The Journal of Chemical Physics

RESEARCH ARTICLE | MAY 17 2023

NMR spectrum prediction for dynamic molecules by machine learning: A case study of trefoil knot molecule

Special Collection: Machine Learning Hits Molecular Simulations

Mikhail Tsitsvero ; Jenny Pirillo ; Yuh Hijikata ; Tamiki Komatsuzaki

J. Chem. Phys. 158, 194108 (2023)
https://doi.org/10.1063/5.0147398

The Journal of Chemical Physics ARTICLE scitation.org/journal/jcp

NMR spectrum prediction for dynamic molecules by machine learning: A case study of trefoil knot molecule

Cite as: J. Chem. Phys. 158, 194108 (2023); doi: 10.1063/5.0147398
Submitted: 22 February 2023 • Accepted: 3 May 2023 •
Published Online: 17 May 2023

Mikhail Tsitsvero,[a] Jenny Pirillo, Yuh Hijikata, and Tamiki Komatsuzaki

AFFILIATIONS

Institute for Chemical Reaction Design and Discovery (WPI-ICReDD), Hokkaido University, Sapporo, Japan

Note: This paper is part of the JCP Special Topic on Machine Learning Hits Molecular Simulations.
[a]**Author to whom correspondence should be addressed:** tsitsvero@icredd.hokudai.ac.jp

ABSTRACT

Nuclear magnetic resonance (NMR) spectroscopy is one of the indispensable techniques in chemistry because it enables us to obtain accurate information on the chemical, electronic, and dynamic properties of molecules. Computational simulation of the NMR spectra requires time-consuming density functional theory (DFT) calculations for an ensemble of molecular conformations. For large flexible molecules, it is considered too high-cost since it requires time-averaging of the instantaneous chemical shifts of each nuclear spin across the conformational space of molecules for NMR timescales. Here, we present a Gaussian process/deep kernel learning-based machine learning (ML) method for enabling us to predict, average in time, and analyze the instantaneous chemical shifts of conformations in the molecular dynamics trajectory. We demonstrate the use of the method by computing the averaged ^{1}H and ^{13}C chemical shifts of each nuclear spin of a trefoil knot molecule consisting of 24 para-connected benzene rings (240 atoms). By training ML model with the chemical shift data obtained from DFT calculations, we predicted chemical shifts for each conformation during dynamics. We were able to observe the merging of the time-averaged chemical shifts of each nuclear spin in a singlet ^{1}H NMR peak and two ^{13}C NMR peaks for the knot molecule, in agreement with experimental measurements. The unique feature of the presented method is the use of the learned low-dimensional deep kernel representation of local spin environments for comparing and analyzing the local chemical environment histories of spins during dynamics. It allowed us to identify two groups of protons in the knot molecule, which implies that the observed singlet ^{1}H NMR peak could be composed of the contributions from protons with two distinct local chemical environments.

I. INTRODUCTION

Within the last decade, there has been tremendous progress in the development of machine learning (ML) methods for fast and accurate prediction of chemical properties of molecules and materials allowing to access molecular scales and simulating time scales that were intractable before. These methods realized accurate models for potential energy surfaces,[1–8] as well as chemical shifts in solids[9–12] and liquid state.[13] One of the current challenges is to enable access to the fast and reliable ML-aided computational chemistry predictions of thermodynamic[14] and time-averaged chemical properties for previously intractable large systems. One computationally demanding task, in particular, is the prediction of nuclear magnetic resonance (NMR) spectra of large and dynamic molecules that requires computing the time-averaged chemical shifts of nuclear spins.

NMR spectroscopy quantifies the response of the atomic nuclei spins placed in a strong magnetic field to the radio-frequency electromagnetic radiation.[15] Due to the nuclear magnetic resonance, spins reemit signals at the shifted frequencies depending on the local environment of each spin, making it possible to capture the relative positions of neighboring nuclear spins as well as local electronic structure. Therefore, NMR measurements are powerful and indispensable, in particular in chemistry and material science. Furthermore, most importantly, the resulting NMR spectrum reflects the time-averaged local atomic environments due to the intrinsic dynamics of the molecule. Access to the local structural, electronic, and dynamic information makes NMR spectroscopy one of the most versatile and precise spectroscopic techniques in the analysis of not only chemical structure but also dynamic behaviors of molecules.

The computation of NMR spectra of molecules has a long history and now is well established.[15] On the other hand, for some molecules, such as the trefoil knot molecule of paraphenylene reported recently,[16] the computation of the NMR spectrum for a representative structure at minimum energy cannot reproduce the single peak of experimentally observed ^{1}H NMR spectrum, even at a qualitative level, presumably because of the intrinsic dynamics of the knot. Interpretation of the experimental NMR spectra can be supported by simulated spectra, which should be obtained by time-averaging the chemical shifts for each nuclear spin over the space of accessible molecular conformations. It is, however, computationally challenging. First, the computation of the chemical shifts for a single molecular conformation is high-cost since it requires the calculation of the second-order derivative of energy of the molecule with respect to the external magnetic field and nuclear magnetic momentum; this is a very time-consuming calculation with currently available quantum chemistry computer software. Second, the chemical shifts should be computed for all instantaneous conformations of the dynamic knot molecule. The time-averaging of the simulated chemical shifts over the entire accessible conformational space poses a computational challenge, while the time-averaged chemical shifts have been directly computed only for a limited number of relatively small molecules.[17]

Several computational approaches exist for the prediction of NMR spectra of dynamic molecules. For example, in the work of Kwan and Liu,[17] the dynamic contributions to predicting the NMR spectra of small molecules were studied and their method was applied to predict the equilibrium structure of the [18]-annulene molecule, consisting of 36 atoms. Next, the Conformer-Rotamer Ensemble Sampling Tool[18] aims to predict the averaged NMR spectra of small molecules by combining the computed spectra for a Boltzmann weighted set of equilibrium structures, where the nuclear permutations due to the rotatory motions of parts of the molecule are addressed. Even though these approaches demonstrated the necessity to account for dynamic effects, as pointed out in the work of Grimme *et al.*,[18] in most cases, it is computationally too expensive to calculate the averaged chemical shifts for each nuclear spin since it requires the computation of chemical shifts for a large number of conformations sampled from the molecular dynamics (MD) trajectory as we also mentioned above. On the other hand, recent ML-based approaches for NMR prediction[9–13,19] allowed to significantly reduce the computational cost by employing a surrogate ML model for chemical shifts trained on a set of reference configurations, therefore, significantly reducing the amount of high-cost quantum chemical calculations. Although time-averaged chemical shifts can be computed with simple ML models, the richer multiparametric probabilistic models could offer a convenient set of tools to obtain finer insights into the characterization of chemical environments of individual spins during dynamics.

