Chemical Science

EDGE ARTICLE

View Article Online
View Journal | View Issue

Cite this: *Chem. Sci.*, 2017, **8**, 6924

Machine learning molecular dynamics for the simulation of infrared spectra†

Michael Gastegger,[a] Jörg Behler [b] and Philipp Marquetand *[a]

Machine learning has emerged as an invaluable tool in many research areas. In the present work, we harness this power to predict highly accurate molecular infrared spectra with unprecedented computational efficiency. To account for vibrational anharmonic and dynamical effects – typically neglected by conventional quantum chemistry approaches – we base our machine learning strategy on *ab initio* molecular dynamics simulations. While these simulations are usually extremely time consuming even for small molecules, we overcome these limitations by leveraging the power of a variety of machine learning techniques, not only accelerating simulations by several orders of magnitude, but also greatly extending the size of systems that can be treated. To this end, we develop a molecular dipole moment model based on environment dependent neural network charges and combine it with the neural network potential approach of Behler and Parrinello. Contrary to the prevalent big data philosophy, we are able to obtain very accurate machine learning models for the prediction of infrared spectra based on only a few hundreds of electronic structure reference points. This is made possible through the use of molecular forces during neural network potential training and the introduction of a fully automated sampling scheme. We demonstrate the power of our machine learning approach by applying it to model the infrared spectra of a methanol molecule, *n*-alkanes containing up to 200 atoms and the protonated alanine tripeptide, which at the same time represents the first application of machine learning techniques to simulate the dynamics of a peptide. In all of these case studies we find an excellent agreement between the infrared spectra predicted *via* machine learning models and the respective theoretical and experimental spectra.

Received 19th May 2017
Accepted 8th August 2017

DOI: 10.1039/c7sc02267k

rsc.li/chemical-science

1 Introduction

Machine learning (ML) – the science of autonomously learning complex relationships from data – has experienced an immensely successful resurgence during the last decade.[1,2] Increasingly powerful ML algorithms form the basis of a wealth of fascinating applications, with image and speech recognition, search engines or even self-driving cars being only a few examples. In a similar manner, ML based techniques have lead to several exciting developments in the field of theoretical chemistry.[3–7]

ML potentials are an excellent example of the benefits ML algorithms can offer when paired with theoretical chemistry methods.[8–16] These potentials aim to accurately reproduce the potential energy surface (PES) of a chemical system (and its forces) based on a number of data points computed with quantum chemistry methods. Due to the powerful non-linear learning machines at their core, ML potentials are able to retain the accuracy of the underlying quantum chemical method, but can be evaluated several orders of magnitude faster. This combination of speed and accuracy is especially advantageous in situations where a large number of costly quantum chemical calculations would be required.

One such case is *ab initio* molecular dynamics (AIMD), a simulation technique used to describe the evolution of chemical systems with time.[17] In AIMD, the motion of the nuclei is described classically according to Newton's equations of motion[18] and depends on the quantum mechanical force exerted by the electrons and nuclei. AIMD is a highly versatile tool and has been used to model a variety of phenomena like photodynamical processes or the vibrational spectra of molecules.[19–23]

The latter application is of particular interest in the field of vibrational spectroscopy. With the development of more and more sophisticated experimental techniques, it is now possible to use methods like infrared (IR) and Raman spectroscopy to obtain highly accurate spectra of macromolecular systems (*e.g.* proteins).[24,25] As a consequence, vibrational spectra have become increasingly complex and theoretical chemistry simulations are now an indispensable aid in their interpretation. Unfortunately, the standard approach to model vibrational

[a]*University of Vienna, Faculty of Chemistry, Institute of Theoretical Chemistry, Währinger Str. 17, 1090 Vienna, Austria. E-mail: philipp.marquetand@univie.ac.at; Fax: +43 1 4277 9527; Tel: +43 1 4277 52764*

[b]*Universität Göttingen, Institut für Physikalische Chemie, Theoretische Chemie, Tammannstr. 6, 37077 Göttingen, Germany*

† Electronic supplementary information (ESI) available. See DOI: 10.1039/c7sc02267k

spectra, static calculations based on the harmonic oscillator (HO) approximation, suffers from several inherent limitations.[21,26] Due to the HO approximation, anharmonic vibrational effects are neglected, which are of great importance in molecular systems with high degrees of flexibility and/or hydrogen bonding, such as biological systems. Moreover, HO based calculations are unable to account for conformational and dynamic effects, due to their restriction to one particular conformer. This also makes it hard to accurately model temperature effects, which have a large influence on conformational dynamics and are highly relevant for spectra recorded at room temperature.[20] These deficiencies lead to disagreements between experimental and theoretical spectra, thus complicating consistent analysis.

Different strategies, like the variational self-consistent field (VSCF) approach and its extensions,[26] as well as quantum dynamics based methods,[27,28] have been developed to account for these effects, but they either neglect dynamical effects or are computationally intractable for systems containing more than a few tens of atoms. Consequently, AIMD, which is able to describe anharmonicities and dynamic effects at manageable computational costs, is an invaluable tool for the practical simulation of vibrational spectra.[20,21]

Yet, standard AIMD is still comparatively expensive, placing severe restrictions on the maximum size of the systems under investigation (approximately 100 atoms) and on the quality of the quantum chemical method. Various techniques, such as compressed sensing[29] or harmonic inversion,[30] can be used to reduce the amount of AIMD samples required to obtain good quality spectra. However, these approaches are not able to overcome system size limitations. A more general alternative to significantly accelerate AIMD simulations without sacrificing chemical accuracy is to replace most electronic structure calculations with much cheaper ML computations. This opens the way for exciting new possibilities, making it possible to simulate larger systems and longer timescales in only a fraction of the original computer time.

The goal of the current work is to use ML accelerated AIMD calculations to simulate accurate IR spectra of different organic molecules. This is achieved by harnessing the synergies between established techniques, improvements to existing schemes and new developments: (I) a special kind of ML potential, called high-dimensional neural network potential (HDNNP), is used to model the PES.[31] (II) Molecular forces are employed in the construction of these HDNNPs, using a method based on the element decoupled Kalman filter.[32] (III) Electronic structure reference data points are selected *via* an enhanced adaptive sampling scheme for molecular systems. (IV) A HDNNP based fragmentation method is used to accelerate reference computations for macromolecules.[33] Finally, (V) a new ML scheme to model dipole moments is introduced. A detailed description of all of these individual components is given in the following section.

Three different molecular systems were studied using the strategies described above. First, a single methanol molecule served as a test case to assess the overall accuracy of the HDNNP based simulations compared to the spectra obtained with standard AIMD. Second, the ability of HDNNPs to efficiently deal with macromolecular systems was demonstrated by (a) constructing a HDNNP of a simple alkane chain based only on small fragments of the macromolecule and then (b) using the resulting model to predict the IR spectra of alkanes of varying chain lengths. In order to probe the suitability of HDNNPs for systems of biological relevance, a final study was dedicated to the protonated trialanine peptide. This also served as an excellent test case for the ML based dipole moment model.

Separate reference data sets are generated for each of the three systems. The system specific HDNNPs are constructed using density functional theory (DFT) as an electronic structure reference method. Generalized gradient functionals are used for methanol and the tripeptide. In the case of alkanes, we demonstrate that in principle highly accurate double-hybrid density functionals[34] can also be used. The simulations carried out with these latter HDNNPs would be next to impossible using on-the-fly AIMD. In all cases, comparisons to experimental IR spectra are shown.

2 Theoretical background

In AIMD, vibrational spectra are computed *via* the Fourier transformation of time autocorrelation functions.[21] Different physical properties give rise to different types of spectra. IR spectra depend on the molecular dipole moments:

$$I_{\mathrm{IR}} \propto \int_{-\infty}^{+\infty} \langle \dot{\boldsymbol{\mu}}(\tau)\dot{\boldsymbol{\mu}}(\tau+t)\rangle_{\tau} \mathrm{e}^{-\mathrm{i}\omega t}\mathrm{d}t \ , \qquad (1)$$

where $\dot{\boldsymbol{\mu}}$ is the time derivative of the molecular dipole moment, ω is the vibrational frequency, τ is a time lag and t is the time.

