Predicting quantum-accurate electron densities for DNA with equivariant neural networks

Alex J. Lee
*Department of Chemical and Biological Engineering,
University of New Mexico, Albuquerque, NM 87131, USA.*

Joshua A. Rackers
Center for Computing Research, Sandia National Laboratories, Albuquerque, New Mexico 87123, USA.[*]

William P. Bricker
*Department of Chemical and Biological Engineering,
University of New Mexico, Albuquerque, NM 87131, USA.* [†]

(Dated: March 9, 2022)

One of the fundamental limitations of accurately modeling biomolecules like DNA is the inability to perform quantum chemistry calculations on large molecular structures. We present a machine learning model based on an equivariant Euclidean Neural Network framework to obtain quantum-accurate electron densities for arbitrary DNA structures that are much too large for conventional quantum methods. The model is trained on representative B-DNA base pair steps that capture both base pairing and base stacking interactions. The model produces accurate electron densities for arbitrary B-DNA structures with typical errors of less than 1%. Crucially, the error does not increase with system size, which suggests that the model can extrapolate to large DNA structures with negligible loss of accuracy. The model also generalizes well to other DNA structural motifs such as the A- and Z-DNA forms, despite being trained on only B-DNA configurations. We show that this machine learning electron density model can be used to calculate electrostatic potentials of DNA with quantum accuracy. These electrostatic potentials produce more accurate results compared to classical force fields and do not show the usual deficiencies at short range. Lastly, the model is used to calculate electron densities of several large-scale DNA structures, and we show that the computational scaling for this model is linear.

INTRODUCTION

Quantum molecular modeling is a powerful tool for predicting and understanding the properties of molecular systems from first principles. Despite advances in methodologies and computing power, however, the broad application of quantum modeling to biological systems of interest remains hindered by the steep computational scaling of solving Schrödinger's equation. [1] Large biological macromolecules such as DNA typically consist of thousands to tens of thousands of atoms or more, making them far too large for traditional quantum calculations.

Nevertheless, quantum density functional theory (DFT) calculations have been used in numerous DNA studies. In particular, they have been useful for understanding the energetics of various DNA binding complexes with proteins [2] and metals. [3, 4] Benefits from studying these complexes range from a deeper understanding of fundamental biological processes to the development of technologically advanced biomedical applications. However, due to the high computational cost of DFT, complexes tend to be modeled in a limited capacity, restricted to modeling a few DNA base pairs or including only small portions of the complex. A DFT benchmarking study was able to accurately investigate DNA stacking interactions at the base pair step level. [5] In addition, a recent DFT study was able to model a full turn of A-DNA in a vacuum environment at the hybrid functional level. [6] But given the steep scaling of DFT, calculations on larger DNA structures would be extremely costly if not unfeasible. As such, molecular simulations of full DNA structures typically rely on classical force fields [7–9] or coarse-grained methods. [10–13]

Recently, machine learning approaches have been developed to sidestep the prohibitive scaling of quantum chemistry methods. These approaches aim to construct an alternative model to solving Schrödinger's equation, making them a promising avenue for studying large biological systems that would typically fall outside the range of quantum simulations. To date, machine learning methods have been applied to predict accurate energies [14–16], forces [17, 18], and electron densities [19–23] on mostly small molecules and crystals. For DNA in particular, machine learning has been used to develop a DFT functional that correctly describes charge delocalization in base pairing. [24] But a machine learning model has not yet been developed to predict properties for arbitrary, large-scale DNA structures.

Out of the properties that can be predicted, we focus on the electron density. The central problem to be solved by any machine learning electronic structure method is the ability to train on small fragments without losing ac-

[*] E-mail: jracker@sandia.edu
[†] E-mail: wbricker@unm.edu

curacy when evaluating the model on larger structures. Recent work has shown that learning the electron density offers a promising way to solve this conundrum. [25] Moreover, the electron density is a fundamental property from which useful quantities such as energies, forces, and electrostatic potentials can be derived.

Accurate modeling of large DNA structures provided by quantum calculations would be especially useful in the burgeoning field of DNA nanotechnology, which utilizes a repeated four-way junction motif [26, 27] to program large-scale synthetic DNA assemblies often referred to as “DNA origami.” [28–33] An understanding of how to control these customized, scaffolded DNA structures has led to applications such as biological light-harvesting and energy transfer [34–37], biomedical sensing [38, 39], and molecular-scale memory storage. [40–42] However, the sheer sizes of these DNA origami structures restrict their theoretical and computational study to coarse-grained methods [12] or classical all-atom molecular dynamics, [43–48] which are both dependent on the accuracy of the underlying force fields. In fact, a recent study has shown that commonly used DNA force fields often give incorrect results in describing the folding mechanisms of four-way junctions. [49] It is now easier than ever to sequence and build novel DNA origami structures using available strand routing and structural design software, [50–52] but a straightforward tool to accurately analyze these DNA structures at the molecular and electronic level does not yet exist.

As previously mentioned, the electrostatic potential is a useful property that can be derived from the electron density. The electrostatic potential is a key component for studying binding and solvent interactions and is a particularly important property of DNA due to its negatively charged phosphate groups. [53] The electrostatic potential is strongly linked to the DNA base sequence and has been used to explain the site-specific binding of proteins [10, 11, 13, 54–57] and counterions [58, 59] that are essential for maintaining DNA structure. However, electrostatic potentials are often not calculated from first principles due to system size limitations and may therefore neglect important physics, such as screening and polarization effects. [59, 60]

In this article, we present what is to our knowledge the first machine learning model that produces accurate electron densities for large scale DNA structures without having to run costly–or in many cases, unfeasible–quantum simulations. A key factor to proving the utility of the model is developing an effective training protocol and carefully quantifying and contextualizing its error. Finally, with the groundwork of the model established, we present an application to calculate quantum-accurate electrostatic potentials from the machine-learned densities, which can be useful for characterizing DNA-protein binding and solvent interactions.

METHODS

The training protocol for constructing the DNA machine learning model involves four steps that can be summarized as follows (Fig. 1):

1. The training set is constructed from every possible combination of base pair step in a B-DNA structure (10 combinations total).

2. Molecular dynamics is used to obtain representative snapshots of base pair steps that make up the main training unit for the DNA model.

3. Static quantum calculations are run on these representative base pair steps to produce ground-state electron densities for training the model.

4. The quantum electron densities are used to train a Euclidean neural network, a type of graph convolutional neural network that can understand and exploit the properties of Euclidean symmetry for more efficient learning. Once the network has been trained, it can take as input an arbitrary DNA structure and output an electron density without directly performing a quantum calculation.