In this study, we develop a new algorithmic scheme for the elucidation of NMR spectra of flexible molecules by using a classical MD simulation and Gaussian process (GP)/deep kernel learning (DKL) model in conjunction with continuous three-dimensional Smooth Overlap of Atomic Positions (SOAP) descriptors, which is capable of predicting the time-averaged chemical shifts of each nuclear spin as well as comparing the local chemical environments of individual spins over the accessible conformational space, see Fig. 1. Differently from other chemical shift prediction models,[9,10,13] the presented method leverages the learned low-dimensional deep kernel representation of local environments allowing to compare the histories of local chemical environments of spins during dynamics as well as visually and numerically control the uncertainty of predictions. As an illustrative example, we employed a trefoil knot molecule,[16] which is topologically interlocked and consists of 24 linked benzene rings (144 carbons and 96 protons). Although local chemical environments of atoms within a single conformation of the knot are quite diverse, its experimental ^{1}H NMR and ^{13}C NMR spectra provided just a singlet peak and two peaks, respectively.[16] It was suggested that a few peaks observed in the ^{1}H and ^{13}C NMR spectra are due to the intrinsic dynamics of the knot.

II. RESULTS

While making predictions of chemical shifts for individual spins by an ML method, it would be beneficial to have (1) a representation of local nuclear environments that is continuous with respect to atomic positions; (2) good prediction quality on the test set together with uncertainty estimates for each test sample; (3) a low-dimensional representation of dataset allowing for data interpretation and uncertainty control. To fit all of these criteria, we use a combination of GP regression with DKL[20,21] together with SOAP continuous descriptors in a three-dimensional molecular space.

The overall prediction scheme is given in Fig. 1, and we refer to it as SOAP/DKL/GP model. In the model, first, training/test conformations were selected from the MD trajectory data obtained from semiempirical and classical MD simulations. Then, density functional theory (DFT) calculations of chemical shifts were performed for the selected molecular conformations. After the preparation of the training data, the SOAP/DKL/GP model was trained with the chemical shift data. Next, model performance was examined in terms of the predictive quality of chemical shifts and uncertainty estimates; a decision was taken on whether more training data are needed. Uncertainty quantification for the model was performed by the learned variance parameter of the GP acting over the low-dimensional output of the deep kernel. The attractive feature of the low-dimensional DKL representation is that we can visually and numerically examine how well the environments of the selected training conformations cover the environments present in the MD trajectory. Finally, the low-dimensional mappings of local nuclear chemical environment histories, provided by the trained deep kernel, were computed together with the time-averaged chemical shifts of each nuclear spin.

The prediction scheme allowed us not only to provide a possible explanation of the formation of the experimental ^{1}H and ^{13}C NMR spectra but also to get theoretical insight into the local chemical environment histories of nuclear spins during the dynamics of the molecule. By comparing the local chemical environment distributions of each ^{1}H nuclear spin, we categorized protons into two groups that correspond to distinct distributions of the local chemical environments, although resulting in almost indistinguishable time-averaged chemical shifts based on classical MD simulation for a 100 ns that is even shorter than the timescale of NMR measurement.

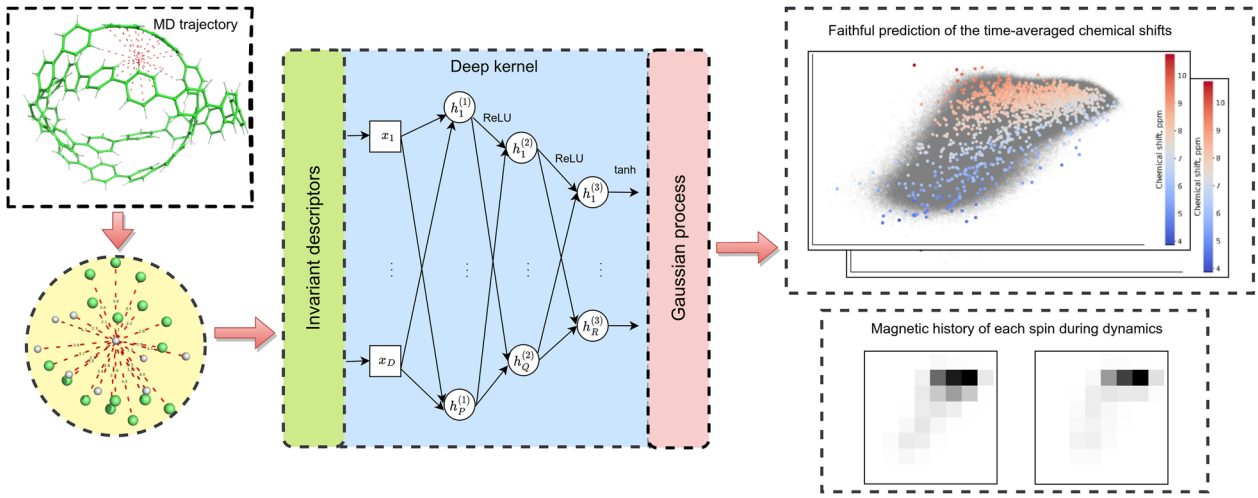

FIG. 1. Machine learning (ML) scheme for predicting time-averaged chemical shifts for each nuclear spin. First, local environments of nuclear spins, based on MD trajectory of the knot (left), are mapped to the invariant vectors. Then, invariant vectors are mapped to the low-dimensional space by the deep kernel (D = 840, P = 1000, Q = 50, R = 2 were used), where predictions are performed by the Gaussian process (middle). Low-dimensional mapping at the output of the deep kernel is used to construct fingerprints characterizing the local chemical environments of the nuclear spins during dynamics (right).

A. Results: Trefoil knot dynamics and NMR spectrum

During the dynamics, each nuclear spin of the knot moves through diverse local chemical environments. This is easily inferred from the calculated chemical shifts for one of the optimized conformations at the DFT level of theory, see Fig. 2, where the calculated chemical shifts cover a wide range of values,[22] while the experimentally observed spectra are given by the narrow singlet peaks in both ^{1}H and ^{13}C NMR spectra. As a consequence, it was suggested that the merging of signals from different nuclear spins of the same species into a narrow peak may appear due to the dynamics of nuclear spins within the flexible molecule.

Guided by the classical MD simulations, we illustrate the dynamics of the knot in Fig. 3. Locally, the paraphenylene chain is slipping next to another chain at the three "crossings" within the knot. Because the molecule is a closed loop, the local traversal sliding induces the sliding of another part of the molecule. This dynamics of the molecule as a whole causes interchange of the nuclear relative positions. This process occurs at timescales of hundreds of ps, leading to the appearance of narrow peaks. Estimated correlation times for different groups of spins based on MD simulations are provided in the supplementary material, Figs. 7 and 8. Since all the atoms within the knot molecule fluctuate with the correlation times in the range of hundreds of ps, it allows us to predict the locations of sharp peaks for individual spins by computing the time averages of their chemical shifts. In this case, no line broadening effect appears, and the final spectrum results in a collection of narrow peaks.