Upon closer examination of eqn (1), several challenges to model AIMD quality IR spectra *via* ML become apparent: reliable ML potentials (and especially forces) are required to describe the time evolution of a chemical system. Consequently, reference points need to be selected from representative regions of the PES, while keeping the number of costly electronic structure calculations to a minimum. This also calls for efficient strategies to handle the reference calculations of large molecules. Finally, a method to accurately model molecular dipole moments is required.

2.1 High-dimensional neural network potentials

In a HDNNP (shown in Fig. 1), the total potential energy E_{pot} of a molecule is expressed as a sum of individual atomic energies.[31,35] The contribution E_i of every atom depends on its local chemical environment and is modeled by a neural network (NN). These atomic NNs are typically constrained to be the same for a given element and thus are also termed elemental NNs. Due to this unique structure, HDNNPs can easily adapt to molecules of different sizes and even be transferred between sufficiently similar molecular systems.

The chemical environment of an atom is represented by a set of many-body symmetry functions $\{\boldsymbol{G}_i\}$, so-called atom-centered symmetry functions (ACSFs).[36] ACSFs depend on the positions $\{R_i\}$ of all neighboring atoms around the central atom, up to

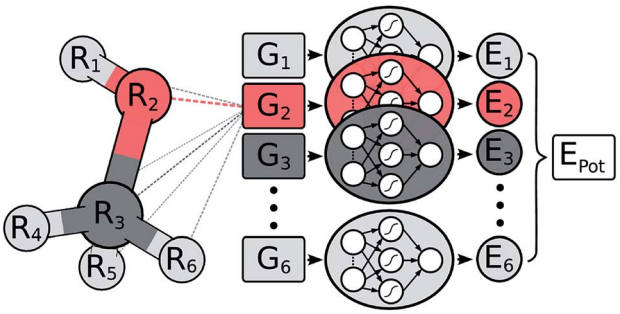

Fig. 1 Schematic representation of a high-dimensional neural network potential (HDNNP). The Cartesian coordinates R are transformed into many-body symmetry functions $\{G_i\}$ describing an atom's chemical environment. Based on these functions, a NN then predicts the energy contribution E_i associated with atom i. The potential energy E_{pot} of the whole molecule is obtained by summing over all individual atomic energies.

a predefined cutoff radius. By introducing a cutoff radius, an atom's environment is restricted to the chemically relevant regions. This brings two distinct advantages: the computational cost of HDNNPs now scales linearly with molecular size and chemical locality can be exploited in their construction and application,[8] which has been demonstrated recently *e.g.* for alkanes.[33] In addition, HDNNPs are well suited for molecular dynamics simulations, since an analytical expression for molecular forces is available due to their well-defined functional form. For a detailed discussion of HDNNPs and ACSFs, see ref. 35.

In order for HDNNPs to yield reliable models of the PES, a set of optimal parameters needs to be determined for the elemental NNs. This is done in a process called training, where a cost function (typically the mean squared error) between the reference data points (*e.g.* energies and forces) and the HDNNP predictions is minimized iteratively. Different algorithms can be used to carry out the minimisation. The current work uses the element-decoupled Kalman filter,[32] a special adaptation of the global extended Kalman filter[37] for HDNNPs.

Besides the energies, it is also possible to include molecular forces in the training process, by minimizing the cost function[35]

$$\mathscr{C}_{\mathrm{E,F}} = \frac{1}{M}\sum_{m}^{M}(\tilde{E}_m - E_m)^2 + \frac{\eta}{M}\sum_{m}^{M}\frac{1}{3N_m}\sum_{\alpha}^{3N_m}(\tilde{F}_{m\alpha} - F_{m\alpha})^2. \quad (2)$$

The first term on the right hand side corresponds to the mean squared error between the reference energies E and HDNNP energies $\tilde{E}$. The second term describes the deviation between the HDNNP $(\tilde{F})$ and quantum chemical forces (F). M is the number of molecules in the reference data set, N the number of atoms in a molecule, and α is an index running over the $3N$ Cartesian force components. η is a constant used to tune the importance of the force error on the update step. Including the forces in the training process leads to substantial improvements in the forces predicted by the HDNNP. Furthermore, instead of only one single energy, $3N$ points of additional information per molecule can now be utilized during training, thus greatly reducing the number of reference points required for a converged potential. An in-depth description of the element-decoupled Kalman filter and its extension to molecular forces can be found in ref. 32.

2.2 Adaptive selection scheme

Ultimately, the quality of a ML potential does not only depend on the underlying ML algorithm and the employed training procedure, but also on how well the reference data set represents the chemical problem under investigation. Ideally, the reference data span all relevant regions of the PES with as few data points as possible to avoid costly electronic structure computations. To this end, different strategies – *e.g.* based on Bayesian inference[38] or geometric fingerprints[39] – have been developed in the past.

A simple but relatively effective procedure to select data points is based on the use of multiple HDNNPs and is described for example in ref. 35. After choosing an initial set of reference data points, a set of preliminary HDNNPs is trained, differing in the initial parameters and/or architectures of their elemental NNs (Fig. 2). These proto-potentials are then used to sample different molecular conformations, using *e.g.* molecular dynamics simulations. Afterwards, the predictions of the HDNNPs are compared to each other. Regions of the PES where the different HDNNPs agree closely are assumed to be represented well, whereas conformations with diverging HDNNP predictions are modeled inaccurately. The inaccurately described conformations are recomputed with the electronic structure reference method and

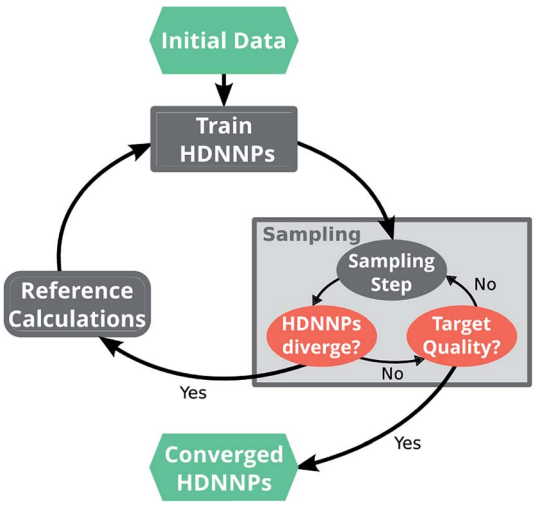

Fig. 2 A typical run of the adaptive selection scheme starts by using a small set of initial reference data points to train a preliminary ensemble of HDNNPs. These HDNNPs are then used to sample new molecular conformations (*e.g. via* molecular dynamics simulations). During sampling, the predictions of the individual potentials are monitored and if divergence is detected, the sampling run is stopped. The conformation for which the HDNNPs disagree is computed with the electronic structure reference method and added to the set of reference points. Subsequently, the HDNNP ensemble is retrained on the expanded data set and sampling is continued with the new potential. This procedure is repeated in an iterative manner, until the divergence stops to exceed a predetermined threshold.

added to the reference data set. The HDNNPs are then retrained using the expanded data set and the process is repeated in a self consistent manner until the HDNNPs reach the desired quality.

The current work introduces small adaptations to this procedure in order to make it more suitable for use with biomolecules and expensive reference methods. Instead of performing independent sampling simulations with the individual HDNNPs, they are combined into an ensemble. In the ensemble, the energy and forces are computed as the average of the J different HDNNP predictions:

$$\overline{E} = \frac{1}{J}\sum_{j=1}^{J}\tilde{E}_j, \tag{3}$$

$$\overline{\mathbf{F}} = \frac{1}{J}\sum_{j=1}^{J}\tilde{\mathbf{F}}_j. \tag{4}$$

Simulations are then carried out using these averaged properties. The prediction uncertainty of the HDNNP ensembles is defined as

$$E_\sigma = \sqrt{\frac{1}{J-1}\sum_{j}^{J}\left(\tilde{E}_j - \overline{E}\right)^2}. \tag{5}$$

The use of the HDNNP ensembles and the above uncertainty measure offers two advantages: first, the reliability of the uncertainty measure increases with the number of basic HDNNPs. The more HDNNPs are used, the more unlikely it becomes that they exhibit similar behavior in underrepresented regions of the PES. Second, ensembles are less susceptible to errors in their individual components, since these errors tend to cancel to a certain degree. This leads to a significant improvement of the prediction accuracy (reducing the error up to a factor of $\frac{1}{\sqrt{J}}$ in some cases) at negligible extra cost. This effect leads to more reliable simulations, especially in the early stages of the PES exploration, hence diminishing the number of electronic structure starting points needed to seed the self-consistent refinement procedure. As a consequence, HDNNPs can now be grown on-the-fly from only a handful of data points in a highly automated manner: starting from *e.g.* a few molecular dynamics steps, HDNNP ensemble simulations are run until E_σ of a visited structure exceeds a predefined threshold. The corresponding conformation is recomputed with the reference method and added to the training set. The HDNNPs are retrained and simulations are continued from the problematic conformation. Finally, once a converged HDNNP ensemble has been obtained in this way, it is used to simulate the properties of interest.

