I. INTRODUCTION

Structure determines function – a hallmark paradigm in the atomistic sciences, ranging from biologists studying protein functions based on x-ray structures to organic chemists discussing reaction mechanisms based on NMR measurements\(^1\). The connection between structure \(\mathcal{R}\) and a compound’s function is given through statistical mechanics averages \(\mathcal{A}\) over the ensemble \(\mathcal{E}\) of Boltzmann weighted configurations,

\[
\mathcal{E} \rightarrow \mathcal{R} \rightarrow \mathcal{A}. \tag{1}
\]

The function of a molecule depends on the biological or chemical context e.g. solubilities or binding affinities – all of which can be expressed as phase space averages \(\mathcal{A}\). Understanding this relation is of fundamental importance as the temperature-dependent balance of configurations dictates the biological function of proteins and their macroscopic behavior (think of egg-white). Unfortunately, to quantitatively predict thermal averages as the temperature-dependent balance of configurations not only requires bond topologies and therefore enables a general machine learning pathway to ensemble properties throughout chemical compound space. We demonstrate AIML for predicting Boltzmann averaged structures after training on hundreds of MD trajectories. AIML output is subsequently used to train machine learning models of free energies of solvation using experimental data, and reaching competitive prediction errors (MAE \(\sim 0.8\ \text{kcal/mol}\) for out-of-sample molecules– within milli-seconds. As such, AIML effectively bypasses the need for MD or MC-based phase space sampling, enabling exploration campaigns throughout CCS at a much-accelerated pace. We contextualize our findings by comparison to state-of-the-art methods resulting in a Pareto plot for the free energy of solvation predictions in terms of accuracy and time.

Equilibrium structures determine material properties and biochemical functions. We propose to machine learn phase-space averages, conventionally obtained by \textit{ab initio} or force-field based molecular dynamics (MD) or Monte Carlo simulations. In analogy to \textit{ab initio} molecular dynamics (AIMD), our \textit{ab initio} machine learning (AIML) model does not require bond topologies and therefore enables a general machine learning pathway to ensemble properties throughout chemical compound space. We demonstrate AIML for predicting Boltzmann averaged structures after training on hundreds of MD trajectories. AIML output is subsequently used to train machine learning models of free energies of solvation using experimental data, and reaching competitive prediction errors (MAE \(\sim 0.8\ \text{kcal/mol}\) for out-of-sample molecules– within milli-seconds. As such, AIML effectively bypasses the need for MD or MC-based phase space sampling, enabling exploration campaigns throughout CCS at a much-accelerated pace. We contextualize our findings by comparison to state-of-the-art methods resulting in a Pareto plot for the free energy of solvation predictions in terms of accuracy and time.

\[\text{Equilibrium structures determine material properties and biochemical functions.} \]

\[\text{We propose to machine learn phase-space averages, conventionally obtained by} \textit{ab initio} \text{or force-field based molecular dynamics (MD) or Monte Carlo simulations. In analogy to} \textit{ab initio} \text{molecular dynamics (AIMD), our} \textit{ab initio} \text{machine learning (AIML) model does not require bond topologies and therefore enables a general machine learning pathway to ensemble properties throughout chemical compound space. We demonstrate AIML for predicting Boltzmann averaged structures after training on hundreds of MD trajectories. AIML output is subsequently used to train machine learning models of free energies of solvation using experimental data, and reaching competitive prediction errors (MAE } \sim 0.8\ \text{kcal/mol}\text{ for out-of-sample molecules– within milli-seconds. As such, AIML effectively bypasses the need for MD or MC-based phase space sampling, enabling exploration campaigns throughout CCS at a much-accelerated pace. We contextualize our findings by comparison to state-of-the-art methods resulting in a Pareto plot for the free energy of solvation predictions in terms of accuracy and time.}\]
FIG. 1. Example of a chemical compound space (CCS) for $N_p = N_e = 32$ protons (and electrons) with a hierarchy of composition, constitution, and conformation. Each level corresponds to distinct temperature-regimes and is described by specific quantum chemical methods. Ab initio machine learning (AIML) can act on all three levels and does not require fixed constitutions but allows general ensemble predictions using ML-based averaged structures and free energy machine learning (FML).

To address the length and timescale problem of conformational space across CCS, we have recently introduced Free energy Machine Learning (FML) which relies on the averaged structure $\mathcal{R}$ as input to predict ensemble averages such as the free energy, $\mathcal{A}_{FML} \rightarrow \mathcal{A}$. Averaging the structure is necessary because ensemble properties inherently depend on multiple configurations and as such using a single geometry per molecule introduces ambiguities to the ML model. Here, we propose to replace the preceding step, i.e. the generation of the averaged input through extensive molecular dynamics runs for any query compound by an ab initio machine learning (AIML) model, $\mathcal{E}_{AIML} \rightarrow \mathcal{R}$.