Symmetries and nuclear permutations arising during the dynamics can be understood by separating nuclear spins into groups. We can partition the whole system into two distinct groups of carbons (i.e., carbons at para-positions and nonpara-positions) and a group of protons with two subgroups. First, there are para-carbons that link phenyl rings to each other and are bonded to three carbons. Second, there are ortho- and meta-carbons that are bonded to two carbons. There are in total 48 para-carbons, and 96 nonpara-carbons, which we can relate to differences in intensities

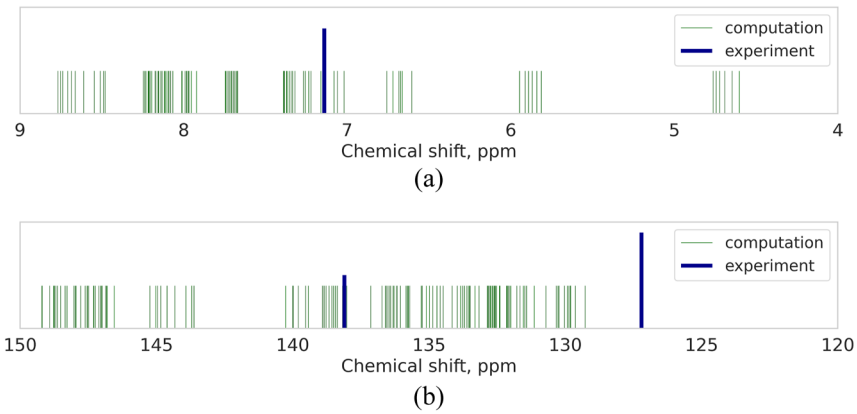

FIG. 2. Experimental NMR spectra read from the reported paper[16] (dark blue color) and calculated chemical shifts for one of the optimized conformations (green color). Relative intensities of the peaks are shown only for clarity. (a) ^{1}H nuclear spins, (b) ^{13}C nuclear spins.

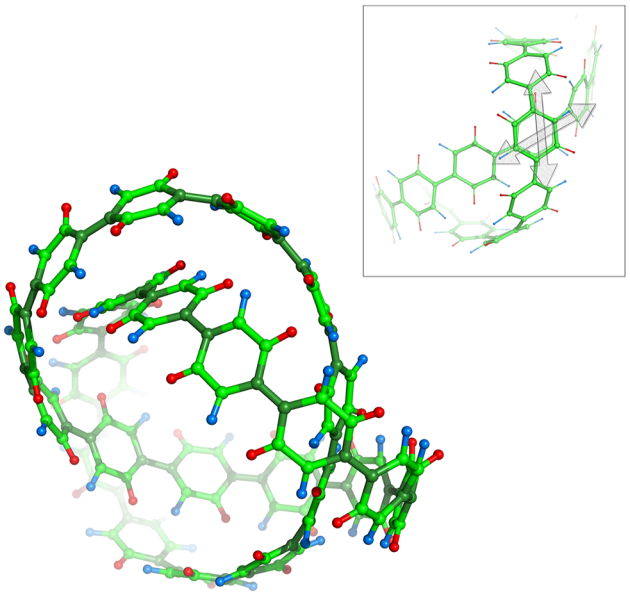

FIG. 3. Trefoil knot molecule. Para-carbons are colored in dark green, nonpara-carbons are colored in light green, and hydrogen atoms, highlighted with red and blue, correspond to two subgroups of protons with distinct chemical environment distributions during dynamics. Typical motions of the knot molecule are shown in the inset with arrows.

of peaks in the ^{13}C NMR spectrum. As it will be shown, despite there being a single observed peak in the experimental ^{1}H NMR spectrum, we found there are two distinct subgroups of protons with different distributions of local chemical environments during the simulated time length of MD. This partition of protons into two subgroups during dynamics appears at timescales of 1 ns and is present for the 100 ns MD simulation. As illustrated in Fig. 3, the protons in the outer (or upper) chain that cross another chain (inner or lower) in a parallel way are highlighted in blue and the protons in the inner (or lower) chain that cross another chain (upper or outer) in a parallel way are highlighted in red. Similar to the protons, the nonpara-carbons also can be categorized into two subgroups correspondingly.

B. Results: Merging of the time-averaged chemical shifts

Next, we show how the time-averaged ^{1}H and ^{13}C NMR chemical shifts become closer and group together when the length of time-averaging window is increased. This is expected to result in sharp peaks observed in the experiment. Figure 4 illustrates the merging of the time-averaged ^{1}H and ^{13}C NMR chemical shifts for each proton and carbon at three simulation times for 1 ns (200 snapshots), 10 ns (2000 snapshots), and 100 ns (20 000 snapshots) of the classical MD. Additionally, in Fig. 6 we report the variance of the time-averaged chemical shifts corresponding to each peak as a function of the MD simulation time: Consequently, for given range of simulation times up to 100 ns, variances of the time-averaged chemical shifts approximately follow the power law as a function of simulation time.

The experimentally observed NMR chemical shifts were 7.14 ppm for ^{1}H and 127.2 and 138.1 ppm for ^{13}C, while the predicted ones were 7.7 ppm for ^{1}H and 134.6 and 139.7 ppm for ^{13}C. This discrepancy between the observed and predicted chemical shifts may be attributed to the following factors: (1) lack of accuracy of the classical force fields to describe the dynamics of the knot appropriately; (2) computational condition for chemical shifts, see