This procedure is effective but highly sequential and calculations using expensive reference methods constitute a significant bottleneck. Under the assumption that the approximate shape of the PES is sufficiently similar for different electronic structure methods, an "upscaling" step is introduced. First, the iterative refinement is carried out using a low-level method until convergence of the HDNNPs. The conformations obtained in this manner are then recomputed using a high-level method. Since these high-level calculations can be done in parallel, the overall procedure is highly efficient. Afterwards, new HDNNPs are trained, now at the quality of the better method. The above assumption with regard to the similar shape of the PES at the different levels of theory is not necessarily valid, hence an upscaling step is typically followed by additional refinement steps at the higher level of theory.

A detailed discussion on the performance of the adaptive selection scheme and the convergence of the ML predictions with ensemble size can be found in the ESI.†

2.3 Fragmentation with high-dimensional neural network potentials

Since the computational cost of electronic structure calculations scales very unfavorably with the system size and accuracy of the underlying method, individual reference computations can still be problematic. Hence, the required reference computations would quickly become intractable for highly accurate HDNNPs describing large molecular systems, despite the efficient sampling scheme.

It is possible to circumvent this problem by exploiting the special structure of HDNNPs. As a consequence of expressing the HDNNP energy as a sum of atomic contributions and introducing a cutoff radius, HDNNPs operate in the same manner as fragmentation methods using a divide and conquer approach: given only the energies of small molecular fragments, HDNNPs can reconstruct the energy of the total system.[8,33] Thus, expensive electronic structure calculations never have to be performed for the whole molecule, but only for small parts of it. The result is a linear scaling of the computational effort with system size. Similar fragmentation strategies are employed by other ML models.[12,40–42]

In practice, a molecule is first divided into its individual fragments. Reference computations are then carried out for these fragments and the resulting data set is used to train a HDNNP. The ML potential is then applied to the geometry of the original molecule and the energy of the full system is recovered in this way. Different strategies can be used to partition the full molecular system. In the current work, every molecule is split into N atom-centered fragments (see Fig. 3). The size and shape of these fragments are determined by a cutoff radius around the central atom. Atoms beyond the cutoff radius are removed and free valencies are saturated with hydrogen atoms. If a free valency is situated on a hydrogen atom or two capping hydrogens overlap, the heavy atom corresponding to this position is instead included in the fragment and the process is repeated iteratively. Typically, the same cutoff radius as that in the ACSFs is used.

HDNNP fragmentation can easily be integrated into the adaptive sampling scheme. Using the deviations in atomic forces predicted by different HDNNPs as uncertainty measures, inaccurately modeled fragments can be identified. These fragments are then added to the reference data set.

2.4 Neural network dipole moments and charge analysis

A vital ingredient in the simulation of IR spectra with AIMD is the molecular dipole moment (see eqn (1)). While strategies to

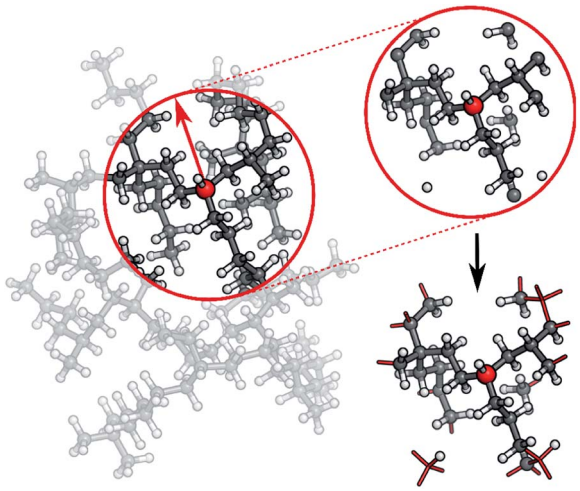

Fig. 3 In order to generate molecular fragments, first all atoms beyond a predetermined cutoff radius from the central atom are removed. Afterwards, free valencies are saturated with hydrogen atoms, unless the valency itself is situated on a hydrogen or corresponds to a double bond in the unfragmented molecule. In this case, the heavy atom connected to this atom in the original molecule is included in the fragment and the process is repeated iteratively. This procedure is performed for the whole system, leading to one fragment per atom.

predict dipole moments using NNs exist,[43,44] HDNNPs themselves have only been used to predict environment dependent charges in full analogy to the atomic energy contributions with the aim to model electrostatic long range interactions.[45,46]

In this work, we extend this approach, by constructing molecular dipole moments as a sum of such environment dependent atomic partial charges:

$$\tilde{\boldsymbol{\mu}} = \sum_{i}^{N} \tilde{q}_i \mathbf{r}_i, \tag{6}$$

where $\tilde{q}_i$ is the charge of atom i modeled by a NN and $\mathbf{r}_i$ is the distance vector of the atom from the molecule's center of mass.

While the elemental charge NNs could in principle be trained to reproduce charges computed with quantum chemical charge partitioning schemes (as was *e.g.* done in ref. 47 to model electrostatic interactions), this approach has the following problems: first, the charge of a given atom obtained with such a partitioning scheme can in principle change along a trajectory in a non-continuous manner (*e.g.* depending on the local minima of the fit found when determining the atomic charges in such methods as CHELPG[48]). The resulting inconsistencies in the reference data can in turn lead to erratic predictions of the final machine learning model. Second, unlike molecular energies and forces, atomic partial charges are not quantum mechanical observable. Hence, there is no physically unique way to determine them and a variety of different partitioning schemes exists.[49] This complicates the choice of a suitable method to compute reference charges, since different schemes often exhibit vastly different behavior and sometimes fail to reproduce the molecular dipole moment accurately.[50]

Both problems can be avoided by training the elemental NNs to reproduce the molecular moments directly, while the environment dependent atomic charges $\tilde{q}_i$ are inferred in an indirect manner. In order to achieve this, a cost function of the form

$$\mathscr{C}_{\mathrm{Q}} = \frac{1}{M}\sum_{m}^{M}\left(\tilde{Q}_m - Q_m\right)^2 + \frac{1}{3M}\sum_{m}^{M}\sum_{l}^{3}\left(\tilde{\mu}_{lm} - \mu_{lm}\right)^2 + \ldots \tag{7}$$

is minimized. Here, Q_m and μ_{lm} are the reference total charge and dipole moment components of molecule m. The index l runs over the three Cartesian components of the dipole moment. $\tilde{Q}$ is the total charge of the composite NN model, computed as $\tilde{Q} = \sum_{i}^{N} \tilde{q}_i$, while $\tilde{\boldsymbol{\mu}}$ is the NN dipole moment (eqn (6)). While the cost function (from eqn (7)) can be easily extended to include higher multipole moments, it was found that including only the total molecular charge and dipoles is sufficient for the purpose of modeling IR spectra. Since this scheme depends exclusively on molecular moments which are quantum mechanical observable, charge partitioning is no longer required. On the contrary, the trained NN model itself constitutes a new kind of partitioning scheme, where the atomic partial charges q_i depend on the chemical environment and are determined on a purely statistical basis. These charges can also be used for additional purposes, *e.g.* to compute electrostatic interactions. Another possible application would be to augment classical force fields,[43] where partial charges typically do not change with the chemical environment.[51] As such, the NN charge scheme presented here constitutes an interesting alternative to static point charges or polarizable models.[52]

3 Computational details

Electronic structure reference calculations were carried out with ORCA[53] at the BP86/def2-SVP[54–59] (methanol and alanine tripeptide), BLYP/def2-SVP[54–56,60] (Ala_3^+) and B2PLYP/def2-TZVPP[34,59] (*n*-alkanes) levels of theory. All calculations were accelerated using the resolution of identity approximation.[61,62]

All HDNNPs were constructed and trained with the RUNNER program.[63] The NN dipole models were implemented in python[64] using the numpy[65] and theano[66] packages. Reference data points were obtained with the adaptive selection scheme, employing molecular dynamics trajectories at a temperature of 500 K with a 0.5 fs timestep to sample relevant conformations. The final ML models are based on 245 (methanol), 534 (*n*-alkanes) and 718 (peptide) reference data points, with a maximum network size of 35-35-1 (two hidden layers with 35 nodes each and one node in the output layer) for the HDNNPs and 100-100-1 for the dipole moment model.