AIML makes use of the Graph-To-Structure (G2S) method to predict three-dimensional structures though chemical compound space. Training data and labels, however, are fundamentally different from G2S: Instead of using a single optimized conformer per molecule, AIML training data consists of averages over complete MD trajectories, enabling the prediction of thermodynamically averaged conformers. This accounts for the fundamentally important difference between a Boltzmann average and a single atomic configuration for ensemble property predictions, as previously discussed in free energy machine learning (FML).

In particular, we replace ensemble sampling with machine learning of ensemble averages $\mathcal{A}$ of an equilibrium property $a(r, p)$ by,

$$\mathcal{A} = \frac{1}{Z} \int a(r, p) e^{-\beta E(r, p)} \, dr \, dp \approx \mathcal{A}_{ML} (\mathcal{R}),$$

(2)

with $\beta$, $Z$, and $E$ being the Boltzmann-factor, the partition function, and the total energy, respectively. The approximate equality of Eq. 2 is achieved by training an ML model $\mathcal{A}_{ML}$ that uses only the averaged conformer $\mathcal{R}$ with values $\mathcal{A}$ that include the rigorous integral over the ensemble.

AIML is a purely ML-based framework and properly accounts for the underlying Boltzmann statistics and can subsequently be used to generate the appropriate input for FML model-based predictions. The goal of our work is not to build an ML model that perfectly reflects thermodynamic expectation values but to construct a surrogate model that can predict these integrals with high accuracy and speed. By including Boltzmann averaged conformers we make sure to include a canonical mapping of the underlying ensemble of each molecule to the ensemble.
FIG. 2. Centered histograms and standard deviation $\sigma_d$ of all distances between ten non-hydrogen atoms of aspirin extracted from an ab initio molecular dynamics (AIMD) trajectory\textsuperscript{11} (a). Conventional AIMD and ab initio ML (AIML) map the ensemble $\mathcal{E}$ to averaged structure $\mathcal{R}$ to the statistical mechanics (SM) average $\mathcal{A}$ (b). Bold labeled index pairs $(i,j)$ define a bond topology (c). Sketch of two principal components PC1, PC2 of ML-based representations\textsuperscript{32–35} on a fictitious free energy surface (d) corresponding to conformers represented by a disconnectivity graph\textsuperscript{36}. Instead of MD average representations\textsuperscript{11} (FML) $\mathbf{X}$ we propose AIML predicted representations of averaged conformers $\mathbf{x}(\mathcal{R})$.

II. METHODS

A. FML and G2S

For the previously published free energy machine learning\textsuperscript{11} (FML) approach first all sampled geometries $\mathbf{r}$ had to be transformed to representation vectors $\mathbf{x}(\mathbf{r})$ before finally the average $\mathbf{X}$ over the representation vectors,

$$\mathbf{X} = \int \mathbf{x}(\mathbf{r}) e^{-\beta E(\mathbf{r})} \, d\mathbf{r},$$

could be computed. The key advantage of AIML is that instead of explicitly sampling conformer space, a single AIML evaluation is required to predict a surrogate vector.
x(R) evaluated for the system average R to replace X (s. Fig. 2d).

Graph-To-Structure (G2S) exploits implicit correlations among relaxed structures in training data sets to infer interatomic distances for out-of-sample compounds across chemical space. G2S effectively enables direct reconstruction of three-dimensional coordinates, thereby allowing circumventing conventional energy optimization. G2S can reach an accuracy on par or better than conventional structure generators. G2S can reach an accuracy on par or better than conventional structure generators. As query input, G2S requires only bond-network and stoichiometry-based information. G2S learns the direct mapping from a chemical graph to that structure that had been recorded in the training data set. For the prediction of new structures, only molecular connectivity is needed, which can be provided e.g. via SMILES\textsuperscript{39} or SELFIES\textsuperscript{40}. The G2S machines predict all pairwise distances. The full 3D geometry is then reconstructed using DGSOL\textsuperscript{41} for heavy atoms and a Lebedev sphere optimization scheme for hydrogen atoms.

**B. Workflow**

![Workflow diagram](image)

FIG. 3. To generate ab initio machine learning (AIML) training data, conformer space is sampled to obtain Boltzmann weighted average distance matrices for given molecular graphs. AIML predicts the averaged distance matrix, which is then converted to a three-dimensional geometry R. Finally, the ensemble average A is predicted using free energy machine learning\textsuperscript{11} (FML).

An essential ingredient for AIML was extending the previous Graph-To-Structure\textsuperscript{37} (G2S) method via the introduction of an averaged geometry R. This enables a computationally efficient ML-based map between ensemble and free energies and addresses the conformer sampling bottleneck. AIML provides an effective representation for the conformer ensemble by mapping all degrees of freedom to a single averaged structure to approximate the ensemble-averaged structure with corresponding AIML representation vector x (cf. Fig. 2). A schematic overview of the steps for AIML training and prediction is given in Fig. 3. As we will discuss in the following two steps are needed to combine both methods: i) use Boltzmann weighted distance matrices D for training ii) use AIML average conformer predictions as input for an ensemble average model (s. Fig. 3 right).