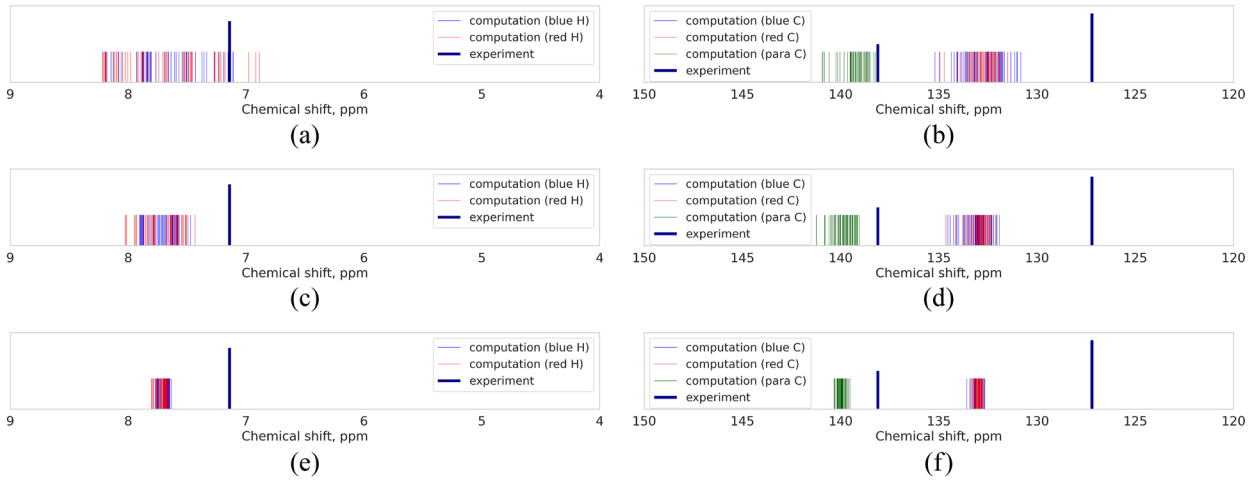

FIG. 4. Merging of the averaged ^{1}H and ^{13}C NMR chemical shifts at different timescales. The computed time-averaged chemical shifts for individual nuclear spins are shown as bars on the rug plot: in blue color for "blue" protons and nonpara-carbons bonded to such protons (denoted "blue C"), in red color, for "red" protons and nonpara-carbons bonded to such protons (denoted "red C"), in green color for para-carbons, together with experimental NMR spectra read from the paper[16] (dark blue color). (a) ^{1}H chemical shifts averaged over 1 ns, (b) ^{13}C chemical shifts averaged over 1 ns, (c) ^{1}H chemical shifts averaged over 10 ns, (d) ^{13}C chemical shifts averaged over 10 ns, (e) ^{1}H chemical shifts averaged over 100 ns, (f) ^{13}C chemical shifts averaged over 100 ns.

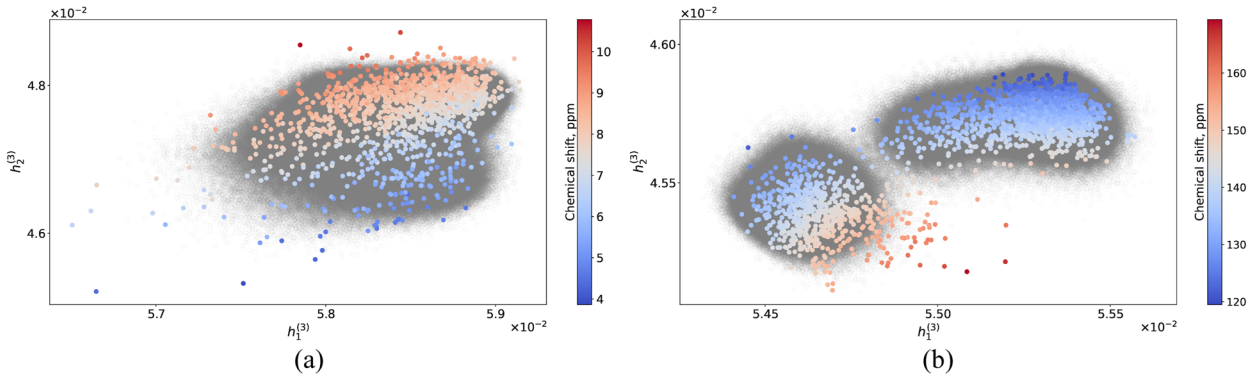

FIG. 5. Mappings of ^{1}H (a) and ^{13}C (b) local environments from classical MD trajectory by the deep kernel. $h_1^{(3)}$ and $h_2^{(3)}$ are the outputs at the last (third) layer of the deep kernel. Colored points correspond to the environments from the training instantaneous conformations with computed chemical shifts; gray circles correspond to the environments from all conformations within the trajectory for which SOAP/DKL/GP prediction scheme was employed. (a) ^{1}H spins (b) ^{13}C spins.

Ref. 17; (3) missing of solvent effects in DFT calculation of chemical shifts; (4) averaging chemical shifts within shorter time scales compared with the time resolution of NMR measurements; and (5) bias error of the ML model. Although the discrepancy originated from various factors, the proposed prediction scheme allowed us to elucidate the merging of the time-averaged chemical shifts, which could explain the experimentally observed NMR spectra of the knot molecule.

Even though the merging of the time-averaged chemical shifts showed agreement with the experimental measurements, a simple question arises: How much faith we should put into ML predictions for a given trajectory and given training data? The low-dimensional mapping provided by the deep kernel may give us a good hint. Parameters of the deep kernel are trained to statistically fit the set of training data and the prediction uncertainty is captured by the base kernel acting on the two-dimensional output of the deep kernel. Additionally, it is valuable to have a visual representation that will indicate where the predictions may become less credible.

In Fig. 5, we provide the mappings of the training local atomic environments on top of all the local atomic environments that appear within the MD trajectory. In a nutshell, the prediction of chemical shift for each unknown environment (a gray circle) is made by interpolating among the set of its colored neighbors. The farther a gray circle from the nearest colored training points is the higher the uncertainty of the prediction becomes. Deep kernel mapping of the instantaneous local environments of individual protons reveals the structure of the types of environments present in the system. There are *seemingly* a single group of environments for protons and two groups for carbons. Overall, the proposed scheme can be used to predict chemical shifts, understand the histories of the local chemical environments of nuclear spins, compute the time-averaged chemical shifts, and estimate the uncertainty of predictions. The local chemical environments of proton and carbon nuclear spins from the MD are mapped to the continuous blobs of gray points covered by the training points (in color), where no gray points stay outcast, thus providing us with a clue that most of the conformational space of the molecule during classical MD is covered by the environments present in the training dataset. Section II C discusses the groupings of the local chemical environment histories of protons in detail.

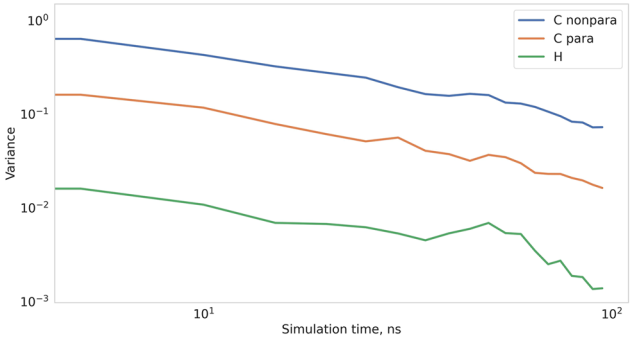

FIG. 6. Variance of the time-averaged chemical shifts for nuclear spins corresponding to each peak as a function of MD simulation time for which time-averaging was performed.

C. Results: Analysis of local chemical environment histories of protons

Since only a single peak was observed in ^{1}H NMR spectrum, we may need to investigate that all the protons move through identical local chemical environments at the accessible MD timescale. How can we quantify the histories of local chemical environments each proton experiences during dynamics? To answer this question, we performed the following analysis.