IR spectra were obtained with molecular dynamics simulations in the gas phase employing the same timestep as the sampling procedure. After a short initial equilibration period (3 ps for methanol and 5 ps otherwise), constant temperature molecular dynamics simulations were run for 30 ps in the case of methanol and 50 ps in the case of the other molecules. In addition to ML accelerated dynamics, AIMD simulations were carried out for methanol using the BP86 level of theory described above.

Detailed information regarding the setup of the electronic structure calculations and molecular dynamics simulations as well as the ML models can be found in the ESI.†

4 Results and discussion

4.1 Methanol

Due to its small size, the methanol molecule constitutes an excellent test system, not only for the direct comparison between the IR spectra obtained *via* standard AIMD and ML simulations, but also to investigate the overall accuracy of the ML approximations.

The final ML model for methanol consists of two HDNNPs and a NN dipole moment model trained on the BP86 data for 245 configurations. To assess the errors associated with the individual components of the model, a standard AIMD simulation is run for 30 ps, producing 60 000 configurations. For the sampled geometries, energies, forces and dipoles are predicted with the ML model. These predictions are then compared to the respective electronic structure results. The distribution of errors between the ML predictions and the BP86 method are shown in blue in Fig. 4.

Excellent agreement between BP86 calculations and the ML model is found for all investigated properties. In the case of energies (Fig. 4a), the mean absolute error (MAE) of 0.048 kcal mol^{-1} (range of energies 13.620 kcal mol^{-1}) is well below the commonly accepted limit for chemical accuracy (1 kcal mol^{-1}) and is expected to be negligible compared to the intrinsic error of the BP86 reference method in practical applications. The components of the force vectors are reproduced equally well (Fig. 4b), with a MAE of 0.533 kcal mol^{-1} $Å^{-1}$ (range 242.34 kcal mol^{-1} $Å^{-1}$). These findings are comparable with other state of the art ML learning strategies developed specifically for the modeling of forces[67] and demonstrate the excellent capabilities of HDNNPs to create potentials suitable for the dynamical simulation of molecules. This conclusion is also supported by a comparison of the normal mode frequencies obtained for the optimized methanol structure at the ML- and BP86-level (see Table 1). Although the HDNNP model was never explicitly trained to reproduce normal mode frequencies, its predictions agree well with the reference frequencies, exhibiting a maximum deviation of only 31.38 cm^{-1} (0.090 kcal mol^{-1}). The new NN dipole model is also found to provide an accurate description of the molecular dipole moments (Fig. 4c). The total dipole moment shows an overall MAE of 0.016 D (over a range of 0.723 D) and the spatial orientation of the dipole vector is modeled equally reliably, with the MAEs of the individual Cartesian components ranging from 0.0173 D to 0.0200 D. The small shift of the dipole error distribution towards negative values is due to the fact that the atomic charges fluctuate around values other than zero. This effect is enhanced further, by the final summation to obtain the dipole moment model (see eqn (6)). Further evidence for the high efficacy of the dipole moment model is provided by the small deviations between the static IR intensities obtained for the optimized methanol at the ML- and BP86-level (see Table 1). However, care should be taken, as these values have been derived within the harmonic oscillator approximation and serve the sole purpose of analyzing the accuracy of the ML model.

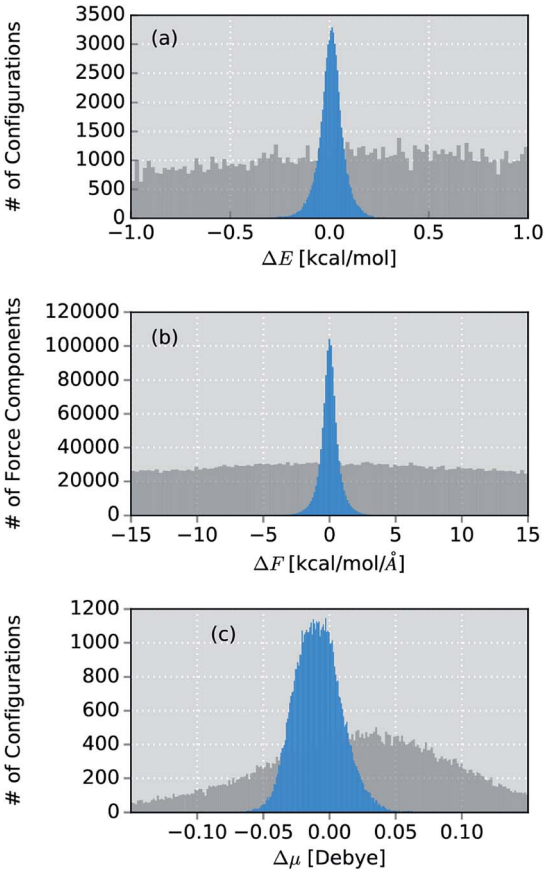

Fig. 4 Distribution of errors between the ML model based on the adaptive sampling scheme and the BP86 reference (blue). The deviations were computed based on the energies, forces and dipole moments (from top to bottom) of 60 000 configurations of methanol sampled with an AIMD simulation. The deviations obtained with a ML model trained on data points selected at random from a force field simulation are shown in grey (see ESI†).

Table 1 Comparison of the normal mode frequencies and IR intensities of methanol obtained with DFT and the ML model within the harmonic oscillator approximation

	$\tilde{\omega}$ [cm^{-1}]			I [km mol^{-1}]		
#	BP86	ML	$\Delta\tilde{\omega}$	BP86	ML	ΔI
1	331.70	346.94	−15.24	119.94	117.96	1.99
2	1037.82	1030.00	7.82	90.89	81.72	9.16
3	1080.46	1092.09	−11.63	34.31	53.33	−19.02
4	1135.08	1138.21	−3.13	0.35	0.08	0.27
5	1328.95	1320.84	8.11	23.97	44.70	−20.73
6	1420.02	1416.42	3.60	1.74	8.15	−6.41
7	1427.64	1422.59	5.05	5.96	2.31	3.66
8	1449.79	1449.02	0.77	8.63	3.24	5.39
9	2880.76	2892.94	−12.18	74.67	65.10	9.58
10	2930.10	2961.48	−31.38	85.43	67.65	17.78
11	3034.15	3054.08	−19.93	29.45	31.19	−1.75
12	3707.93	3707.73	0.20	21.29	19.89	1.39

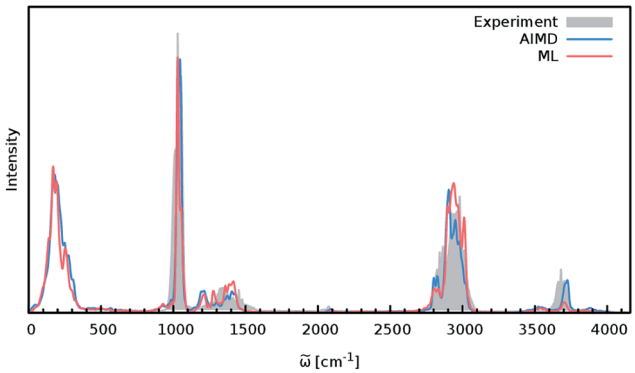

Fig. 5 IR spectra of the methanol molecule. The ML spectrum (red) is able to reproduce the AIMD spectrum (blue) obtained with BP86 with high accuracy. In addition, both theoretical spectra agree well with the experimental one recorded in the regions between 600 cm^{-1} to 4100 cm^{-1} (grey).

