker, M. Teter, and J. Hutter, Phys. Rev. B **54**, 1703 (1996).

[71] C. Hartwigsen, S. Gœdecker, and J. Hutter, Phys. Rev. B **58**, 3641 (1998).

[72] M. Krack, Theor. Chem. Acc. **114**, 145 (2005).

[73] J. P. Perdew, K. Burke, and M. Ernzerhof, Phys. Rev. Lett. **77**, 3865 (1996).

[74] S. R. Bahn and K. W. Jacobsen, Computing in Science & Engineering **4**, 56 (2002).

[75] E. Weinan and E. Vanden-Eijnden, Journal of Statistical Physics **123**, 503 (2006).

[76] P. Metzner, C. Schütte, and E. Vanden-Eijnden, Multiscale Modeling & Simulation **7**, 1192 (2009).

[77] F. Noé, C. Schütte, E. Vanden-Eijnden, L. Reich, and T. R. Weikl, Proc. Natl. Acad. Sci. USA **106**, 19011 (2009).

[78] M. K. Scherer, B. Trendelkamp-Schroer, F. Paul, G. Pérez-Hernández, M. Hoffmann, N. Plattner, C. Wehmeyer, J.-H. Prinz, and F. Noé, J. Chem. Theory Comput. **11**, 5525 (2015).

[79] A. M. Cooper, P. P. Hallmen, and J. Kästner, J. Chem. Phys. **148**, 094106 (2018).

[80] C. Qu, P. L. Houston, R. Conte, A. Nandi, and J. M. Bowman, J. Phys. Chem. Lett. **12**, 4902 (2021).

[81] H. Lampert, W. Mikenda, and A. Karpfen, J. Phys. Chem. **100**, 7418 (1996).

[82] G. Gilli and P. Gilli, *The nature of the hydrogen bond: Outline of a comprehensive hydrogen bond theory*, Vol. 23 (Oxford University Press, 2009).

[83] R. N. Karingithi, C. L. Shaw, E. W. Roberts, and P. A. Molina, J. Mol. Struct. THEOCHEM **851**, 92 (2008).

[84] B. E. Husic and V. S. Pande, J. Am. Chem. Soc. **140**, 2386 (2018).

[85] J. Xiao and F. R. Salsbury, Phys. Chem. Chem. Phys. **21**, 4320 (2019).

[86] E. V. Anslyn and D. A. Dougherty, *Modern physical organic chemistry* (Mill Valley, CA: University science books, 2006).

[87] L. M. M. Quijano and D. A. Singleton, J. Am. Chem. Soc. **133**, 13824 (2011).

[88] O. Exner, in *Correlation Analysis in Chemistry. Recent Advances*, edited by N. B. Chapman and J. Shorter (Plenum Press, New York, 1978) pp. 439–540.

[89] L. Kleinman and D. M. Bylander, Phys. Rev. Lett. **48**, 1425 (1982).