Selection of training data for the machine learning model

In order for a machine learning model to make predictions with any accuracy, it must be properly trained. Practically speaking, to make predictions for a large biomolecular structure such as DNA, the training data must be small enough to calculate quickly yet complex enough to capture the interesting physics that undergird the larger structure. In principle, the machine learning model learns generalizable features about atomic environments, such as bonding interactions, that can be thoroughly sampled by the training set. With these considerations, we decided on the base pair step of B-DNA as the fundamental training unit. The base pair step is well-suited for training DNA because it is the smallest unit that captures three key physical interactions: the hydrogen bonds between complementary base pairs (A/T and G/C) that hybridize the two strands, the base stacking interactions between adjacent bases along a strand that stabilize the DNA double helical structure, and the covalent linkages between nucleotide components that form the DNA backbone.

The strategy behind choosing the base pair step as the fundamental training unit is that any arbitrary DNA structure can be broken down into a sequence of overlapping base pair steps. A model that understands the features of every possible base pair step should be able to reasonably reconstruct the electron density for any DNA sequence. This training protocol is validated by

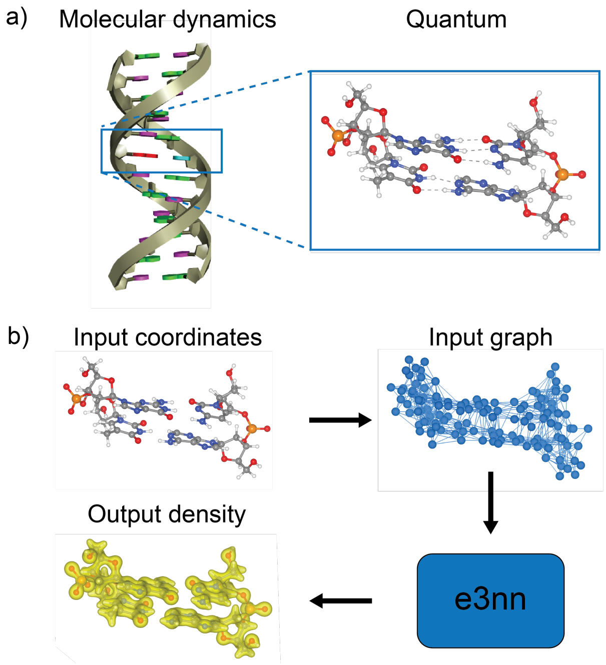

FIG. 1 Schematic for training a machine learning model for DNA. Illustrations were made using Discovery Studio Visualizer [61] and VESTA. [62] (a) Molecular dynamics simulations are run on B-DNA 12-mers. Quantum calculations are performed on representative snapshots extracted from the central two base pairs to obtain training densities. (b) The e3nn model takes atomic coordinates as input, converts them into a graph which is fed through the network, and outputs the electron density for the given input coordinates. The model is trained on all ten combinations of base pair steps that make up a full DNA base sequence.

comparing model predictions to quantum calculation in a test set of B-DNA structures. In the test set, arbitrary structures ranging from two to five base pairs in length were included (referred to as 2-mers, 3-mers, and so on). For 2-, 3-, and 4-mers, every possible combination of base pairs (10, 32, and 136 combinations, respectively) was included. For 5-mers, the sheer number of base pair combinations as well as the size of the structures makes exhaustive sampling costly. Therefore, we selectively sampled the 5-mer space using ten randomly generated base sequences. Table I summarizes the contents of the test set (see the ESI† for details on the 5-mer base sequences which are located in Table S1). We emphasize that the training and test sets are both based off of the B-form of DNA. We later demonstrate the ability of the model to extrapolate to other structural motifs such as A- and Z-DNA.

TABLE I Contents of the B-DNA test set for various base sequence lengths. Combinations refer to unique sequences of base pairs for a given sequence length. For 2, 3, and 4 base pairs, combinations are sampled exhaustively

sequence length	base pair combinations	samples per combination	total samples
2	10	30	300
3	32	10	320
4	136	2	272
5	10	3	30

Molecular dynamics sampling of DNA configurational space

To sample representative geometries of all base pair steps, molecular dynamics (MD) simulations were run with Amber 20 [63] using the BSC1 force field. [9] An initial 12mer B-DNA structure was placed in a periodic truncated octahedral box with a 10 Å buffer and solvated with TIP3P water. [64] Mg^{2+} counterions were added to neutralize the structure, and then an excess of Mg^{2+} and Cl^{-} at about 100 mmol/L were added to the simulation box. Since we are only interested in the dynamics of the central base pair step, restraints were added to keep intact the base pairs at both ends of the DNA.

Prior to MD production runs, structures were first minimized then allowed to heat up from 0 K to 300 K for 40 ps. Production runs were done with the NPT ensemble at 300 K. We ran three 50 ns simulations with different starting trajectories for a total simulation time of 150 ns. Snapshots from these simulations were taken every 2 ps, and all atoms were stripped away apart from the central two base pairs (Fig. 1a). Dangling bonds were capped with hydrogens. To ensure representative sampling of the DNA configurational space, snapshots were processed using K-means clustering with a cluster size of five and a metric of distance-RMSD. The samples closest to the cluster centers were included in the training data. Given that there are four DNA bases (A, C, G, and T) and taking directionality into account, there are a total of ten unique combinations of base pair steps, requiring ten independent 12-mer molecular dynamics runs.

Quantum electron density of DNA configurations

Ab initio quantum calculations were performed on the MD sampled base pair step configurations to obtain the ground-state electron densities used for training the model. Quantum calculations were performed using `psi4` [65] with the DFT hybrid functional `PBE0` [66] and the `aug-cc-pvdz` basis. [67] The outputs from the machine learning model should reflect the same level of theory used to calculate the training data. In principle, the model can be trained on any level of theory.

Typically, electron densities from quantum calculations are constructed from the wave functions of the occupied states. This so-called one-particle density matrix, however, grows as the square of the system size. To keep the size of the density representation linear, we chose to express the density in terms of an atom-centered, auxiliary, "density fitting" basis: [68, 69]

$$\rho(\boldsymbol{r}) = \sum_{i=0}^{N_{atoms}} \sum_{k=0}^{N_{basis}} \sum_{l=0}^{l_{max}} \sum_{m=-l}^{+l} C_{iklm} Y_{lm} e^{-\alpha_{ikl}(\boldsymbol{r}-\boldsymbol{r_i})^2}, \tag{1}$$

where α_{ikl} control the Gaussian widths, Y_{lm} are the spherical harmonics, and C_{iklm} are the coefficients for the auxiliary basis. These coefficients are the data from which the model is trained and are also the outputs of the model. We used the `def2-universal-jfit` auxiliary basis for this study. [69] Expressing the density in this form has been shown to be highly efficient in previous machine learning studies. [14, 20–22]

Choice of machine learning algorithm (`e3nn`)