Each spin continuously moves through various local chemical environments during dynamics. This motion is reflected by the continuous path each nuclear spin tracks over the two-dimensional output of the deep kernel. Therefore, each proton has the corresponding distribution of points over the low-dimensional deep kernel output, reflecting the series of the local environments the proton experienced throughout the dynamics. These distributions can

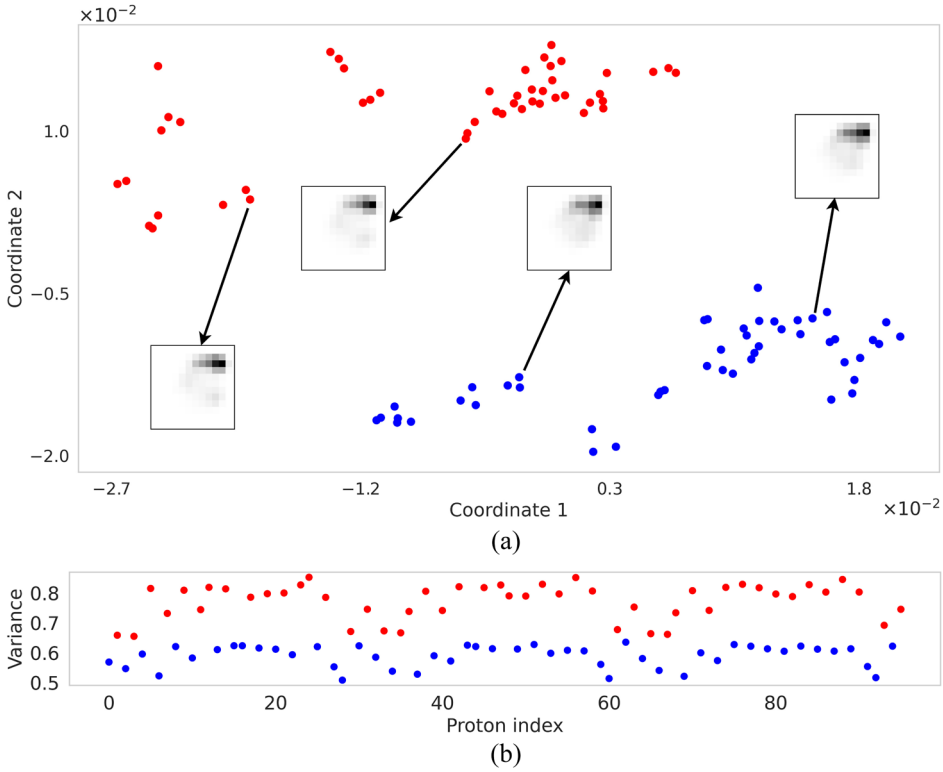

FIG. 7. (a) Similarity between the distributions of proton chemical environments during dynamics at the output of the deep kernel visualized by the multidimensional scaling algorithm. Each point corresponds to a distribution of local environments at the output of the deep kernel experienced by a single proton during the dynamic motion of the knot. Distance between a pair of points corresponds to the earth mover's distance between the corresponding discretized distributions. Example distributions are highlighted within squares. Local chemical environment histories of protons form two distinct groups highlighted by red and blue colors. (b) Variances of the predicted chemical shifts experienced by protons during dynamics.

thus serve to characterize the overall local chemical environment histories of individual protons along the MD trajectory.

First, we computed the distributions of sets of points over the two-dimensional output of the deep kernel by binning points into 12×12 discrete grids. Second, to compare the distributions corresponding to the individual protons, we computed the earth mover's distance[24] between them, which resulted in a 96×96 distance matrix for 96 protons. Finally, using this distance matrix, we applied a multidimensional scaling algorithm to visualize the similarity between the protons' chemical environment distributions as shown in Fig. 7(a). We observe two distinct groups of protons, defined as red and blue protons in Fig. 3, which implies that the observed singlet ^{1}H NMR peak could be composed of the contributions from protons with two distinct local chemical environments. Indeed, we obtain a very similar value of the time-averaged chemical shifts averaged among all red and all blue protons, respectively, as reported in Table S1 in the supplementary material. In addition, it was noted that red protons experience a larger variance of chemical shifts during dynamics than blue ones, as shown in Fig. 7(b). This is because red and blue protons are affected differently via CH/π type interactions by the neighboring chain. As a note of caution, this grouping of protons may depend on the simulated MD timescale. Here, we used a 100 ns classical MD trajectory, where no flip of a phenyl ring was observed (discussion on the ring flip barrier is available in the supplementary material, Sec. S2 G). In case there are such ring-flips for longer simulation times, all protons would be indistinguishable, which is also in agreement with experimental data. Similar to the proton case, the nonpara-carbons can also be clustered into two subgroups during the dynamics (supplementary material, Fig. S14).

III. CONCLUSION

In this work, we have presented an ML-based method for predicting chemical shifts, computing the time-averaged chemical shifts for each nuclear spin, analyzing the local chemical environment histories of spins during dynamics, and applied it to the example of a large knot molecule. We were able to obtain averaged chemical shifts for such a large molecule and found that the observed singlet ^{1}H NMR peak could be composed of the contributions from protons with two distinct local chemical environments. The distinctive feature of the presented method is the use of the learned low-dimensional deep kernel representation of local environments allowing to compare the histories of local chemical environments of spins during dynamics and to evaluate the prediction uncertainty both numerically and visually over the output layer of the deep kernel.

From the perspective of machine learning, it might be surprising that we successfully predicted the chemical shifts for several thousand snapshots using calculated chemical shifts for only 18 conformations as a training dataset. This is due to the characteristics of the knot structure: First, there is similarity between the local environments of different spins; second, the dynamics of the interlocked knot molecule is somewhat restricted due to the knot structure.

These facts enabled us to capture the repetitive atomic environments with a small number of training conformations required for accurate predictions. We selected the knot molecule to prove that the SOAP/DKL/GP scheme can provide the time-averaged chemical shifts for each nuclear spin; the prediction of the time-averaged chemical shifts for large molecules with more complex dynamics will indeed motivate the design of more advanced multiparametric probabilistic predictive schemes that scale easily with the addition of a large number of training points.

Some additional points in methodological development are the following. First, the computational cost of the chemical shifts for many conformations of a large system at a sufficient level of theory can be prohibitively high. Since chemical shifts represent a local chemical property, we may need to design approximate calculation schemes together with transfer learning methods for training the model on one collection of systems while applying it to the other. Second, performing MD simulations at sufficient time lengths comparable to the acquisition time of NMR measurements is also high-cost. Third, fitting chemical shifts for a large number of local atomic neighborhoods may be problematic due to scalability issues of GPs; this point currently is an active research topic on its own. These issues are under consideration for the following work.

IV. METHODS

A. Computational studies

1. Classical MD

First, we performed quantum mechanical calculations at HF/6-31G* level of theory by Gaussian 16 Rev. C.01[26] software to evaluate atomic charges. The topology files of the trefoil knot molecule and solvent were prepared by the antechamber package[27] with the general AMBER force field.[28] AMBER topology files generated by the package were converted into a GROMACS format using the Python script, Acpype.[29]

Second, equilibration of the solvent under the presence of the knot molecule was conducted by the following procedure. One knot molecule was placed at the center of 8 Å cubic box and solvated by 7669 dichloromethane molecules. Energy minimization of the systems was performed based on the conjugate gradient algorithm with the threshold of the maximum force, 10.0 kJ/mol/nm. Next, canonical ensemble (NVT ensemble) equilibration for 0.1 ns at 300 K with 0.2 fs timestep followed by isothermal–isobaric ensemble (NPT ensemble) equilibration for 0.1 ns at 1 bar pressure with 0.2 fs timestep was carried out for relaxation of solvent molecules, while the atomic positions of the knot are restrained with a harmonic potential. The temperature and the pressure were kept constant by using the velocity rescale thermostat and the Berendsen barostat,[30] respectively.