The required training steps for the AIML structure prediction use Boltzmann-averaged intramolecular distance matrices. More specifically, as a first step, the molecule is transformed to a graph-based representation\textsuperscript{37} g with the average distance matrix D as training labels resulting in ML models for heavy atom pairs as well as heavy and hydrogen atoms. Before entering the next step of free energy prediction, AIML is trained with the maximal number of molecules (N = 512) to construct the average training and test set conformers. This process is repeated for the complete dataset with consistent training test splits between AIML structure and free energies prediction. Next, a machine (s. sec. II C) for learning free energies is trained using AIML predicted average geometries. Finally, AIML can be used for out-of-sample predictions (s. Fig. 3). Based on the molecular graph, the Boltzmann weighted distance matrix D is predicted.

Next, a distance-geometry solver\textsuperscript{41} (DGSOL) is used to convert the distance matrix to three-dimensional coordinates. The predicted average conformer then serves as a link between the graph and corresponding three-dimensional geometry (s. Fig. 2). Subsequently, the average predicted conformer R is transformed to a single Bag-of-Bonds\textsuperscript{42} (BoB) representation vector x(R) and used to predict the free energy.

**C. Machine learning**

ML, in particular of solutions to the Schrödinger equation i.e. quantum machine learning\textsuperscript{43–47} (QML), allows navigating chemical compound space (CCS) with high efficiency and precision. ML has become a popular avenue to material science with applications to atomization energies\textsuperscript{48,49}, crystal formation energies\textsuperscript{50}, carbenes\textsuperscript{51}, excited states\textsuperscript{52,53}, oxidation states\textsuperscript{54}, nuclear magnetic resonance spectra\textsuperscript{55}, reaction barriers\textsuperscript{56,57}, magnetic systems\textsuperscript{58} and charge transfer\textsuperscript{59} or molecular fragments\textsuperscript{60,61}. Many recent approaches are based on individual levels of CCS, i.e. composition\textsuperscript{62,63}, constitution\textsuperscript{64–67} (s. Fig. 1). Moreover ensemble properties such as the free energy of solvation\textsuperscript{11,68–74}, melting
points, magnetic anisotropy tensors, phases of water, have previously been addressed with ML.

Our ML approach is based on Kernel-Ridge Regression (KRR) a supervised learning method that allows approximating arbitrary functional relationships between input data given as molecular representations \(x\) and properties \(A(x)\). Using KRR \(x\) is mapped into a high dimensional feature space rendering the regression problem linear. A remarkable result of KRR is that the mapping does not need to be carried out explicitly, instead, the distances between representations \(x_i\) and \(x_j\) are computed e.g. using Gaussian kernel functions,

\[
 k(x_i, x_j) = \exp\left(-\frac{||x_i - x_j||^2}{2\sigma^2}\right), \tag{4}
\]

that measure the similarity between two compounds \(i\) and \(j\) resulting in the kernel matrix \(K\) where \(\sigma\) is the kernel-width hyperparameter. We use KRR to predict the vector of all interatomic distances \(D_q\) of a query compound \(q\) and a distance geometry solver (DSGOL) for subsequent reconstruction of the three-dimensional geometry.

The vector \(D_q\) contains all distances of atoms in the query molecule. The average conformer prediction \(R(x_q = g_q)\) of a query molecule \(q\) represented by a graph \(g_q\) is given by,

\[
 R(g_q) = \text{DGSOL}(D(g_q)), \tag{5}
\]

where the distance prediction is as follows,

\[
 D_q = K \cdot \alpha. \tag{6}
\]

Here, the kernel matrix \(K\) is evaluated between query and training compounds \(g_i\) and \(g_q\) with regression coefficients \(\alpha\). The optimal regression coefficients \(\alpha\) are obtained by solving a set of equations,

\[
 \alpha = (K + \lambda \cdot I)^{-1} D, \tag{7}
\]

where the vector \(D\) contains all distances between atoms for each of the training molecules and \(\lambda\) is a regularization parameter. The ensemble property prediction is given by:

\[
 A_q = \bar{K} \cdot \bar{\alpha}. \tag{8}
\]

As before, for training \(\bar{K}\) is evaluated between all training compounds now using molecular representation vectors \(x(R)\):

\[
 \bar{\alpha} = (\bar{K} + \lambda \cdot I)^{-1} A, \tag{9}
\]

where the vector \(A\) contains the values of the ensemble property in the training set. Learning curves quantify the model prediction error, often measured as mean absolute error (MAE), against the number of training samples \(N\) and are key to understand the efficiency of ML models. It is generally found that they are linear on a log-log scale,

\[
 \log\left(\frac{\text{MAE}}{\text{unit}}\right) \approx I - S \cdot \log(N), \tag{10}
\]

where \(I\) is the initial error and \(S\) is the slope indicating model improvement given more training data.