A key property of molecular systems is that they are equivariant with respect to Euclidean symmetry. Intuitively, what this means is that when a molecule is translated, rotated, or reflected in 3D space, it behaves as an equivalent molecule. The symmetries are thus extended to properties of the molecule. For the sake of neural network training efficiency, it is advantageous to use an architecture that can understand and exploit these symmetries. As such, we use the `e3nn` machine learning framework, which is structured as a graph convolutional neural network that has equivariance in three dimensions built in. [25, 70, 71] Because `e3nn` understands that a rotated form of a molecule is the same molecule, it learns much faster than a model that does not account for this symmetry. In fact, it has been shown that `e3nn` can reduce the amount of training data needed by a factor of 1000 compared to models without built-in equivariance. [18]

For the task of learning electron densities represented in an auxiliary basis, equivariance is an essential property. Each C_{iklm} coefficient corresponds to a different spherical harmonic function of degree l and order m. This means that the coefficients for all functions with $l > 0$ will change under rotation or reflection of the molecule. Equivariance is required to perform this transformation correctly. The `e3nn` framework implements equivariance by representing learned features in the hidden layers of the network as combinations of irreducible representations of 3D space. These features can themselves be interpreted as spherical harmonics. The networks trained in this work use learned features in the hidden layers up to $l_{max} = 3$.

A detailed description of `e3nn` can be found in the literature [25, 70, 71] or online: `https://e3nn.org/` . A brief, non-technical explanation of the model is given here (Fig. 1b). The `e3nn` model is initialized with a structure's atomic coordinates. The coordinates are converted into a three-dimensional graph that gets passed as the input layer to the neural network. Each node in the graph represents an atomic position that marks the center of a gated radial convolution that learns geometric features (such as bonding interactions) about its environment. The learned features are stored as combinations of irreducible representations in the hidden layers of the network. After passing through the hidden layers, the model produces as output the coefficients C_{iklm} that make up the 3D charge density. Because `e3nn` has equivariance built in, a molecule that is translated, rotated, or reflected in space will produce the same charge density–in fact, a translated, rotated, or reflected form of the original charge density. Prior to training the `e3nn` model, network hyperparameters were tuned using a 2-mer validation set independent from the training and test sets. The network hyperparameters can be found in the ESI†.

RESULTS

Assessing the accuracy of the machine learning model

We first show that the training protocol described in the Methods is effective and produces an accurate model for predicting DNA electron densities. Specifically, we test if a machine learning model trained only on base pair steps can make accurate predictions for arbitrary DNA structures. We evaluate the fit of the model by directly comparing to a test set of electron densities calculated from *ab initio* quantum calculations. Since quantum calculations are restricted by system size, the test set was limited to DNA structures with sequence lengths ranging from two to five base pairs.

A quantitative measure of the density prediction error can be given by the equations:

$$\epsilon_{\rho_{ML}}(\%) = 100 \times \frac{\int \mathrm{d}\boldsymbol{r} |\rho_{QM_{projected}}(\boldsymbol{r}) - \rho_{ML}(\boldsymbol{r})|}{\int \mathrm{d}\boldsymbol{r} \rho_{QM_{projected}}(\boldsymbol{r})}, \tag{2}$$

or

$$\epsilon_{\rho_{true}}(\%) = 100 \times \frac{\int \mathrm{d}\boldsymbol{r} |\rho_{QM_{true}}(\boldsymbol{r}) - \rho_{ML}(\boldsymbol{r})|}{\int \mathrm{d}\boldsymbol{r} \rho_{QM_{true}}(\boldsymbol{r})}. \tag{3}$$

ρ_{QM} is the density calculated from quantum mechanics, where the subscript indicates whether the density is in its true form or projected form in the "density fitting" basis (Eq. 1). ρ_{ML} is the machine learning predicted density, which is always expressed in the density fitting basis and so does not need a subscript. $\epsilon_{\rho_{ML}}$ is the error that comes from fitting the trained model, whereas $\epsilon_{\rho_{true}}$ includes the "density fitting" contribution to the error and so will always be higher than $\epsilon_{\rho_{ML}}$. For all tests with the `def2-universal-jfit` basis, the density fitting error was essentially constant, at around 0.73%.

The learning curve for training the model with increasing numbers of training samples is plotted in Fig. 2a. The reported error $\epsilon_{\rho_{ML}}$ is calculated against the test set of 2-mers. For a functional machine learning model that meaningfully learns from training data, this learning curve must be linear on a log-log scale. [72] The linear trend of our data proves that our proposed training protocol produces a functional machine learning model. The model trained on 4000 samples predicts 2-mer test set densities with an average error of $\epsilon_{\rho_{ML}} = 0.62\%$. Including the "density fitting" error, the total prediction error averages to $\epsilon_{\rho_{true}} = 1.00\%$.

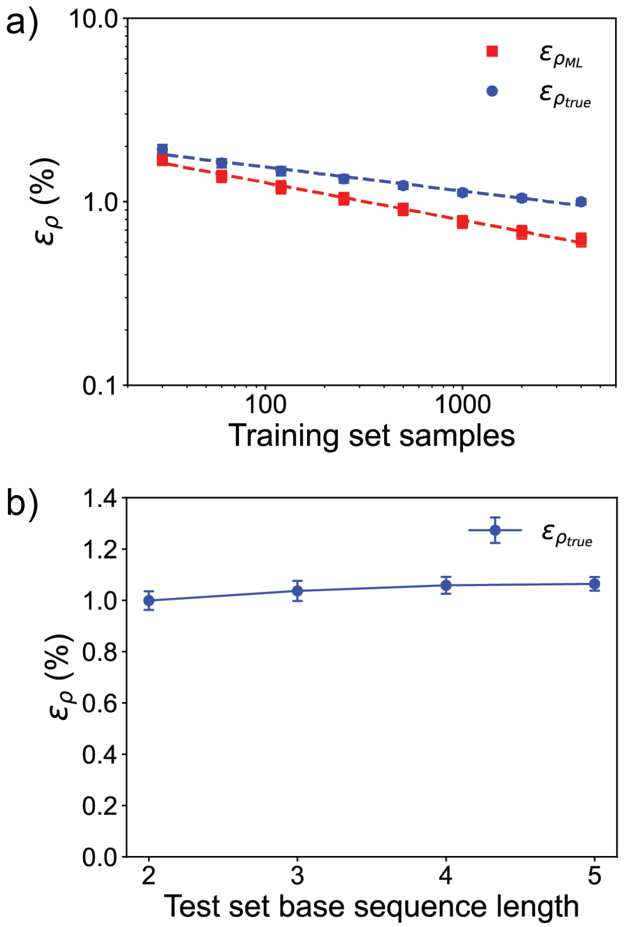

FIG. 2 Training and test results for the DNA machine learning model. (a) Learning curves for increasing numbers of training samples. Dashed lines show the linear regression curves. The errors are calculated against the test set of 2-mers. (b) Error $\epsilon_{\rho_{true}}$ with respect to increasing base sequence length in the test set. The same model, trained on 4000 samples, was used in each case.