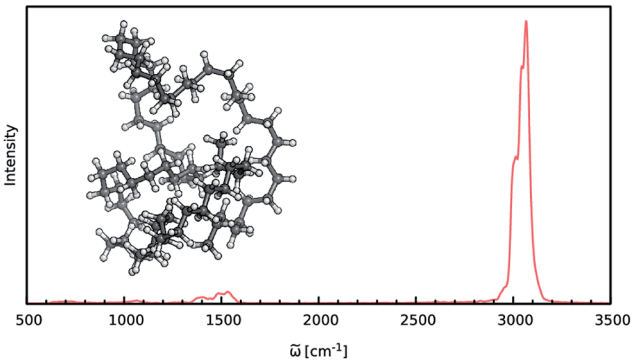

Fig. 6 IR spectrum of the $C_{69}H_{140}$ alkane as predicted by the ML model based on the B2PLYP method.

In order to study the quality of an IR spectrum modeled with the composite ML model, it is compared directly to the spectrum obtained *via* the BP86 AIMD simulation. Fig. 5 shows both IR spectra alongside an experimental spectrum of methanol recorded in the gas phase.[68] While the whole spectral range is covered for both theoretical spectra, the experiment only spans the region from 600 cm^{-1} to 4100 cm^{-1}. The overall shape of the ML spectrum, as well as the peak positions and intensities, shows excellent agreement with the electronic structure reference. The most distinctive difference between the QM and ML spectra is the intensity of the stretching vibration of the O–H bond observed at 3700 cm^{-1}. This relatively minor deviation is most likely caused by small deviations of the dipole moment model. Overall, the ML approach presented here is able to reproduce the AIMD IR spectrum of methanol with high accuracy. These results are remarkable insofar as the final ML model is based on only 245 electronic structure calculations. This demonstrates the effectiveness of the combination of HDNNPs and the NN dipole model, as well as the power of the improved sampling scheme.

Finally, both simulations agree well with the experimental spectrum, serving as an example of the utility of AIMD and ML accelerated AIMD for the prediction of accurate vibrational spectra.

4.2 *n*-Alkanes

When constructing ML potentials for large molecular systems containing hundreds or thousands of atoms, the necessary electronic structure reference calculations can quickly become intractable, especially for high-level methods. HDNNPs, as well as the dipole moment model presented in this work, can overcome this limitation *via* their implicit use of fragmentation (see Section 2.3). In order to demonstrate the potential of this approach, it is used to predict the IR spectrum of an *n*-alkane with the chemical formula $C_{69}H_{140}$ (depicted in Fig. 6) *via* ML accelerated AIMD simulations based on the B2PLYP double-hybrid density functional method.

The two HDNNPs and NN dipole moment model constituting the final ML model were trained on reference calculations for 534 fragments of the *n*-alkane. These fragments use a cutoff radius of 4.0 Å and contain 37 atoms on average and a maximum of 70 atoms. After initial adaptive sampling at the BP86/def2-SVP level, the final B2PLYP/TZVPP level ML model is obtained *via* an upscaling step described in Section 2.2. Dispersion interactions, which are expected to play an important role in molecular systems of this size, are accounted for *via* a simple scheme: the HDNNPs are constructed from standard B2PLYP calculations and augmented with the empirical D3 dispersion correction using Becke–Johnson damping[69,70] in an *a posteriori* fashion.

The IR spectrum of the $C_{69}H_{140}$ *n*-alkane predicted *via* ML is shown in Fig. 6. It exhibits all of the spectroscopic features typical for simple hydrocarbons: the intense peak at 3000 cm^{-1} corresponds to symmetric and asymmetric C–H stretching vibrations. Deformations of the CH_2-groups give rise to the bands close to 1500 cm^{-1}, while the extremely weak signals in the vicinity of 1000 cm^{-1} and 600 cm^{-1} are generated by C–C bond stretching and CH_2 rocking vibrations.

Although the general shape and features of the IR spectrum are described well by the ML model, some peak positions deviate from the expected experimental frequencies. This effect is especially pronounced for the C–H stretching vibrations, which are blue-shifted from the typical experimental value of 2900 cm^{-1} to 3040 cm^{-1}. This blue shift is due to the B2PLYP method or the classical equations of nuclear motion (and not an artifact introduced by the ML approximations), as will be explained in the following. Direct AIMD simulations and even static frequency calculations are prohibitively expensive for the $C_{69}H_{140}$ molecule. Instead, we exploit the transferability of the combined HDNNP and dipole model and use it to simulate the IR spectrum of the much smaller *n*-butane, for which theoretical and experimental spectra can be obtained easily. Fig. 7 shows the *n*-butane IR spectra obtained with ML accelerated AIMD and static electronic structure calculations and the experimental spectrum[68] (for a direct comparison of the static ML and B2PLYP spectra, see the ESI†). The blue shift of the C–H stretching vibrations present in the ML spectrum can also be found in the static B2PLYP spectrum. Moreover, both spectra show good agreement with each other with respect to the overall positions of the spectral peaks. These findings support the

conclusion that the observed frequency shifts are indeed a consequence of the underlying electronic structure method or the classical description of the nuclear dynamics[71] and not an artifact of the ML approximation. Furthermore, the ML accelerated AIMD approach is found to accurately reproduce the structure of the experimental vibrational bands (especially the C–H stretching vibrations, see the inset of Fig. 7). This is not the case for the static spectrum and shows that even for relatively small molecules an accurate description of the dynamic effects is important in order to obtain high-quality IR spectra. Both observations demonstrate the excellent accuracy of the HDNNP and NN dipole model, even for molecular systems not encountered during training.

Finally, to demonstrate the power of the ML based approach in general and the fragmentation based approach in particular, a few exemplary timings are given for the $C_{69}H_{140}$ molecule (using a single core of an Intel Xeon E5-2650 v3 CPU): obtaining the relevant molecular fragments using the iterative sampling scheme takes approximately 7 days. The reference calculations of the fragments on the B2PLYP level of theory can be carried out in a highly parallel manner within 1.2 days (using a single CPU per configuration), including the time necessary to construct the final ML model. ML accelerated AIMD simulations for the $C_{69}H_{140}$ molecule which involve the calculation of 110 000 energies and forces (5 ps equilibration and 50 ps simulation) take 3 hours. The NN dipole moments can be obtained within half an hour. Including the generation of the model, the total time to obtain the ML based IR spectrum amounts to a little over 8 days. In contrast, the evaluation of a single energy and gradient at the B2PLYP level for the full *n*-alkane would require 30 days, extrapolating from the timing of the fragment reference calculations. Hence, performing the 110 000 calculations necessary for the AIMD simulation would require a total of 3.3 million days or 9041 years. Using a conventional fragmentation method (*e.g.* the systematic fragmentation method) and assigning every fragment to an individual core, the total computation time of one configuration of the *n*-alkane at the B2PLYP level can be reduced to 1.2 days, leading to an overall simulation time of 361 years. Although this leads to a speedup of a factor of 25 compared to the unfragmented B2PLYP calculations, HDNNP simulations are still several orders of magnitude faster, once again demonstrating their excellent computational efficiency. An even more convincing picture for the efficacy of the current ML approach is painted by comparing the number of finite difference calculations required to obtain a static electronic structure spectrum to the number of samples contained in the ML model (see also the ESI†): using analytical molecular forces to construct finite difference Hessians, 1254 electronic structure computations need to be performed in the case of a static quantum chemical spectrum, while the ML model requires less than half of this number (534) to provide an accurate spectrum.

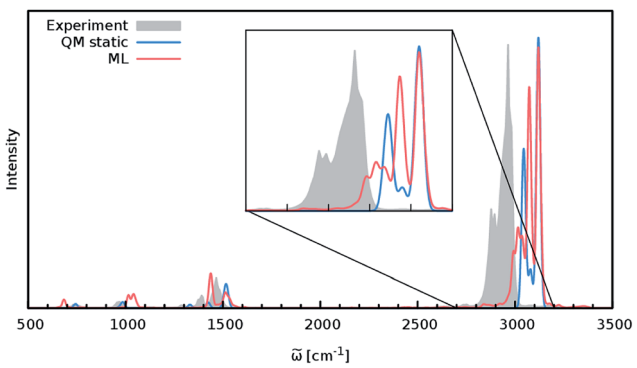

Fig. 7 IR spectrum of *n*-butane obtained *via* the ML model (red), compared to the static quantum mechanical spectrum computed at the B2PLYP level (blue) and convoluted with Gaussians. The peak positions in the ML and B2PLYP spectra agree closely, suggesting that the observed deviations from the experimental spectrum (grey) are not caused by the ML approximation. The overall structure of the peaks is reproduced much better by the ML accelerated AIMD simulation, especially in the region of the C–H stretching vibrations (see inset).