Third, the main MD simulation was performed under NPT ensemble for 100 ns. The temperature at 300 K and pressure at 1 bar were kept constant using the velocity rescale thermostat[31] and the Parrinello–Rahman barostat,[32] respectively. The cutoff value of 14 Å was assigned for nonbonded interactions. Long-range electrostatic interactions were treated using a particle-mesh Ewald scheme.[33] Dynamics were propagated with a leapfrog integrator using a time step of 1 fs, and C–H bonds were constrained using a linear constraint solver.[34] All simulations were carried out under periodic boundary conditions using GROMACS software version 2020.5.[35] The structures at 10, 30, 50, 70, 90, 95, and 100 ns during the main MD simulation were used to calculate the chemical shifts.

2. DFTB-MD

Optimization of trefoil knot was performed using the self-consistent-charge density functional tight-binding (DFTB) method[36] with the third-order expansion[37] as implemented in the DFTB+ package version 20.2.1.[38] The 3ob parameter set[39,40] with the Hubbard parameters −0.1492 for C and −0.1857 for H were employed. After the optimization, DFTB with molecular dynamics (DFTB-MD) simulations were conducted for 200 ps with 0.4 fs time interval, employing an NVT ensemble at 300 K and the Nose–Hoover chain thermostat.[41,42] Grimme's D3 type dispersion[43] was included in all calculations. The structures at 0, 20, 25, 50, 75, 100, 125, 150, 175, and 200 ps were used to calculate the chemical shifts.

3. NMR spectra simulation

Isotropic chemical shieldings were calculated at B3LYP/6-311+G(2d,p) level of theory with GIAO method by the Gaussian 16 Rev. C.01.[26] Chemical shieldings were converted to chemical shifts by employing the $SiMe_4$ as a reference (values are available in the supplementary material). Chemical shifts were calculated for the selected snapshots from the classical MD, DFTB-MD, one optimized structure at DFTB level of theory, and one optimized structure at the B3LYP/6-311+G(2d, p) level of theory. The calculation for a single snapshot required around three days on 16 cores.

B. Invariant descriptors and ML model

For a continuous and rotationally invariant representation of local nuclear environments, we used the SOAP descriptors.[44] Here, we briefly describe the basic steps behind SOAP calculations for the simplest case of the three-body density correlation function. SOAP first maps the atomic positions of a molecule within the local neighborhood of some central atom onto the sum of smooth and localized functions, typically Gaussians, which is referred to as atomic density. The density is then projected onto the basis sets of radial functions and spherical harmonics. The types of radial basis functions and size of radial/spherical basis sets are hyperparameters and are selected depending on the diversity of molecular conformations within the dataset. The feature vector that continuously depends on the positions of neighbor atoms, and which is rotation, translation, and permutation invariant, is obtained by contraction of density expansion coefficients over magnetic quantum numbers. Details on SOAP calculations are provided in the supplementary material and the complete theory can be found, for example, in Ref. 45.

The GPs are known to be efficient and mathematically well-formalized regression models that provide predictions jointly with uncertainty estimates.[46] The GP is fully defined by its mean function and kernel or covariance function. While the mean function simply encodes the prior knowledge into the model, the covariance function captures the correlations between the pairs of points within the dataset. The representation of local nuclear environments provided by SOAP vectors is high-dimensional. For example, 840-dimensional vectors were computed in this work. While training GPs in high-dimensional space for simple datasets does not

necessarily pose a problem, we gain additional information if we obtain a low-dimensional representation of data. Such representation of data points may naturally be learned by the DKL,[20,21] where the GP covariance function is given by the composition of the deep neural network with the low-dimensional base kernel: The low-dimensional output of the neural network serves as an input for the base kernel. Weights of the neural network, referred to as a deep kernel, are trained jointly with the base kernel parameters to maximize the marginal log-likelihood, which provides a natural measure of how well the statistical model describes the training data. In the end, the trained deep kernel provides a mapping from the high-dimensional descriptor space to the low-dimensional space, where similar environments having similar chemical shifts are mapped close by.

The proposed SOAP/DKL/GP scheme applied to the dynamic knot molecule allowed us (1) to compute the time-averaged chemical shifts; (2) to compare the histories of local chemical environments of nuclear spins during dynamics; (3) to get visual and numerical estimates where predictions may fail for a given training set of conformations and a given MD trajectory. We used the cross-validation technique to assess the performance of the model and then tested the model on the unseen test snapshot randomly selected from the available snapshots from the classical MD trajectory. During the tests of the ^{1}H SOAP/DKL/GP model, the Mean Square Error (MSE) on the unseen test snapshot was 0.066 ppm, the Mean Absolute Error (MAE) was 0.202 ppm, and the maximum error was 0.658 ppm. The ^{13}C SOAP/DKL/GP model had an MSE of 2.402 ppm, an MAE of 1.240 ppm, and a maximum error 5.21 ppm on the unseen test snapshot. Additional details of the model performance are provided in the supplementary material.

From a physical perspective, the chemical shift prediction scheme is based on two main assumptions. First, it is assumed that the chemical shielding tensor varies continuously with the positions of the neighboring atoms, i.e., we can apply the local atom density fitting scheme. Second, it is assumed that the entire conformational space of the molecule was accessible by MD simulations. In cases when high free energy barriers exist between basins of conformations, we may need to apply additional Boltzmann weighted averaging for the set of trajectories from each basin.

Previously, the continuous SOAP descriptors[44] were used together with simple GP models for energy and force-fitting,[1] as well as for the prediction of NMR spectra of solid state structures.[9] Differently from previous studies that applied GPs to fit the chemical properties of molecules, here, we employ a highly multiparametric probabilistic model that combines training of the deep kernel jointly with GP. We use the nonlinear deep kernel map as a way to visualize and compare distributions of local environments of nuclear spins during the dynamics of a molecule.

We used chemical shifts computed for 18 configurations (seven configurations from classical MD, nine configurations from DFTB-MD, one optimized structure at DFTB level of theory, and one optimized structure at B3LYP/6-311+G(2d, p) level of theory without including neither implicit nor explicit solvent) as a dataset. We trained the ML model on 17 configurations and used one configuration (taken from classical MD) as test dataset. SOAP parameters, the architecture of the deep kernel, and analysis of the model performance in terms of log-likelihood loss and root-mean-square error are given in the supplementary material. We used the library[47] for computing the SOAP representations and the library[48] for GP model training.