Recall that the training set contains only configurations with two base pairs. For the model to predict accurate densities for arbitrary DNA structures longer than two base pairs, the error must not significantly increase for longer base sequences. Figure 2b shows the model prediction error against the complete test set with DNA sequence lengths up to five base pairs. Because it is computationally burdensome to project quantum mechanical densities on to the density fitting basis for the larger systems in the test set, we compare all results to the true quantum mechanical densities using $\epsilon_{\rho_{true}}$ in Eq. 3, noting that the machine learning error $\epsilon_{\rho_{ML}}$ would be lower as in Fig. 2a. The error increases slightly from the test set of 2-mers but flattens out as DNA sequence length is increased, reaching an average error of $\epsilon_{\rho_{true}} = 1.06\%$ for the test set of 5-mers.

Machine-learned quantum electrostatic potentials

As described in the Introduction, the electrostatic potential is a property that can be derived from the electron density and has been used to explain the site-specific binding of proteins and counterions. [10, 11, 13, 54–57] To assess the accuracy of model-calculated DNA electrostatic potentials, single base pair structures of both the A/T and G/C base pairs were investigated, as the electrostatic potentials for these base pairs are well-studied. [13, 73, 74] The major and minor grooves carry electronic signatures that distinguish the base pairs and make them uniquely recognizable to interacting molecules. On the major groove side, the G/C pair shows a strong polarity across the hydrogen bond (Fig. 3b) whereas the A/T base pair has a positive amine group in its center, resulting in an overall neutral to weakly negative potential (Fig. 3a). On the minor groove side, while both base pairs have negative potentials, the A/T base pair is more strongly negative since it lacks the positive amine group.

For a more quantitative assessment, the 2-mer test set was used to calculate the root-mean-square-deviations (RMSDs) of machine-learned electrostatic potentials compared to reference *ab initio* quantum calculations at varying isovalues of the density (Fig. 3c). During this assessment, the averages of the electrostatic potentials were aligned for better comparison. By changing the isovalue, the distance of the potential surface from the molecule changes as well, which is quantified here as the average distance of the potential from the nearest atom. Smaller or larger isovalues of the density result in a potential surface farther away from or closer to the molecule, respectively. For comparison, the RMSDs for the electrostatic portion of the Amber BSC1 classical force field, [9] which consists of parameterized partial atomic charges, are shown alongside the machine-learned potentials. At longer distances from the molecule, the machine learning and classical potentials agree closely. Conversely, at shorter distances the model performs much better as the error in the classical potential increases sharply. This behavior is expected since at short distances, the classical potential enters the region of the electron cloud, which is approximated here using the van der Waals radius of hy-

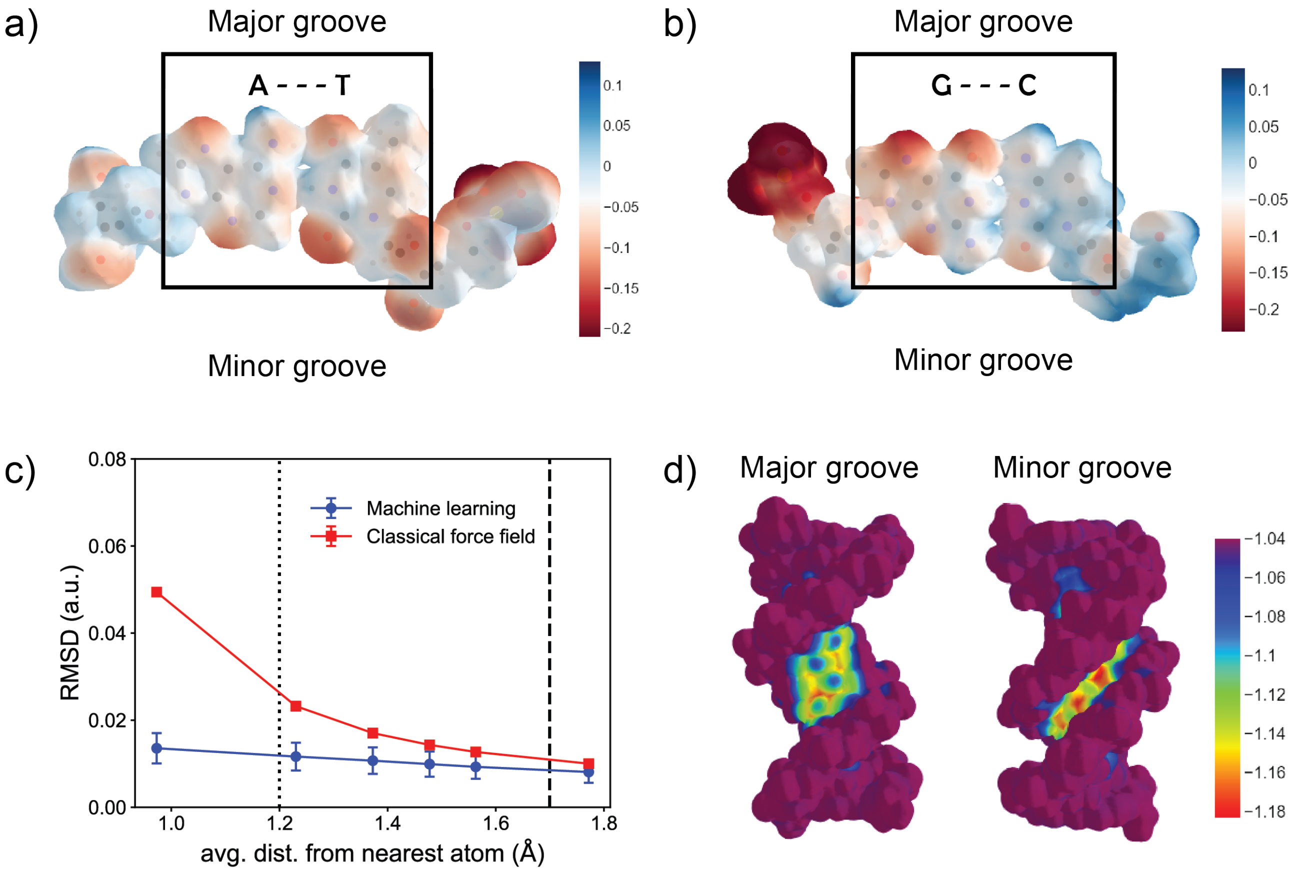

FIG. 3 Electrostatic potentials derived from the machine learning density model. Electrostatic potentials for the (a) A/T and (b) G/C base pairs. Units for the potential are given in a.u. The dark red portions at the ends of the structures are the negatively charged phosphate groups. (c) RMSDs against quantum reference calculations on the 2-mer test set for the machine learning electrostatic potential and the classical BSC1 force field. [9] To get a sense of the range of interaction, the van der Waals radius for hydrogen is plotted at $r = 1.2$ Å (dotted line), and one half of the distance for DNA base stacking is plotted at $r = 1.7$ Å (dashed line). (d) Machine-learned electrostatic potentials on the major and minor groove sides of a DNA structure with an A-tract (PDB code: 264d [75]). The isovalue of the density is set such that the potential is calculated at an average distance of 1.7 Å from the nearest atom. Electrostatic potential plots were made using Plotly. [76]