4.3 Protonated alanine tripeptide

Vibrational anharmonicities as well as conformational and dynamic effects play a crucial role in the vibrational spectra of biomolecules. In order to investigate the ability of ML accelerated AIMD to account for these effects, the composite ML model is used to simulate the IR spectrum of the protonated alanine tripeptide molecule (Ala_3^+) in the gas phase. Modeling the Ala_3^+ molecule poses several challenges: an accurate description of the complicated PES depends crucially on the ability of the adaptive sampling scheme and the HDNNPs to reliably identify and interpolate relevant electronic structure data points. Moreover, the changing charge distribution and dipole moment of the protonated species need to be captured by the NN dipole model. Since the IR spectrum of Ala_3^+ has been studied extensively, both experimentally and theoretically,[72,73] the quality of the ML approach can be assessed directly.

The composite Ala_3^+ ML model consists of two HDNNPs and a NN dipole model and was constructed from 717 reference geometries selected with the adaptive sampling scheme. The model exhibits overall RMSEs of 1.56 kcal mol^{-1}, 3.40 kcal mol^{-1} Å^{-1} and 0.26 Debye for the energies, forces and dipoles respectively. This increase in the RMSEs and number of required data points compared to the previous systems is an indicator for the chemical complexity of the peptide. Long range dispersion interactions were accounted for in the same manner as in the case of the *n*-alkanes.

Previous theoretical studies by Vaden and coworkers[73] have found that the experimental IR spectrum of Ala_3^+ is primarily composed of the contributions of three different conformers: (1) an elongated Ala_3^+ chain with the proton situated at the N-terminal amine group, (2) a folded chain protonated at the same site and (3) an elongated form in which the proton is located at the carbonyl group of the N-terminus (see Fig. 8), which will be referred to as the NH_3, folded and NH_2 families henceforth. In order to account for these effects, ML accelerated AIMD simulations were carried out for all three conformers at 350 K, the estimated experimental temperature. The final ML IR spectrum was then obtained by averaging. Fig. 8 shows the overall spectrum, as well as the contributions of the individual

conformations alongside the experimental spectrum.[73] Due to the range of the recorded spectrum and the high congestion of spectral bands in the regions of the lower vibrational modes, we restrict our discussion only to the stretching modes involving hydrogens (*ca.* 2700 cm^{-1} to 3700 cm^{-1}). An analysis of the static spectra computed for the full spectral range at the ML and BLYP level can be found in the ESI.†

As can be seen, the ML model correctly captures the features present in the experimental spectrum. The intense peak at 3570 cm^{-1} is due to the O–H stretching vibrations of the carboxylic acid group of the C-terminus. The position as well as the slight asymmetry of this band is almost perfectly reproduced in the ML spectrum. The region from 3300 cm^{-1} to 3500 cm^{-1} is populated by signals arising from the stretching modes of N–H bonds not participating in hydrogen bonds (*e.g.* NH_2 terminus in the NH_2 family). The free N-terminal N–H groups of the NH_3 and folded family give rise to the intense feature at 3420 cm^{-1}. Compared to the experimental spectrum, the region ranging from 3250 cm^{-1} to 3350 cm^{-1} is underpopulated in the BLYP simulation. This deviation is primarily a consequence of the employed electronic structure method. As can be seen in Fig. 8, the description of this region is extremely method dependent and changing to a model based on the BP86 functional (see discussion below) leads to an increased population of the corresponding bands. In addition to the choice of methods, temperature effects seem to play a role, as increasing the simulation temperature also populates these bands to a certain extent. The experimental temperature of 350 K reported for this systems is only an averaged estimate and higher temperatures for the individual conformers might indeed be possible. While this distribution of temperatures might be accounted for *via* a trial and error procedure, the exact reproduction of the experimental conditions is not the ultimate goal of this study. Vibrations associated with the N–H groups directly involved in hydrogen bonds are situated in the region from 3100 cm^{-1} to 3300 cm^{-1}, where the ML spectrum captures several experimental subpeaks. Finally, the region from 2800 cm^{-1} to 3100 cm^{-1} corresponds to the C–H stretching vibrations. Here, the most distinct features are the peak at 2930 cm^{-1} due to the C–H vibrations of the C_α groups and the peak at 2970 cm^{-1}, which is caused by the vibrations of the methyl group hydrogens. The generally good agreement between the ML and experimental spectrum and the ability to reliably resolve individual bands is a testament for the efficacy of the composite ML scheme introduced in this work: the dipole model is able to describe the charge distribution of Ala_3^+ accurately, while the HDNNP ensemble provides a reliable approximate PES.

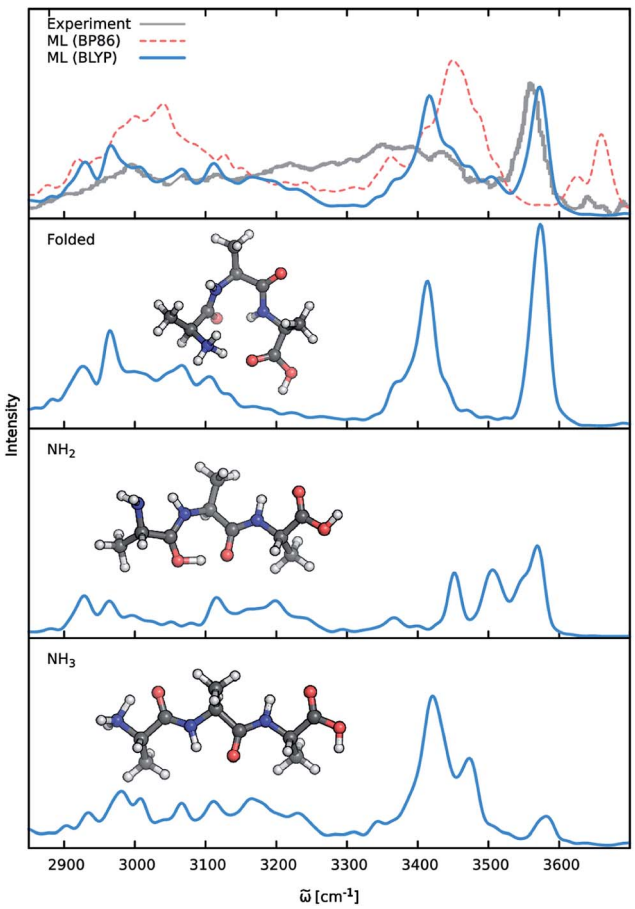

Fig. 8 IR spectra of the protonated alanine tripeptide. The top panel shows the experimental spectrum (gray), as well as the ML spectra based on the BLYP (blue) and BP86 (red) reference methods. The lower panels depict the structures of the three main Ala_3^+ conformers, along with their respective contributions to the averaged BLYP ML spectrum.

A good perspective on the accuracy of the ML approach can also be gained by comparing the current ML model to one based on a different electronic structure reference method. The top panel of Fig. 8 shows the averaged IR spectrum predicted by a ML model based on the BP86 density functional next to the previously discussed BLYP spectrum. Although one would expect the closely related BLYP and BP86 methods to give similar results, significant differences can be found: besides a blue shift of the signal caused by the C-terminal COOH group by almost 80 cm^{-1} compared to the BLYP spectrum and experimental spectrum, large deviations are also found in the shape and positions of the bands corresponding to the N–H stretching vibrations. Here, we investigate the cause of the latter effect by a closer examination of the NH_3 conformer. Since the hydrogens of the N-terminal NH_3 group can be involved in a proton transfer event to the neighboring carbonyl group, different spectra can arise depending on how often this transfer occurs. The transfer rate is directly correlated to the energy barrier associated with the transfer, suggesting that BLYP and BP86 differ significantly in the description of this event, which in turn leads to differences in the ML spectra. Whether this phenomenon is caused by the ML approximations or due to the BP86 method itself can easily be verified by computing the proton transfer barriers with both electronic structure methods and ML models. As can be seen in Fig. 9, the barrier height is indeed underestimated by the BP86 functional compared to BLYP, giving rise to the observed behavior. At the same time, the ML models faithfully reproduce the barriers found with their respective reference methods. This is an excellent demonstration for the reliability of the ML approach, since the deviation between the ML model and reference method is actually

negligible compared to the differences between two closely related electronic structure methods. The ease with which ML of different QM methods can be generated also suggests a potential use of the ML approach presented here as an efficient tool for extensively comparing and thus benchmarking electronic structure methods. Additional ML models can simply be constructed by recomputing the representative conformations selected by the sampling scheme with a different method in a parallel fashion and subsequent retraining of the new model (see Section 2.2). Possible applications of this finding will be explored in the future.