III. RESULTS

A. Concept of equilibrium structure prediction

Within AIML, we view the averaged structure as an ensemble property, \(R = A\), representing the connection to the aforementioned overarching theme of structure determining function. In addition, the thermal equilibrium structure is relevant for NMR spectroscopy which accounts for protein flexibility by resulting in time-averaged structures equivalent to \(R\) due to the ergodic theorem. Purely ML-based implementation bypasses routine sampling issues – de facto replacing MD simulations with predicted averages as ensemble fingerprints for subsequent property prediction.

We first use kernel-ridge regression (KRR) to predict the symmetric matrix of averaged interatomic distances of a query compound \(q\), i.e. \(D_q \approx K \cdot \alpha\) which contains all inferred averaged distances of atoms in the query molecule. \(K\) and \(\alpha\) correspond to the kernel matrix and training weights obtained for a training set consisting of MD trajectories and averaged interatomic distances as labels. In analogy to Graph-To-Structure (G2S), we subsequently rely on the distance geometry solver (DSGOL) for reconstruction of the three-dimensional structure \(R\), as well as on graph-based representations, \(g_q\). To exemplify the AIML approach, consider the \(ab\ initio\) molecular dynamics (AIMD) trajectory published in Ref. (31) of the aspirin molecule at 300 K, resulting atomic distance histograms in Fig. 2a. In order to establish a graph-based representation to replace \(ab\ initio\) MD with \(ab\ initio\) ML (AIML) as illustrated in Fig. 2b, it is necessary to assign bonds. This is straightforward using the distance histogram as covalently bonded atoms will not move far from each other (see Fig. 2c) – a concept that can be generalized via coarse-graining (21,84–86). While AIML requires a suggested molecular graph it is not restricted to a single fixed bond topology but allows adapting the molecular graph depending on the relevant degrees of freedom depending on temperature or the environment of the molecule. Thus AIML can include the formation and breaking of bonds (cf. Fig. 1) corresponding to adding or erasing a one in the bond topology matrix \(g_{\delta p}\) (cf. Fig. 2). Note that the predicted pairwise distance matrix does not only account for nearest neighbor effects but the complete many-body description since it includes all cross combinations of atomic distances.

Using the graph as the representation for construct-
ing kernels, the AIML model then learns the center of each off-diagonal element in the distance histogram as a label. Next, the AIML predicted distances are used to reconstruct the average conformer. Subsequently the ML representation vector $\mathbf{x}(\mathbf{R})$ is computed. Therefore, AIML allows exchanging the order of average evaluation compared to the previous FML approach (illustrated Fig. 2) resulting in a dramatic reduction of computational costs.

AIML proposes a different paradigm by connecting all hierarchies of CCS with ensemble properties into a single ML-based framework. Because of its generality, AIML does not require a priori information about bonds but only averaged atomic distances. In analogy to the hierarchy of CCS, AIML has several special cases with fundamental physical interpretations (s. Fig. 1): At temperatures higher than most bond energies AIML operates on atomic clusters corresponding to a topology matrix that (mostly) contains zeros i.e. the composition. At moderate temperatures, bonds exist but may occasionally break corresponding to adding or removing a zero in the topology matrix. In this case, multiple molecular graphs can be extracted and AIML predicts the constitution averaged structure.

B. Application: Ensemble to structure to property

In this section, we demonstrate the usefulness of AIML for the problem of accurate predictions of free energies of solvation. In particular, we focus on experimental free energies of solvation of 642 charge-neutral small to medium-sized bio-organic molecules, as encoded in the FreeSolv database\(^5\). Solvation free energies of fundamental importance for chemistry, and the FreeSolv database has become a popular benchmark for performance testing of novel models\(^37,88,89\). To use AIML to predict averaged structures and FML to predict free energies of solvation for out-of-sample molecules, we first trained AIML models as described above using molecular graphs as input and as labels for the averaged distances.

These were obtained from extended force-field based MD runs and density functional theory (DFT) for conformer sampling with Def2TZVPD-FINE\(^101,104\) basis set and Becke-Perdew\(^87,88\) (BP) functional in the gas-phase (s. sec. IV A for details). Note that we neglect the effect of water on the phase space of the solute (e.g. through hydrogen bonds). We believe that this aspect warrants further in-depth investigation within subsequent studies in the future.

Depending on the temperature (cfg. Fig. 1), some degrees of freedom do not get averaged out by the phase space integral (Eq. 2), restricting the domain to certain local basins of the total free energy. These remaining degrees of freedom can be identified in the way described in Fig 2a). We demonstrate the idea of AIML for ambient temperatures in the gas phase for which conformational molecular graph topologies as in biochemistry hold – without any loss of generality. Within such a regime, we can safely assume that any distance histogram matrix would be consistent with one topology which represents the coarse-grained back-bone that is not averaged out by the phase space integral. Of course, this approach could also be applied to any other set of conditions presuming that there is some way to easily infer valid topologies as a function of conditions (like temperature and composition). The latter is a separate problem that goes beyond the scope of this work.