drogen at $r = 1.2$ Å. Classical force fields do not contain direct information about the electron density distribution and only approximate it with partial atomic charges, so they tend to show deficiencies at short-ranges. [77] Since machine-learned electrostatic potentials are calculated directly from electron densities, they should produce accurate short- and long-range interactions, as reflected by the stable RMSD curve in Fig. 3c.

Moving beyond the test set, we demonstrate that the model can produce a machine-learned electrostatic potential for a full size DNA structure. Figure 3d shows a machine-learned electrostatic potential for a DNA structure with an A-tract (PDB code: 264d [75]). A-tracts have received considerable attention due to their unique structural properties, including an unusually narrow minor groove with an enhanced negative electrostatic potential. Both the shape and the negativity of the potential are thought to enable site-specific binding of certain proteins. [10, 13, 56] Although most computational studies on DNA A-tracts were carried out in solvent environments, the machine-learned electrostatic potential for DNA in a vacuum environment was also able to capture the enhanced negativity in the narrow minor groove (Fig. 3d).

DISCUSSION

Contextualizing the machine learning error

In the Results, we reported that the model trained on 4000 samples gave a density prediction error of $\epsilon_{\rho_{ML}} = 0.62\%$ on the 2-mer test set (Fig. 2a). To get a better sense of what this error means, we compare this value to other density prediction errors in the literature. We caution that an exact comparison is difficult to make due to differences in methodologies, neural network architectures, and types of systems studied. For example, in the study that reported the lowest error of 0.3%, the test systems were simple dimers that did not exceed 25 atoms. By comparison, a structure in our DNA 2-mer test set contains around 126 atoms. Given that reported errors in the literature ranged from 0.3% to 2.5%, [19–23] the accuracy of our DNA machine learning model is consistent with the most accurate machine learning density studies.

Another way to contextualize the density prediction error is to compare it to the error from an isolated atom model in which the density is obtained as a superposition of non-interacting, isolated atoms. Representative electron densities produced by the machine learning and isolated atom models (Fig. 4a) for a randomly selected 4-mer in the test set are shown alongside density differences with the true *ab initio* densities (Fig. 4b). The true density prediction error in the machine learning model is $\epsilon_{\rho_{true}} = 1.02\%$ compared to $\epsilon_{\rho_{true}} = 9.68\%$ in the isolated atom model. An important point about these figures is that even though the machine learning and isolated atom densities appear very similar at a glance, the differences are actually drastic, judging from the disparity of the true errors and the density difference plots.

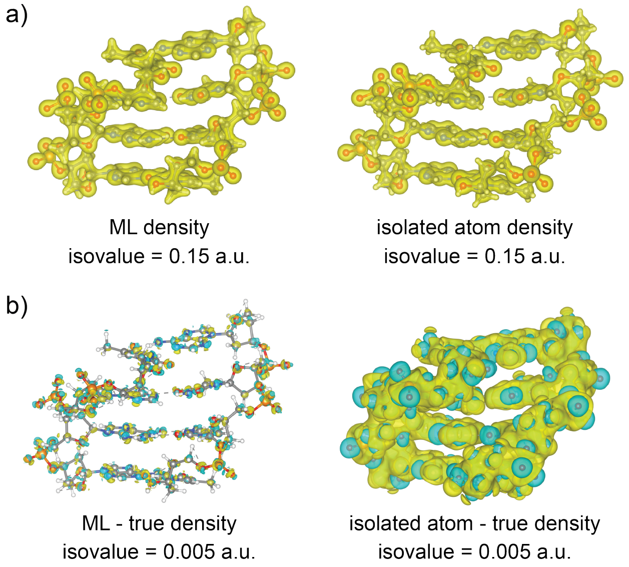

FIG. 4 Comparing machine learning predicted densities to an isolated atom model for a representative test set 4-mer (base sequence: ATCT). Illustrations were made using VESTA. [62] (a) Machine learning predicted and isolated atom densities at an isovalue of 0.15 a.u. The density prediction errors are $\epsilon_{\rho_{true}} = 1.02\%$ and 9.68%, respectively. (b) Density differences with the true quantum density at an isovalue of 0.005 a.u. Yellow and cyan surfaces represent positive and negative values, respectively.

The fact that the machine learning error shows only a slight increase when testing on longer DNA sequences ($\epsilon_{\rho_{true}} = 1.00\%$ for a sequence length of 2 vs. $\epsilon_{\rho_{true}} = 1.06\%$ for a sequence length of 5) suggests that the current machine learning model can make accurate predictions on arbitrary DNA structures of any length. In this sense, it is assumed that the base pair step used for training the model encompasses the majority of the interactions–particularly the base pairing and stacking interactions–that affect a DNA structure's electron density. Still, the slight increase in error may be attributable to longer-range interactions that are not sampled in the two base pair training unit. These longer-range interactions could be accounted for by expanding the test set to include 3-mers, 4-mers, and so on. However, note that both the number of base pair combinations and the size of the training unit will increase for larger DNA sequences, which would greatly increase the computational effort for building the training set.

Another possible source for the small increase in error when testing on longer DNA sequences is that the atomic configurations themselves might not be represented as well by the training set. Recall that the training structures were sampled from the center of a DNA strand (Fig. 1a). Longer DNA chains might adopt configurations that are more similar to those at the edges of DNA, farther from the training data of only central base pairs used for this model. Again, this could be accounted for by including "edge" base pair configurations in addition to the "center" configurations that are currently in the training set.

Of course, the machine learning error could also be improved by simply including more training samples. As long as the learning curve in Fig. 1a stays linear, the model will meaningfully learn when more data is added to the training set. If the current trend holds, doubling the size of the training set from 4000 to 8000 samples would reduce the machine learning density error from $\epsilon_{\rho_{ML}} = 0.62\%$ to 0.52%. Whether this increase in accuracy is worth the extra cost of building up the training set depends on the application at hand.