C. Comparing chemical environment distributions

To compare the chemical environment distributions mapped by the deep kernel, first we computed the discretized distributions of sets of points for individual spins over the two-dimensional output of the deep kernel by binning points into 12×12 discrete grid. Then, we computed the earth mover's distance[49] between pairs of distributions with POT library,[24] which resulted in a square distance matrix. Earth mover's distance quantifies the best way to move the mass between two distributions. Next, we applied the multidimensional scaling (MDS) algorithm to visualize the similarity between the chemical environment distributions of spins. MDS is a nonlinear dimensionality reduction algorithm that aims to map data points with a given distance matrix to a low-dimensional Cartesian space by attempting to preserve the distances between the data points.[50] We used MDS algorithm implemented in the scikit-learn library.[51]

SUPPLEMENTARY MATERIAL

The Python code for training the machine learning model, the dataset including the chemical shifts training data and the classical MD trajectory, and additional figures and details of the model training are available in the supplementary material.

ACKNOWLEDGMENTS

This work was supported by World Premier International Research Center Initiative (WPI), MEXT, Japan.

AUTHOR DECLARATIONS

Conflict of Interest

The authors have no conflicts to disclose.

Author Contributions

Mikhail Tsitsvero: Conceptualization (lead); Data curation (equal); Formal analysis (equal); Investigation (equal); Methodology (equal); Project administration (equal); Software (equal); Supervision (equal); Validation (equal); Visualization (equal); Writing – original draft (equal); Writing – review & editing (equal). **Jenny Pirillo**: Conceptualization (supporting); Data curation (equal); Investigation (equal); Resources (equal); Validation (equal); Writing – review & editing (equal). **Yuh Hijikata**: Conceptualization (supporting); Data curation (equal); Investigation (equal); Resources (equal); Validation (equal); Writing – review & editing (equal). **Tamiki Komatsuzaki**: Conceptualization (supporting); Investigation (equal); Validation (equal); Writing – review & editing (equal).

DATA AVAILABILITY

The data that support the findings of this study are available within the article and its supplementary material.

REFERENCES

[1] A. P. Bartók, M. C. Payne, R. Kondor, and G. Csányi, "Gaussian approximation potentials: The accuracy of quantum mechanics, without the electrons," Phys. Rev. Lett. **104**, 136403 (2010).

[2] J. Vandermause, Y. Xie, J. S. Lim, C. J. Owen, and B. Kozinsky, "Active learning of reactive Bayesian force fields applied to heterogeneous catalysis dynamics of H/Pt," Nat. Commun. **13**, 5183 (2022).

[3] S. Chmiela, A. Tkatchenko, H. E. Sauceda, I. Poltavsky, K. T. Schütt, and K.-R. Müller, "Machine learning of accurate energy-conserving molecular force fields," Sci. Adv. **3**, e1603015 (2017).

[4] S. Chmiela, H. E. Sauceda, K.-R. Müller, and A. Tkatchenko, "Towards exact molecular dynamics simulations with machine-learned force fields," Nat. Commun. **9**, 3887 (2018).

[5] S. Chmiela, H. E. Sauceda, I. Poltavsky, K.-R. Müller, and A. Tkatchenko, "sGDML: Constructing accurate and data efficient molecular force fields using machine learning," Comput. Phys. Commun. **240**, 38 (2019).

[6] K. T. Schütt, S. Chmiela, O. A. von Lilienfeld, A. Tkatchenko, K. Tsuda, and K.-R. Müller, *Machine Learning Meets Quantum Physics,* Lecture Notes in Physics Vol. 968 (Springer, 2020).

[7] V. L. Deringer, A. P. Bartók, N. Bernstein, D. M. Wilkins, M. Ceriotti, and G. Csányi, "Gaussian process regression for materials and molecules," Chem. Rev. **121**, 10073 (2021).

[8] J. Broad, S. Preston, R. J. Wheatley, and R. S. Graham, "Gaussian process models of potential energy surfaces with boundary optimization," J. Chem. Phys. **155**, 144106 (2021).

[9] F. M. Paruzzo, A. Hofstetter, F. Musil, S. De, M. Ceriotti, and L. Emsley, "Chemical shifts in molecular solids by machine learning," Nat. Commun. **9**, 4501 (2018).

[10] S. Liu, J. Li, K. C. Bennett, B. Ganoe, T. Stauch, M. Head-Gordon, A. Hexemer, D. Ushizima, and T. Head-Gordon, "Multiresolution 3D-DenseNet for chemical shift prediction in NMR crystallography," J. Phys. Chem. Lett. **10**, 4558 (2019).

[11] M. Cordova, M. Balodis, A. Hofstetter, F. Paruzzo, S. O. Nilsson Lill, E. S. E. Eriksson, P. Berruyer, B. Simões de Almeida, M. J. Quayle, S. T. Norberg *et al.*, "Structure determination of an amorphous drug through large-scale NMR predictions," Nat. Commun. **12**, 2964 (2021).

[12] R. Gaumard, D. Dragún, J. N. Pedroza-Montero, B. Alonso, H. Guesmi, I. Malkin Ondík, and T. Mineva, "Regression machine learning models used to predict DFT-computed NMR parameters of zeolites," Computation **10**, 74 (2022).

[13] S. M. Aguilera-Segura, D. Dragún, R. Gaumard, F. Di Renzo, I. M. Ondík, and T. Mineva, "Thermal fluctuation and conformational effects on NMR parameters in β-O-4 lignin dimers from QM/MM and machine-learning approaches," Phys. Chem. Chem. Phys. **24**, 8820 (2022).

[14] M. Ceriotti, "Beyond potentials: Integrated machine-learning models for materials," MRS Bull. **47**, 1045-1053 (2022).

[15] D. J. Tantillo, *Applied Theoretical Organic Chemistry* (World Scientific, 2018).

[16] Y. Segawa, M. Kuwayama, Y. Hijikata, M. Fushimi, T. Nishihara, J. Pirillo, J. Shirasaki, N. Kubota, and K. Itami, "Topological molecular nanocarbons: All-benzene catenane and trefoil knot," Science **365**, 272 (2019).

[17] E. E. Kwan and R. Y. Liu, "Enhancing NMR prediction for organic compounds using molecular dynamics," J. Chem. Theory Comput. **11**, 5083 (2015).

[18] S. Grimme, C. Bannwarth, S. Dohm, A. Hansen, J. Pisarek, P. Pracht, J. Seibert, and F. Neese, "Fully automated quantum-chemistry-based computation of spin–spin-coupled nuclear magnetic resonance spectra," Angew. Chem., Int. Ed. **56**, 14763 (2017).

[19] M. Lin, J. Xiong, M. Su, F. Wang, X. Liu, Y. Hou, R. Fu, Y. Yang, and J. Cheng, "A machine learning protocol for revealing ion transport mechanisms from dynamic NMR shifts in paramagnetic battery materials," Chem. Sci. **13**, 7863 (2022).

[20] A. G. Wilson, Z. Hu, R. Salakhutdinov, and E. P. Xing, *Deep Kernel Learning*, *Artificial Intelligence and Statistics* (PMLR, 2016), pp. 370–378.