As numerical evidence of the functionality of the AIML idea, we present in Fig. 3a) prediction errors of the three-dimensional Boltzmann averaged structures $\mathbf{R}$ as a function of training set size (aka learning curves\(^81,105,106\)).

Numerical results shown in Fig. 4a) indicate a systematic linear improvement on a log-log scale\(^107\) as a function of the size of the training set $N_s$, that is, the number of averaged training structures. Note that $N$ stands for the number of training points for free energy values and that we have trained and evaluated two different machines after each other, for structure and free energy prediction respectively. For the maximal training set size considered ($N_s = 512$ molecules, 80% of FreeSolv), the average root-mean-square deviation\(^108-110\) (RMSD), a measure of structural distance between structures, has decayed to only 0.80 Å for predicted FF and 0.82 Å for DFT averaged conformers, the slope of the learning curve, however, indicates that learning has not yet been saturated. We find structure prediction of large molecules a particularly hard problem i.e. on average the RMSDs increase with the size of the structure. This was also observed for a random subset of the GDB17 molecular database\(^111\) (s. SI Fig. 7b for the scatter plot of molecular size vs. RMSD indicating a rough correlation).

Note that the corresponding learning curve predicting optimal (not Boltzmann-averaged) distances is less steep for the molecules in FreeSolv, and exhibits a higher offset (s. SI Fig. 4). This could be due to the fact that learning thermal averages is less ambiguous for AIML than learning potential energy minima as it is in the case of G2S. This also holds for the heavy atom-hydrogen distances (s. SI Tab. 1 and Fig. 1 in SI). From a different point of view, by computing the Boltzmann average over distances conformer flexibility is effectively integrated out, thus simplifying learning compared to the optimized structures.

Next, we demonstrate the learning efficiency (s. Fig. 4b) of AIML for free energy prediction based on previously predicted structures $\mathbf{R}$. The main disadvantage of the preceding free energy machine learning\(^11\) (FML) model was that it requires explicit conformer sampling for free energy prediction. The novel advantage of AIML is that no sampling is required. Instead, after prediction of the average conformer $\mathbf{R}$, the free energy prediction is based on the ML representation vector $\mathbf{x}(\mathbf{R})$ where $x$ is the ML-based representation Bag-of-Bonds\(^42\) (BoB). Encouragingly, FML (requiring explicit MD sampling) and AIML exhibit similar learning curves, achiev-
FIG. 4. Learning curve of the Root-Mean-Square Deviation (RMSD), a measure of structural distance, of ab initio machine learning (AIML) three-dimensional structure predictions as a function of the number of training structures $N_s$ using GAFF2 force field and density functional theory (DFT) with Becke-Perdew (BP) functional for training structure sampling (a). Two predicted average conformers show the improvement of structure prediction along the learning curve (at $N_s = 32, 128, 512$). The mean absolute error (MAE) of predicted free energies $F$ of the FreeSolv database as a function of $N$ for free energy machine learning (FML) (b) using MD sampling versus AIML (no sampling) and the Bag-of-Bonds (BoB) representation. The AIML $N_s = 32$ model was trained with only 32 structures for conformer prediction.

We find AIML to perform consistently better when the training distances result from Boltzmann sampling instead of using optimized structures - similar to what we also noted above for the structure prediction (see SI Fig. 9). Even more surprising is the observation that training free energy prediction with AIML predicted structures resulted in slightly better models than training with the ground-truth averaged conformers resulting from MD. This indicates that AIML effectively smoothens conformer space by isolating the most important degrees of freedom, thus facilitating structure-based regression of thermal averages.
C. Assessment of efficiency

To gain a more comprehensive idea of AIML’s value to the field, we have assessed the cost accuracy trade-off. Testing AIML on the FreeSolv database we have measured average prediction times of 41 ms/molecule. These prediction times are dominated by the structure reconstruction task (40 ms), while only 1 ms is required to yield the free energy estimate (on a single-core AMD EPYC 7402P compute chip). For comparison, the corresponding prediction based on a classical force-field MD simulation protocol would have consumed three to four orders of magnitude more time, not to mention the costs associated with quantum chemistry based \textit{ab initio} MD. To gain a comprehensive overview of the field, we have performed solvation free energy calculations for all of the FreeSolv molecules using the following methods (all listed free energies available, s. sec. V, MAEs listed in SI Tab. II):