Additional applications for the machine learning model

Along with predicting electrostatic potentials, we suggest additional applications for the machine learning electron density model. One application is to use the model to develop more accurate classical force fields to simulate processes such as DNA self-assembly that cannot be adequately captured by more general DNA force fields. [33] Partial charges in classical force fields are typically parameterized from quantum calculations using techniques such as RESP, [78] but the underlying quantum calculations are often restricted in size and might not best reflect the macromolecule as a whole. The machine learning density model would allow larger reference calculations on a wider variety of relevant conformations to be used for charge fitting, making it possible to fit partial charges much more accurately. In addition, the model could be used for "on-the-fly" charge reassignment similar to QM/MM approaches. [79] This would enable force fields to more accurately model the diversity of conformations that large DNA molecules can sample.

In addition to parameterizing classical force fields, it is also possible to directly calculate forces from the electron density itself using the Hellmann-Feynman theorem. [80] An advantage for using machine learning models to do this is that long-range forces calculated from density-based machine learning models converge much quicker with respect to training cluster size compared to force-based models. [25] While the `aug-cc-pvdz` basis set used to train the model in this work is not large enough to fulfill the requirements for accurate Hellmann-Feynman forces, [81, 82] the accuracy of the machine learning electron densities suggests that with more complete basis sets the model could be directly applied for *ab initio* molecular dynamics.

Finally, the machine learning model could be used to aid experimentalists in the area of X-ray crystal structure refinement. Refinement techniques typically involve iteratively improving the agreement between experimental reflection data and some structural model. Due to computational limitations, most structural models rely on crude approximations such as the isolated atom model, which we have discussed does not accurately represent a structure's true electron density (Fig. 4). The isolated atom model works reasonably well at intermediate resolutions but starts to show deficiencies for high-resolution images (better than 1.0 Å). [83] Because computational scaling is less of an issue for machine learning models, they can be efficiently used to replace crude models with more sophisticated ones. The need for more sophisticated models in refinement will only become more prevalent as experimental techniques continue to improve and high-resolution images become the standard.

Extrapolating the trained model to other forms of DNA

So far, the density prediction model has only been applied to structures in the canonical B-DNA form, which seems appropriate since the model was trained on data from B-DNA structures. To get a sense for how well the model can extrapolate to the other structural forms of DNA, the model was additionally tested on A- and Z-DNA structures. Experimentally resolved DNA structures from the Protein Data Bank (PDB) were trimmed into 6-mers to make them tractable for quantum reference calculations. The error in the model-predicted densities was quantified using $\epsilon_{\rho_{true}}$ in the same way as before. The results are summarized in Table II. The predicted densities of these structures can be found in the ESI† as Fig. S1.

TABLE II Machine learning model performance on various DNA structural motifs

PDB code	Description	$\epsilon_{\rho_{true}}$ (%)
251d [84]	B-DNA	1.15
1fzx [85]	B-DNA w/ A-tract	1.24
440d [86]	A-DNA	1.73
4fs5 [87]	Z-DNA	1.77

As expected, the lowest error is reported for the B-DNA structures. As a test case, we included a B-DNA structure with an A-tract. With its base sequence of AAAAAA, the A-tract represents a structural extreme for B-DNA. It is, therefore, not surprising that the model error is slightly higher ($\epsilon_{\rho_{true}} = 1.24\%$) although it remains close to that of the regular B-DNA structure ($\epsilon_{\rho_{true}} = 1.15\%$). For the A- and Z-DNA structures, the error increases to $\epsilon_{\rho_{true}} \sim 1.75\%$. Considering that A- and Z-DNA have significantly different structural motifs (Z-DNA features a left-handed helix, for instance), the increase in error is expected. Even so, because of the relatively small errors, we conclude that the model trained only on B-DNA structures can reasonably extrapolate to other forms of DNA. The error is comparable to the $\epsilon_{\rho_{ML}} = 1.5\%$ reported in another extrapolation study, which predicted polypeptide densities from a model trained on a database of biofragments. [20] Of course, the error can be decreased by broadening the training set to include samples from A- and Z-DNA forms. The modest increase in error for A-DNA and Z-DNA suggests that the amount of additional training data necessary to accurately model these structures is likely to be small.

Finally, we showcase the capability of our constructed density prediction model and point the way toward future machine learning studies by presenting model-predicted densities for a handful of extremely large DNA structures. These include the Drew-Dickerson dodecamer [88] (Fig. 5a), a stacked four-way junction [89] (Fig. 5b), a nucleosome core particle [90] (Fig. 5d), and an example of a 2D wireframe DNA origami structure [47, 52] (Fig. 5e). The largest of these structures contains 108,654 electrons, placing it far beyond the scope of traditional quantum methods. The times to compute a density for these structures with the model are presented in Fig. 5c. Plotted on a log-log scale, the slope of the trendline should correspond to the order of scaling $\mathcal{O}(N)$. Based on a slope of $N = 1.17$, the scaling for the machine learning model is essentially linear.

CONCLUSIONS

We constructed a machine learning model based on the equivariant `e3nn` neural network framework that can calculate electron densities for arbitrary DNA structures that well exceed the scope of traditional *ab initio* quantum calculations. The model is trained on B-DNA base pair steps and shows a remarkably low error ($\epsilon_{\rho_{ML}} = 0.62\%$ on the test set of 2-mers) comparable to the most accurate machine learning electron density studies. This error does not significantly increase with

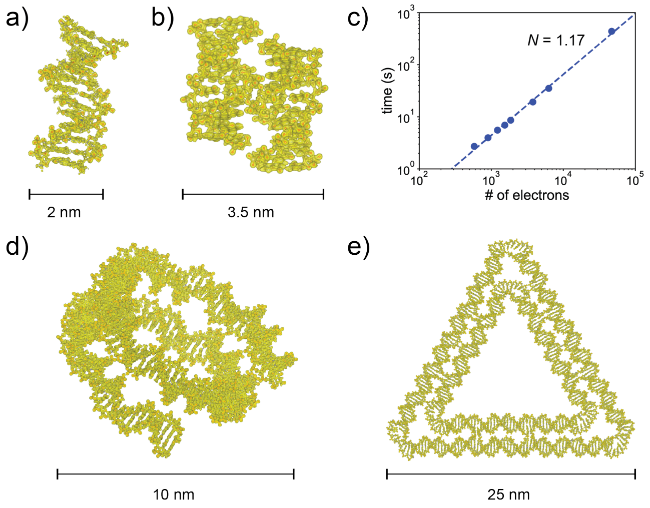

FIG. 5 Machine learning densities for large DNA structures. (a) Drew-Dickerson dodecamer, 758 atoms, 3780 electrons, (PDB code: 4c64 [88]). (b) Stacked four-way junction, 1260 atoms, 6280 electrons (PDB code: 1dcw [89]). (c) Timings for running the machine learning models. The machine learning model scales linearly, $N = 1.17$. (d) Nucleosome core particle, 147 base pairs, 9346 atoms, 46980 electrons (PDB code: 1kx5 [90]). (e) DNA origami triangle [47, 52], ~ 340 base pairs, 21658 atoms, 108654 electrons. Illustrations were made using VESTA. [62]

base sequence length but tends to flatten out, which suggests the model can predict densities for larger DNA structures with negligible loss of accuracy. The model also extrapolates reasonably well to structural forms beyond B-DNA such as the A- and Z-DNA forms, giving errors of around $\epsilon_{\rho_{true}} = 1.75\%$. The learning curves for this model indicate that the error can be systematically decreased by expanding the training set. To demonstrate the power of the model, we computed electron densities for structures ranging from a DNA duplex to a large DNA origami object containing over 100k electrons on timescales of seconds to minutes.