[21] R. Calandra, J. Peters, C. E. Rasmussen, and M. P. Deisenroth, "Manifold Gaussian processes for regression," in *2016 International Joint Conference on Neural Networks (IJCNN)* (IEEE, 2016), pp. 3338–3345.

[22] In this work, we used a typical DFT/GIAO computational method for computing the chemical shifts. We need to examine various functionals and basis sets to show improved agreement with the range of experimental chemical shifts, and we are not focusing on such quantitative analysis here.

[23] T. Stadelmann, C. Balmer, S. Riniker, and M.-O. Ebert, "Impact of solvent interactions on ^{1}H and ^{13}C chemical shifts investigated using DFT and a reference dataset recorded in $CDCl_3$ and CCl_4," Phys. Chem. Chem. Phys. **24**, 23551 (2022).

[24] R. Flamary, N. Courty, A. Gramfort, M. Z. Alaya, A. Boisbunon, S. Chambon, L. Chapel, A. Corenflos, K. Fatras, N. Fournier *et al.*, "POT: Python optimal transport," J. Mach. Lear. Res. **22**, 1 (2021).

[25] M. Dračínskỳ, M. Buchta, M. Buděšínskỳ, J. Vacek-Chocholoušová, I. G. Stará, I. Starỳ, and O. L. Malkina, "Dihydrogen contacts observed by through-space indirect NMR coupling," Chem. Sci. **9**, 7437 (2018).

[26] M. Frisch, G. Trucks, H. Schlegel, G. Scuseria, M. Robb, J. Cheeseman, G. Scalmani, V. Barone, G. Petersson, H. Nakatsuji *et al.*, Gaussian 16 Revision C. 01. 2016, Gaussian Inc, Wallingford CT, 2016, p. 421.

[27] J. Wang, W. Wang, P. A. Kollman, and D. A. Case, "Automatic atom type and bond type perception in molecular mechanical calculations," J. Mol. Graphics Modell. **25**, 247 (2006).

[28] J. Wang, R. M. Wolf, J. W. Caldwell, P. A. Kollman, and D. A. Case, "Development and testing of a general amber force field," J. Comput. Chem. **25**, 1157 (2004).

[29] A. W. Sousa da Silva and W. F. Vranken, "ACPYPE - AnteChamber PYthon Parser interfacE," BMC Res. Notes **5**, 367 (2012).

[30] H. J. C. Berendsen, J. P. M. Postma, W. F. Van Gunsteren, A. DiNola, and J. R. Haak, "Molecular dynamics with coupling to an external bath," J. Chem. Phys. **81**, 3684 (1984).

[31] G. Bussi, D. Donadio, and M. Parrinello, "Canonical sampling through velocity rescaling," J. Chem. Phys. **126**, 014101 (2007).

[32] M. Parrinello and A. Rahman, "Polymorphic transitions in single crystals: A new molecular dynamics method," J. Appl. Phys. **52**, 7182 (1981).

[33] T. Darden, D. York, and L. Pedersen, "Particle mesh Ewald: An N·log(N) method for Ewald sums in large systems," J. Chem. Phys. **98**, 10089 (1993).

[34] B. Hess, H. Bekker, H. J. C. Berendsen, and J. G. E. M. Fraaije, "LINCS: A linear constraint solver for molecular simulations," J. Comput. Chem. **18**, 1463 (1997).

[35] M. J. Abraham, T. Murtola, R. Schulz, S. Páll, J. C. Smith, B. Hess, and E. Lindahl, "GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers," SoftwareX **1**, 19 (2015).

[36] M. Elstner, D. Porezag, G. Jungnickel, J. Elsner, M. Haugk, T. Frauenheim, S. Suhai, and G. Seifert, "Self-consistent-charge density-functional tight-binding method for simulations of complex materials properties," Phys. Rev. B **58**, 7260 (1998).

[37] Y. Yang, H. Yu, D. York, Q. Cui, and M. Elstner, "Extension of the self-consistent-charge density-functional tight-binding method: Third-order expansion of the density functional theory total energy and introduction of a modified effective Coulomb interaction," J. Phys. Chem. A **111**, 10861 (2007).

[38] B. Aradi, B. Hourahine, and T. Frauenheim, "DFTB+, a sparse matrix-based implementation of the DFTB method," J. Phys. Chem. A **111**, 5678 (2007).

[39] M. Gaus, A. Goez, and M. Elstner, "Parametrization and benchmark of DFTB3 for organic molecules," J. Chem. Theory Comput. **9**, 338 (2013).

[40] M. Kubillus, T. Kubař, M. Gaus, J. Řezáč, and M. Elstner, "Parameterization of the DFTB3 method for Br, Ca, Cl, F, I, K, and Na in organic and biological systems," J. Chem. Theory Comput. **11**, 332 (2015).

[41] S. Nosé, "A unified formulation of the constant temperature molecular dynamics methods," J. Chem. Phys. **81**, 511 (1984).

[42] W. G. Hoover, "Canonical dynamics: Equilibrium phase-space distributions," Phys. Rev. A **31**, 1695 (1985).

[43] S. Grimme, J. Antony, S. Ehrlich, and H. Krieg, "A consistent and accurate *ab initio* parametrization of density functional dispersion correction (DFT-D) for the 94 elements H-Pu," J. Chem. Phys. **132**, 154104 (2010).

[44] A. P. Bartók, R. Kondor, and G. Csányi, "On representing chemical environments," Phys. Rev. B **87**, 184115 (2013).

[45]M. J. Willatt, F. Musil, and M. Ceriotti, "Atom-density representations for machine learning," J. Chem. Phys. **150**, 154110 (2019).
[46]C. E. Rasmussen and C. K. Williams, *Gaussian Processes for Machine Learning*, Vol. 1 (MIT Press, Cambridge, MA, 2006), p. 95.
[47]L. Himanen, M. O. J. Jäger, E. V. Morooka, F. Federici Canova, Y. S. Ranawat, D. Z. Gao, P. Rinke, and A. S. Foster, "DScribe: Library of descriptors for machine learning in materials science," Comput. Phys. Commun. **247**, 106949 (2020).
[48]J. R. Gardner, G. Pleiss, D. Bindel, K. Q. Weinberger, and A. G. Wilson, "GPyTorch: Blackbox matrix-matrix Gaussian process inference with GPU acceleration," in *Advances in Neural Information Processing Systems* (Neural Information Processing Systems Foundation, 2018).
[49]G. Peyré, M. Cuturi *et al.*, "Computational optimal transport: With applications to data science," Found. Trends Mach. Learn. **11**, 355 (2019).
[50]I. Borg and P. J. Groenen, *Modern Multidimensional Scaling: Theory and Applications* (Springer Science and Business Media, 2005).
[51]F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg *et al.*, "Scikit-learn: Machine learning in Python," J. Mach. Learn. Res. **12**, 2825 (2011).