1. \textit{Solvation model based on density} (SMD) at M06-2X/Def2-TZVP\textsuperscript{112} (timing for SMD implemented in Gaussian\textsuperscript{113} not to be published)
2. COSMO-RS-B1 and COSMO-RS-B2 referring to COSMO-RS\textsuperscript{95-96,114-116} with Def2TZVP\textsuperscript{101-102} and Def2TZVPD-FINE\textsuperscript{101-104} basis sets respectively and Becke-Perdew\textsuperscript{57,88} (BP) functional in the gas-phase
3. \textit{Reaction mechanism generator group} (RMG) solvation\textsuperscript{117}
4. \textit{ddCOSMO}\textsuperscript{118} results obtained with PySCF\textsuperscript{119} and PBE-0\textsuperscript{120,121}/Def2TZVPD\textsuperscript{101-104}
5. Generalized Born\textsuperscript{122,123} (GBSA) model, results obtained using AMBER\textsuperscript{2,6}
6. Free energy machine learning\textsuperscript{11} (FML) with explicit conformer sampling on FreeSolv database

To complete the picture, we also included literature values for FreeSolv concerning the methods ARROW-PIMD\textsuperscript{20}, Thermodynamic integration (TI) with GAFF\textsuperscript{2,5,6} extracted from the FreeSolv\textsuperscript{59} database and reference interaction site model\textsuperscript{100,124-126} (3D-RISM). The trade-off between cost and accuracy, including an outline of the resulting Pareto front, is displayed in Fig. 5.

We note that AIML adds to the convexity of the Pareto front, representing a meaningful compromise: Although roughly twice as slow, it is slightly more accurate at 0.82 kcal mol\textsuperscript{-1} than the Reaction Mechanism Generator\textsuperscript{127} (RMG) model (MAE of 0.98 kcal mol\textsuperscript{-1}), but still four orders of magnitude faster than the next best \textit{ab initio} method COSMO-RS\textsuperscript{95,96,114-116}. A list of all MAE is provided in SI Tab. 2. Thus, AIML is positioned on the Pareto front of the available solvation methods located in a sweet spot between speed and accuracy, providing the fastest predictions at the given accuracy of about 0.82 kcal mol\textsuperscript{-1}. Note that the AIML learning curves have not yet saturated and that its accuracy will likely further improve if more training samples are included (c.f. Fig. 4). Improving the AIML model will hardly worsen the prediction time due to the linear scaling of KRR predictions w.r.t. training set size (s. sec. II C) and therefore shift the Pareto front towards higher accuracy. Furthermore, it is important to note that arbitrary accurate \textit{ab initio} trajectories can be used for training while the prediction time is independent of the level of theory. We expect that recently published ML models tailored towards solvation such as SoluteML\textsuperscript{128} may outperform the presented AIML models’ accuracy, but we note that only a small training set of \(N = 512\) molecules was used and we expect MAE to decay further with the training set size. A3D-PNAConv-F\textsuperscript{129} combines the 2D and 3D structure and transfer learning achieving an MAE of 0.417 kcal mol\textsuperscript{-1} for the FreeSolv data set. Predictions require conformer sampling using a FF and the lowest-energy conformer. In contrast, AIML does not require sampling and is in fact replacing the functionality of a force-field or of an \textit{ab-initio} calculation.

Note that we have also tried to combine RMG and AIML/FML via the \(\Delta\)-ML approach\textsuperscript{130} where RMG is used as a baseline for AIML, but unfortunately, the prediction errors did not improve (s. SI Fig. 8). Moreover, AIML performs worse for large molecules with many conformers (s. SI Fig. 7b). Unfortunately, combining a random sampled GDB17\textsuperscript{111} dataset with 10000 molecules with the FreeSolv average conformers did not lead to improved structure predictions due to the small overlap of
the two data sets (s. SI Fig. 6). Specifically, the small training set size and very high chemical diversity of the FreeSolv database including the elements C, H, O, S, N, F, I, Br, P, Cl, and up to 24 non-hydrogen atoms per molecule limit the accuracy of structure prediction for large compounds. Instead of adding random structures to improve structure prediction for the FreeSolv database (s. SI Fig. 6) we could show that sampling\textsuperscript{131} the local chemical space of the largest FreeSolv can help to improve the models’ accuracy (s. SI Fig. 8). Alternatively, these problems may be resolved by improved graph-based representations that include information about local chemical substructures, leading to better structure and improved free energy predictions.

Recent graph-based ML models\textsuperscript{132,133} can achieve a competitive accuracy with root mean squared errors (RMSEs) around 1 kcal mol\textsuperscript{−1}. We have achieved a similar RMSE of 1.35 kcal mol\textsuperscript{−1} for a training set size of \( N = 512 \). We emphasize that the AIML approach is very different: First, AIML uses three-dimensional conformations which can lead to a much-improved accuracy (MAE of 0.57 kcal mol\textsuperscript{−1} for \( N = 490 \)) as we have shown earlier\textsuperscript{11} and allows going beyond fixed graphs. Second, AIML also predicts ensemble-based representation whereas SMILES-based ML use molecular graphs as input. We found a direct comparison with the two previously mentioned other ML methods\textsuperscript{132,133} difficult because they either use a different training-test split\textsuperscript{132} or neglect\textsuperscript{133} certain molecules of the FreeSolv database\textsuperscript{89}. The comparison shows that our method might have a slightly higher initial offset due to having to learn the representation i.e. the averaged conformer before predicting the free energy. Our learning curves (s. Fig. 4) do not indicate saturation of the MAE with training set size \( N \) and might still surpass graph-based models for large training set sizes because AIML contains information about molecular conformations.