Having a tool that can obtain accurate electron densities for DNA in all its forms opens the door for applications that were previously restricted by computational scaling. In principle, any application that requires an electron density should be able to make use of the model. The model has a direct application for calculating quantum-accurate electrostatic potentials, which have been used to study DNA binding interactions with proteins and counterions. When comparing model-derived electrostatic potentials to classical force fields, the machine learning potentials do not show deficiencies at short range. Lastly, this machine learning electron density model has exciting future uses in molecular dynamics. The model could be used to develop more accurate force fields for classical molecular dynamics, or applied directly to calculate Hellmann-Feynman forces for *ab initio* molecular dynamics.

AUTHOR CONTRIBUTIONS

A.J.L., J.A.R., and W.P.B. conceived of and designed the study. A.J.L. carried out all simulations and analysis. J.A.R. developed the groundwork for the methodology and contributed scripts for analysis. J.A.R. and W.P.B. supervised the project. A.J.L. wrote the manuscript, and all authors edited the manuscript.

CONFLICTS OF INTEREST

There are no conflicts to declare.

DATA AND MATERIALS AVAILABILITY

All code needed to reproduce the experiments and analysis is publicly available on GitHub at `https://github.com/JoshRackers/equivariant_electron_density`.

ACKNOWLEDGMENTS

We thank LDRD ACORN under OSP number A21-0245 and the Sandia Academic Alliance for supporting this work. We thank the UNM Center for Advanced Research Computing, supported in part by the National Science Foundation, for providing the high performance computing and large-scale storage resources used in this work. Additional high performance computing resources, including the GPU-based machine for training and running the machine learning model, were provided by Sandia National Laboratories. Sandia National Laboratories is a multimission laboratory managed and operated by National Technology & Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525. This paper describes objective technical results and analysis. Any subjective views or opinions that might be expressed in the paper do not necessarily represent the views of the U.S. Department of Energy or the United States Government. We thank Mitchell Wood for helpful suggestions.

[1] D. J. Cole and N. D. M. Hine, *J. Phys.: Condens. Matter*, 2016, **28**, 393001.

[2] J. González, I. Baños, I. León, J. Contreras-García, E. J. Cocinero, A. Lesarri, J. A. Fernández and J. Millán, *J.*

Chem. Theory Comput., 2016, **12**, 523–534.

[3] X. W. Liu, J. Li, H. Deng, K. C. Zheng, Z. W. Mao and L. N. Ji, *Dalton Trans.*, 2003, 1352–1359.

[4] G. H. Shahnazari and M. D. Ganji, *Sci. Rep.*, 2021, **11**, 435.

[5] H. Kruse, P. Banáš and J. Šponer, *J. Chem. Theory Comput.*, 2019, **15**, 95–115.

[6] Y. Liu, X. Ren and L. He, *J. Chem. Phys.*, 2019, **151**, 215102.

[7] K. Vanommeslaeghe, E. Hatcher, C. Acharya, S. Kundu, S. Zhong, J. Shim, E. Darian, O. Guvench, P. Lopes, I. Vorobyov and A. D. Mackerell Jr., *J. Comput. Chem.*, 2010, **31**, 671–690.

[8] L. Etheve, J. Martin and R. Lavery, *Nucleic Acids Res.*, 2016, **44**, 9990–10002.

[9] I. Ivani, P. D. Dans, A. Noy, A. Pérez, I. Faustino, A. Hospital, J. Walther, P. Andrio, R. Goñi, A. Balaceanu, G. Portella, F. Battistini, J. L. Gelpí, C. González, M. Vendruscolo, C. A. Laughton, S. A. Harris, D. A. Case and M. Orozco, *Nat. Methods*, 2016, **13**, 55–58.

[10] R. Rohs, S. M. West, A. Sosinsky, P. Liu, R. S. Mann and B. Honig, *Nature*, 2009, **461**, 1248–1253.

[11] R. Rohs, X. Jin, S. M. West, R. Joshi, B. Honig and R. S. Mann, *Annu. Rev. Biochem.*, 2010, **79**, 233–269.

[12] B. E. K. Snodin, F. Randisi, M. Mosayebi, P. Sulc, J. S. Schreck, F. Romano, T. E. Ouldridge, R. Tsukanov, E. Nir, A. A. Louis and J. P. K. Doye, *J. Chem. Phys.*, 2015, **142**, 234901.

[13] T.-P. Chiu, S. Rao, R. S. Mann, B. Honig and R. Rohs, *Nucleic Acids Res.*, 2017, **45**, 12565–12576.

[14] F. Brockherde, L. Vogt, L. Li, M. E. Tuckerman, K. Burke and K.-R. Müller, *Nat. Commun.*, 2017, **8**, 872.

[15] M. Bogojeski, L. Vogt-Maranto, M. E. Tuckerman, K.-R. Müller and K. Burke, *Nat. Commun.*, 2020, **11**, 5223.

[16] R. Nagai, R. Akashi and O. Sugino, *npj Comput. Mater.*, 2020, **6**, 43.

[17] O. T. Unke, S. Chmiela, H. E. Sauceda, M. Gastegger, I. Poltavsky, K. T. Schütt, A. Tkatchenko and K.-R. Müller, *Chem. Rev.*, 2021, **121**, 10142–10186.

[18] S. Batzner, A. Musaelian, L. Sun, M. Geiger, J. P. Mailoa, M. Kornbluth, N. Molinari, T. E. Smidt and B. Kozinsky, *arXiv*, preprint, arXiv:2101.03164, 2021, `https://arxiv.org/abs/2101.03164`.

[19] A. Chandrasekaran, D. Kamal, R. Batra, C. Kim, L. Chen and R. Ramprasad, *npj Comput. Mater.*, 2019, **5**, 22.