The above observations also serve to highlight the ability of the ML model to automatically infer the chemistry underlying the Ala_3^+ system. Proton transfer events are essential in characterizing the experimental spectrum.[72] Driven by the automated sampling scheme, the composite ML approach gradually learns to describe these relevant chemical events, as is nicely demonstrated based on the reaction barrier previously obtained for the NH_3 transfer (Fig. 9): although the description of this event was never explicitly targeted in the training procedure, the barrier is nevertheless reproduced to an excellent degree of accuracy. This feat is impressive insofar as the ML model is based on a relatively small set of *ab intio* computations. These findings also serve to highlight an important advantage of HDNNPs over typical classical force fields, which is the ability to describe bond breaking and formation reactions.

Once again, the excellent computational efficiency of the composite ML model should be stressed: while the computational chemistry method employed for Ala_3^+ is already considered to be relatively cheap, the speedup gained is still significant. A single step in the BP86 simulation takes approximately 1.5 minutes (on a single Intel Xeon E5-2650 v3 CPU). The dynamics of every Ala_3^+ conformer are simulated for 55 ps, requiring a total of 110 000 steps. This amounts to a simulation time of 114 days for full AIMD. In contrast, using the ML model one can perform the same simulation in only one hour.

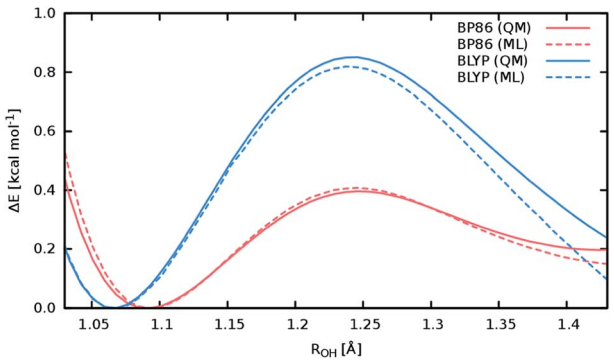

Fig. 9 Reaction barriers associated with the proton transfer from the N-terminal NH_3 group in the NH_3 conformer of Ala_3^+ to the neighboring carbonyl. The reaction coordinate is the distance between the transferred NH_3 hydrogen and the carbonyl oxygen. The barriers computed with the electronic structure reference methods are shown as solid lines colored red for the BLYP method and blue in the case of the BP86 method. The dashed curves correspond to the predictions of the respective ML models, maintaining the above color scheme.

5 Conclusions

Here, we present the first application of machine learning (ML) techniques to the dynamical simulation of molecular infrared spectra. We find that our ML approach is able to predict infrared spectra of various chemical systems in a highly reliable manner, correctly describing anharmonicities, as well as dynamic effects, such as proton transfer events. The excellent accuracy – which is only limited by the underlying computational chemistry method – is paired with high computational efficiency, reducing the overall computation time by several orders of magnitude. This makes it possible to treat molecular systems that are usually beyond the scope of standard electronic structure methods. As a proof of principle, we have simulated *n*-alkanes containing several hundreds of atoms, as well as the protonated alanine tripeptide. However, even larger systems can in principle be handled easily by our ML approach. To realize the above simulations, we combined neural network potentials (NNPs) of the Behler–Parrinello type[31] with a newly developed ML model for molecular dipole moments. This neural network based model constitutes a new form of a charge partitioning scheme based purely on statistical principles and offers access to environment dependent atomic charges. For the efficient selection of electronic structure data points, a new adaptive sampling scheme is introduced. By employing this scheme, it is possible to incrementally grow ML potentials for specific applications in a highly automated manner based on only a small initial seed of reference data. When combined with the ability of NNPs to include molecular forces in their training procedure, the amount of reference data points required to construct a ML potential is reduced (*e.g.* 717 configurations are sufficient for a converged potential of the tripeptide). Furthermore, we demonstrate the ability of NNPs to model macromolecules based only on the information contained in small fragments, making it possible to treat even these systems with highly accurate electronic structure methods in a divide and conquer fashion. The above findings are not restricted to the simulation of infrared spectra *via* dynamics simulations, but apply to ML potentials in a broader sense. The ML approach presented here thus constitutes an alternative to the currently prevailing trend of fitting potentials to more and more reference data points. The latter strategy suffers from the disadvantage that electronic structure reference calculations become prohibitively expensive for highly accurate methods and/or large molecular systems. Here we show that these problems can be overcome through the efficient use of data, bringing the dream of simulating the dynamics of *e.g.* enzymatic reactions with highly accurate methods one step closer.

Conflicts of interest

There are no conflicts of interest to declare.

Acknowledgements

Allocation of computer time at the Vienna Scientific Cluster (VSC) is gratefully acknowledged. JB is grateful for a DFG Heisenberg Professorship (Be3264/11-1).