**IV. CONCLUSION**

We have introduced \textit{ab initio} machine learning (AIML) allowing for efficient predictions of ensemble averages which systematically improve in accuracy as training set sizes grow. To the best of our knowledge, for the first time, AIML effectively bypasses the need for extensive MD or MC simulations to directly infer Boltzmann averaged geometries. Unlike all other solvation models (shown in Fig. 5) the AIML framework could easily be applied to other ensemble properties, e.g. melting points, without much adaptation since no manual pre-selection of features for molecular fingerprints is required. AIML does not require any additional sampling for inferring ensemble averages of new out-of-sample query molecules: Instead AIML accounts for multiple Boltzmann weighted configurations implicitly through its training data. We have exemplified AIML for estimating experimental solvation free energies, and our numerical results amount to evidence showing that the conformer ensemble can effectively be linked to a single averaged conformer that serves as a canonical representative. AIML predictions are consistent with the previous free energy machine learning\textsuperscript{131} (FML) approach without the need to run an MD simulation for each prediction, reaching errors as low as 0.82 kcal mol\textsuperscript{−1} for 41 CPU-ms/molecule prediction cost.

Further analysis has revealed that AIML does not yet work well with all available molecular representations\textsuperscript{72}. More specifically, we find that representations, tailored toward atomization energies and including explicit angular dependencies, such as FCHL19\textsuperscript{134,135}, yield less favorable AIML models (s. SI Fig. 9). Conversely, it might be possible to further improve AIML by tailoring and optimizing representations and architecture (e.g. using locality, symmetry, neural networks).

The question of uniqueness is very fundamental for molecular representations\textsuperscript{136–141}. For the present dataset, however, this was not an issue as all averaged representations distinguished all data items. It is possible, however, to imagine a scenario where this is not the case: For two different ensembles, \( \mathcal{E} \) and \( \mathcal{E}' \) with the same average conformer \( \mathcal{R} = \mathcal{R}' \), but two different averages \( \mathcal{A} \neq \mathcal{A}' \), the presented AIML model would make the same predictions and could not distinguish between these two systems. However, this problem could be resolved by including higher-order moments for the prediction of the representation, like including an AIML model for the standard deviation of the representation. Future work will deal with this question. Here our main focus was on molecules with uniquely defined bond topologies. For future applications, AIML could be applied to cases where assigning bond topologies is ambiguous or impossible such as transition states\textsuperscript{142,143} or molecules at very high temperatures. These are important cases where AIML can be used, but graph-based ML models cannot be used.

In summary, in comparison to classical or \textit{ab initio} MD-based predictions of free energies of solvation, AIML offers respective speed-ups by four to seven orders of magnitude. AIML achieves such speed-ups by effectively shifting computational cost for the query prediction to the training set generation. However, in light of the sheer scale of the chemical compound space available for molecular queries, this trade-off might be useful.