References

1 C. M. Bishop, *Pattern Recognition and Machine Learning*, Springer, New York, 1st edn, 2006.
2 I. Goodfellow, Y. Bengio and A. Courville, *Deep Learning*, MIT Press, 2016.
3 K. T. Schütt, F. Arbabzadah, S. Chmiela, K. R. Müller and A. Tkatchenko, *Nat. Commun.*, 2017, **8**, 13890.
4 J. N. Wei, D. Duvenaud and A. Aspuru Guzik, *ACS Cent. Sci.*, 2016, **2**, 725–732.
5 F. A. Faber, L. Hutchison, B. Huang, J. Gilmer, S. S. Schoenholz, G. E. Dahl, O. Vinyals, S. Kearnes, P. F. Riley and O. A. von Lilienfeld, arXiv:1702.05532, 2017.
6 R. Gómez Bombarelli, J. Aguilera Iparraguirre, T. D. Hirzel, D. Duvenaud, D. Maclaurin, M. A. Blood Forsythe, H. S. Chae, M. Einzinger, D.-G. Ha, T. Wu, G. Markopoulos, S. Jeon, H. Kang, H. Miyazaki, M. Numata, S. Kim, W. Huang, S. I. Hong, M. Baldo, R. P. Adams and A. Aspuru Guzik, *Nat. Mater.*, 2016, **15**, 1120–1127.
7 F. Hase, S. Valleau, E. Pyzer-Knapp and A. Aspuru Guzik, *Chem. Sci.*, 2016, **7**, 5139–5147.
8 J. Behler, *Phys. Chem. Chem. Phys.*, 2011, **13**, 17930–17955.
9 C. M. Handley and P. L. A. Popelier, *J. Phys. Chem. A*, 2010, **114**, 3371–3383.
10 J. Behler, *J. Phys.: Condens. Matter*, 2014, **26**, 183001.
11 J. Behler, *J. Chem. Phys.*, 2016, **145**, 170901.
12 J. S. Smith, O. Isayev and A. E. Roitberg, *Chem. Sci.*, 2017, **8**, 3192–3203.
13 V. Botu, R. Batra, J. Chapman and R. Ramprasad, *J. Phys. Chem. C*, 2017, **121**, 511–522.
14 A. P. Bartók and G. Csányi, *Int. J. Quantum Chem.*, 2015, **115**, 1051–1057.
15 Z. Li, J. R. Kermode and A. De Vita, *Phys. Rev. Lett.*, 2015, **114**, 096405.
16 K. Yao, J. E. Herr, S. N. Brown and J. Parkhill, *J. Phys. Chem. Lett.*, 2017, **8**, 2689–2694.
17 D. Marx and J. Hutter, *Ab Initio Molecular Dynamics: Basic Theory and Advanced Methods*, Cambridge University Press, Cambridge, reprint edn, 2012.
18 I. Newton, *Philosophiae naturalis principia mathematica*, J. Societatis Regiae ac Typis J. Streater, 1687.
19 M. Barbatti, *WIREs Comput. Mol. Sci.*, 2011, **1**, 620–633.
20 M.-P. Gaigeot, *Phys. Chem. Chem. Phys.*, 2010, **12**, 3336–3359.
21 M. Thomas, M. Brehm, R. Fligg, P. Vöhringer and B. Kirchner, *Phys. Chem. Chem. Phys.*, 2013, **15**, 6608–6622.
22 S. Mai, P. Marquetand and L. González, *Int. J. Quantum Chem.*, 2015, **115**, 1215–1231.
23 P. Marquetand, J. Nogueira, S. Mai, F. Plasser and L. González, *Molecules*, 2016, **22**, 49.
24 J. Simons, *Mol. Phys.*, 2009, **107**, 2435–2458.
25 M. S. de Vries and P. Hobza, *Annu. Rev. Phys. Chem.*, 2007, **58**, 585–612.
26 T. K. Roy and R. B. Gerber, *Phys. Chem. Chem. Phys.*, 2013, **15**, 9468–9492.
27 H.-D. Meyer, *Wiley Interdiscip. Rev.: Comput. Mol. Sci.*, 2012, **2**, 351–374.
28 P. S. Thomas and T. Carrington Jr, *J. Phys. Chem. A*, 2015, **119**, 13074–13091.
29 X. Andrade, J. N. Sanders and A. Aspuru Guzik, *Proc. Natl. Acad. Sci. U. S. A.*, 2012, **109**, 13928–13933.
30 V. A. Mandelshtam and H. S. Taylor, *J. Chem. Phys.*, 1997, **107**, 6756–6769.
31 J. Behler and M. Parrinello, *Phys. Rev. Lett.*, 2007, **98**, 146401.
32 M. Gastegger and P. Marquetand, *J. Chem. Theory Comput.*, 2015, **11**, 2187–2198.
33 M. Gastegger, C. Kauffmann, J. Behler and P. Marquetand, *J. Chem. Phys.*, 2016, **144**, 194110.
34 S. Grimme, *J. Chem. Phys.*, 2006, **124**, 034108.
35 J. Behler, *Int. J. Quantum Chem.*, 2015, **115**, 1032–1050.
36 J. Behler, *J. Chem. Phys.*, 2011, **134**, 074106.
37 T. B. Blank and S. D. Brown, *J. Chemom.*, 1994, **8**, 391–407.
38 J. Li, B. Jiang and H. Guo, *J. Chem. Phys.*, 2013, **139**, 204103.
39 V. Botu and R. Ramprasad, *Int. J. Quantum Chem.*, 2015, **115**, 1074–1083.
40 S. Manzhos, K. Yamashita and T. Carrington Jr, *Comput. Phys. Commun.*, 2009, **180**, 2002–2012.
41 K. Yao, J. E. Herr and J. Parkhill, *J. Chem. Phys.*, 2017, **146**, 014106.
42 M. Malshe, R. Narulkar, L. M. Raff, M. Hagan, S. Bukkapatnam, P. M. Agrawal and R. Komanduri, *J. Chem. Phys.*, 2009, **130**, 184102.
43 M. G. Darley, C. M. Handley and P. L. A. Popelier, *J. Chem. Theory Comput.*, 2008, **4**, 1435–1448.
44 C. M. Handley and P. L. A. Popelier, *J. Chem. Theory Comput.*, 2009, **5**, 1474–1489.
45 N. Artrith, T. Morawietz and J. Behler, *Phys. Rev. B*, 2011, **83**, 153101.
46 S. Faraji, S. A. Ghasemi, S. Rostami, R. Rasoulkhani, B. Schaefer, S. Goedecker and M. Amsler, *Phys. Rev. B*, 2017, **95**, 104105.
47 T. Morawietz, V. Sharma and J. Behler, *J. Chem. Phys.*, 2012, **136**, 064103.
48 C. M. Breneman and K. B. Wiberg, *J. Comput. Chem.*, 1990, **11**, 361–373.
49 S. M. Bachrach, in *Population Analysis and Electron Densities from Quantum Mechanics*, John Wiley & Sons, Inc., 2007, pp. 171–228.
50 K. B. Wiberg and P. R. Rablen, *J. Comput. Chem.*, 1993, **14**, 1504–1518.
51 A. D. Mackerell, *J. Comput. Chem.*, 2004, **25**, 1584–1604.
52 C. M. Baker, *Wiley Interdiscip. Rev.: Comput. Mol. Sci.*, 2015, **5**, 241–254.
53 F. Neese, *Wiley Interdiscip. Rev.: Comput. Mol. Sci.*, 2012, **2**, 73–78.
54 A. D. Becke, *Phys. Rev. A*, 1988, **38**, 3098–3100.
55 P. A. M. Dirac, *Proc. R. Soc. London, Ser. A*, 1929, **123**, 714–733.
56 J. P. Perdew, *Phys. Rev. B*, 1986, **33**, 8822–8824.
57 S. H. Vosko, L. Wilk and M. Nusair, *Can. J. Phys.*, 1980, **58**, 1200–1211.
58 J. C. Slater, *Phys. Rev.*, 1951, **81**, 385–390.
59 F. Weigend and R. Ahlrichs, *Phys. Chem. Chem. Phys.*, 2005, **7**, 3297–3305.

60 C. Lee, W. Yang and R. G. Parr, *Phys. Rev. B*, 1988, **37**, 785–789.
61 K. Eichkorn, O. Treutler, H. Öhm, M. Häser and R. Ahlrichs, *Chem. Phys. Lett.*, 1995, **240**, 283–290.
62 O. Vahtras, J. Almlöf and M. W. Feyereisen, *Chem. Phys. Lett.*, 1993, **213**, 514–518.
63 J. Behler, *RUNNER – A program for constructing high-dimensional neural network potentials*, Universität Göttingen, 2017.
64 *Python Reference Manual*, ed. G. van Rossum and F. L. Drake, PythonLabs, Virginia, USA, 2001, http://www.python.org, access date 06.04.2017.
65 S. van der Walt, S. C. Colbert and G. Varoquaux, *Comput. Sci. Eng.*, 2011, **13**, 22–30.
66 J. Bergstra, O. Breuleux, F. Bastien, P. Lamblin, R. Pascanu, G. Desjardins, J. Turian, D. Warde Farley and Y. Bengio, *Proceedings of the Python for Scientific Computing Conference*, SciPy, 2010.
67 S. Chmiela, A. Tkatchenko, H. E. Sauceda, I. Poltavsky, K. T. Schütt and K.-R. Müller, *Sci. Adv.*, 2017, **3**, e1603015.
68 P. Chu, F. Guenther, G. Rhoderick and W. Lafferty, in *NIST Chemistry WebBook NIST Standard Reference Database Number 69*, ed. P. Linstrom and W. Mallard, National Institute of Standards and Technology, Gaithersburg MD, p. 20899, retrieved April 24, 2017, DOI: 10.18434/t4d303.
69 S. Grimme, J. Antony, S. Ehrlich and H. Krieg, *J. Chem. Phys.*, 2010, **132**, 154104.
70 S. Grimme, S. Ehrlich and L. Goerigk, *J. Comput. Chem.*, 2011, **32**, 1456–1465.
71 S. A. Fischer, T. W. Ueltschi, P. Z. El-Khoury, A. L. Mifflin, W. P. Hess, H.-F. Wang, C. J. Cramer and N. Govind, *J. Phys. Chem. B*, 2016, **120**, 1429–1436.
72 A. Cimas, T. D. Vaden, T. S. J. A. de Boer, L. C. Snoek and M.-P. Gaigeot, *J. Chem. Theory Comput.*, 2009, **5**, 1068–1078.
73 T. D. Vaden, T. S. J. A. de Boer, J. P. Simons, L. C. Snoek, S. Suhai and B. Paizs, *J. Phys. Chem. A*, 2008, **112**, 4608–4616.