**A. Conformer and Free Energy Data**

ML based on a single geometry can lead to ambiguous predictions\textsuperscript{11} ensemble property predictions because predictions can vary substantially depending on the conformer. Many body representations\textsuperscript{134,135,144–146} rely on three-dimensional geometries, which becomes even more relevant if the target property depends on multiple relevant conformers. A solution to this issue is to sample configuration space to obtain a conformer invariant ML representations\textsuperscript{11}. Sampling can be achieved
by different strategies: MD simulations\textsuperscript{5,6}, systematic conformation ensemble scans\textsuperscript{147-149} and conformation generation methods either knowledge or force field (FF) based such as ETKDG6\textsuperscript{150}, Gen3D\textsuperscript{151} and others\textsuperscript{152-156}. A more expensive method but accurate method is to obtain conformations using \textit{ab initio} approaches such as density functional theory (DFT)\textsuperscript{157-159} or tight binding\textsuperscript{145-147} (TB). Despite these advantages, a common pitfall of these methods is that sometimes extension to arbitrary chemistries is not straightforward. To this end, ML-based methods\textsuperscript{57,157-159} hold the promise of providing faster and more general structure predictions. There are only very few ML structure generation methods e.g., based on reinforcement learning\textsuperscript{160} or stochastic normalizing flows\textsuperscript{161} that take energy weights of different conformers into account. Here, to obtain a diverse set of conformers as the AIML training set we have performed MD simulations in vacuum at an elevated temperature of \(T = 350\,\text{K}\) using OpenMM\textsuperscript{26} using a Langevin integrator. GAFF2\textsuperscript{5,6} with a time-step of \(\Delta t = 2\,\text{fs}\) was used with a total simulation time of 2 ns. Partial charges are computed with antechamber\textsuperscript{5,6} at AM1-BCC\textsuperscript{162} level. MD samples are selected with 2 ps time separation. To compare AIML with COSMO-RS\textsuperscript{35,96,114-116} solvation method, we used the COSMO-RS workflow based on \textit{ab initio} DFT calculations with Turbomole\textsuperscript{63} and the Becke-Perdew (BP)\textsuperscript{87,88} functional as implemented in COSMOconf\textsuperscript{164} with two different basis sets, Def2TZVP and Def2TZVPD-FINE\textsuperscript{101-104} (for future reference referred to as B1 and B2). Based on these results, free energies are extracted using the COSMOtherm\textsuperscript{165} program. In addition, the reaction mechanism generator group (RMG) based approach was used to compute free energies of solvation\textsuperscript{117} of the FreeSolv database via the lerulicom API\textsuperscript{166}. The FreeSolv\textsuperscript{99} dataset contains 642 charge neutral compounds and their experimental free energies of solvation. The average unsigned error of the experimental values is 0.57 kcal mol\textsuperscript{-1} close to the level of thermal energy fluctuations (\(k_B \cdot 300\,\text{K} \approx 0.6\,\text{kcal mol}^{-1}\)). All ML models use a maximal training set size of 80% corresponding to \(N = 512\) molecules. Hyperparameters are optimized with nested five-fold cross validation.

V. DATA AND CODE AVAILABILITY

The AIML code and all free energies of solvation of the FreeSolv database (if produced by the authors) are published in a freely available repository https://doi.org/10.5281/zenodo.6401711. We gladly provide more data for specific requests.

VI. ACKNOWLEDGEMENTS

We acknowledge support from the European Research Council (ERC-CoG Grant QML). This project has received funding from the European Union’s Horizon 2020 research and innovation program under Grant Agreement #772834. This research was also supported by the NCCR MARVEL, a National Centre of Competence in Research, funded by the Swiss National Science Foundation (grant number 182892).

VII. DECLARATION OF CONFLICTING INTERESTS

DL, GFvR, and OAvL are shareholders of Leruli GmbH.

[1] A. C. T. van Duin, S. Dasgupta, F. Lorant, and W. A. Goddard, The Journal of Physical Chemistry A 105, 9396 (2001).
[2] R. Car and M. Parrinello, Phys. Rev. Lett. 55, 2471 (1985).
[3] C. Oostenbrink, A. Villa, A. E. Mark, and W. F. Van Gunsteren, Journal of Computational Chemistry 25, 1656 (2004), https://onlinelibrary.wiley.com/doi/10.1002/jcc.20090.
[4] M. Christen, P. H. H{"u}nenberger, D. Bakowies, R. Baron, R. B{"u}rgi, D. P. Geerke, R. B{"u}rgi, D. P. Geerke, T. N. Heinz, M. A. Kastenholz, V. Kr{"a}utler, C. Oostenbrink, C. Peter, D. Trzesniak, and W. F. van Gunsteren, Journal of Computational Chemistry 26, 1719 (2005), https://onlinelibrary.wiley.com/doi/pdf/10.1002/jcc.20303.
[5] J. Wang, W. Wang, P. A. Kollman, and D. A. Case, J. Mol. Graph. Model. 25, 247 (2006).
[6] J. Wang, R. M. Wolf, J. W. Caldwell, and P. A. Kollman, J. Comp. Chem. 25, 1157 (2004).
[7] C. Vega, J. L. F. Abascal, M. M. Conde, and J. L. Aragones, Faraday Discuss. 141, 251 (2009).
[8] W. L. Jorgensen, D. S. Maxwell, and J. Tirado-Rives, Journal of the American Chemical Society 118, 11225 (1996).
[9] W. L. Jorgensen, J. Chandrasekhar, J. D. Madura, R. W. Impey, and M. L. Klein, The Journal of Chemical Physics 79, 926 (1983).
[10] J. B. Abrams, L. Rosso, and M. E. Tuckerman, The Journal of chemical physics 125, 074115 (2006).
[11] J. Weinreich, N. J. Browning, and O. A. von Lilienfeld, The Journal of Chemical Physics 154, 134113 (2021).
[12] D. Voet and J. G. Voet, Biochemistry (John Wiley and Sons, Inc., Hoboken, NJ, USA, 2006).
[13] O. A. von Lilienfeld, K.-R. M{"u}ller, and A. Tkatchenko, Nature Reviews Chemistry 4, 347 (2020).
[14] D. Frenkel and B. Smit, Understanding Molecular Simulation, 2nd ed. (Academic Press, Inc., USA, 2001).
[15] M. N. Rosenbluth, AIP Conference Proceedings 690, 22 (2003).