[20] A. Fabrizio, A. Grisafi, B. Meyer, M. Ceriotti and C. Corminboeuf, *Chem. Sci.*, 2019, **10**, 9424.

[21] A. Grisafi, A. Fabrizio, B. Meyer, D. M. Wilkins, C. Corminboeuf and M. Ceriotti, *ACS Cent. Sci.*, 2019, **5**, 57–64.

[22] B. Cuevas-Zuviría and L. F. Pacios, *J. Chem. Inf. Model.*, 2021, **61**, 2658–2666.

[23] L. Zepeda-Núñez, Y. Chen, J. Zhang, W. Jia, L. Zhang and L. Lin, *J. Comput. Phys.*, 2021, **443**, 110523.

[24] J. Kirkpatrick, B. McMorrow, D. H. P. Turban, A. L. Gaunt, J. S. Spencer, A. G. D. G. Matthews, A. Obika, L. Thiry, M. Fortunato, D. Pfau, L. R. Castellanos, S. Petersen, A. W. R. Nelson, P. Kohli, P. Mori-Sánchez, D. Hassabis and A. J. Cohen, *Science*, 2021, **374**, 1385–1389.

[25] J. A. Rackers, L. Tecot, M. Geiger and T. E. Smidt, *arXiv*, preprint, arXiv:2201.03726, 2022, `https://arxiv.org/abs/2201.03726`.

[26] N. C. Seeman, *J. Theor. Biol.*, 1982, **99**, 237–47.

[27] D. M. J. Lilley and R. M. Clegg, *Annu. Rev. Biophys. Biomol. Struct.*, 1993, **22**, 299–328.

[28] P. W. K. Rothemund, *Nature*, 2006, **440**, 297–302.

[29] H. Dietz, S. M. Douglas and W. M. Shih, *Science*, 2009, **325**, 725–730.

[30] S. M. Douglas, H. Dietz, T. Liedl, B. Hoegberg, F. Graf and W. M. Shih, *Nature*, 2009, **459**, 414–418.

[31] R. M. Zadegan and M. L. Norton, *Int. J. Mol. Sci.*, 2012, **13**, 7149–7162.

[32] R. Veneziano, S. Ratanalert, K. Zhang, F. Zhang, H. Yan, W. Chiu and M. Bathe, *Science*, 2016, **352**, aaf4388.

[33] E.-C. Wamhoff, J. L. Banal, W. P. Bricker, T. R. Shepherd, M. F. Parsons, R. Veneziano, M. B. Stone, H. Jun, X. Wang and M. Bathe, *Annu. Rev. Biophys.*, 2019, **48**, 395–419.

[34] E. A. Hemmig, C. Creatore, B. Wünsch, L. Hecker, P. Mair, M. A. Parker, S. Emmott, P. Tinnefeld, U. F. Keyser and A. W. Chin, *Nano Lett.*, 2016, **16**, 2369–2374.

[35] F. Nicoli, A. Barth, W. Bae, F. Neukirchinger, A. H. Crevenna, D. C. Lamb and T. Liedl, *ACS Nano*, 2017, **11**, 11264–11272.

[36] E. Boulais, N. P. D. Sawaya, R. Veneziano, A. Andreoni, J. L. Banal, T. Kondo, S. Mandal, S. Lin, G. S. Schlau-Cohen, N. W. Woodbury, H. Yan, A. Aspuru-Guzik and M. Bathe, *Nat. Mater.*, 2018, **17**, 159–166.

[37] S. M. Hart, W. J. Chen, J. L. Banal, W. P. Bricker, A. Dodin, L. Markova, Y. Vyborna, A. P. Willard, R. Häner, M. Bathe and G. S. Schlau-Cohen, *Chem*, 2021, **7**, 752–773.

[38] S. Modi, S. M. G., D. Goswami, G. D. Gupta, S. Mayor and Y. Krishnan, *Nat. Nanotechnol.*, 2009, **4**, 325–330.

[39] R. Veneziano, T. J. Moyer, M. B. Stone, E.-C. Wamhoff, B. J. Read, S. Mukherjee, T. R. Shepherd, J. Das, W. R. Schief, D. J. Irvine and M. Bathe, *Nat. Nanotechnol.*, 2020, **15**, 716–723.

[40] L. Ceze, J. Nivala and K. Strauss, *Nat. Rev. Genet.*, 2019, **20**, 456–466.

[41] J. L. Banal, T. R. Shepherd, J. Berleant, H. Huang, M. Reyes, C. M. Ackerman, P. C. Blainey and M. Bathe, *Nat. Mater.*, 2021, **20**, 1272–1280.

[42] G. D. Dickinson, G. M. Mortuza, W. Clay, L. Piantanida, C. M. Green, C. Watson, E. J. Hayden, T. Andersen, W. Kuang, E. Graugnard, R. Zadegan and W. L. Hughes, *Nat. Commun.*, 2021, **12**, 2371.

[43] J. Yoo and A. Aksimentiev, *Proc. Natl. Acad. Sci. U. S. A.*, 2013, **110**, 20099–20104.

[44] C. Maffeo, J. Yoo and A. Aksimentiev, *Nucleic Acids Res.*, 2016, **44**, 3013–3019.

[45] K. Pan, W. P. Bricker, S. Ratanalert and M. Bathe, *Nucleic Acids Res.*, 2017, **45**, 6284–6298.

[46] H. Jun, T. R. Shepherd, K. Zhang, W. P. Bricker, S. Li, W. Chiu and M. Bathe, *ACS Nano*, 2019, **13**, 2083–2093.

[47] H. Jun, X. Wang, W. P. Bricker and M. Bathe, *Nat. Commun.*, 2019, **10**, 5419.

[48] M. R. Adendorff, G. Q. Tang, D. P. Millar, M. Bathe and W. P. Bricker, *Nucleic Acids Res.*, 2021, **50**, 717–730.

[49] J. Yooab and A. Aksimentiev, *Phys. Chem. Chem. Phys.*, 2018, **20**, 8432–8449.

[50] S. Williams, K. Lund, C. Lin, P. Wonka, S. Lindsay and H. Yan, DNA Computing, Berlin, Heidelberg, 2009, pp. 90–101.

[51] S. M. Douglas, A. H. Marblestone, S. Teerapittayanon, A. Vazquez, G. M. Church and W. M. Shih, *Nucleic Acids Res.*, 2009, **37**, 5001–5006.

[52] H. Jun, X. Wang, M. F. Parsons, W. P. Bricker, T. John, S. Li, S. Jackson, W. Chiu and M. Bathe, *Nucleic Acids Res.*, 2021, **49**, 10265–10274.

[53] P. Ren, J. Chun, D. G. Thomas, M. J. Schnieders, M. Marucho, J. Zhang and N. A. Baker, *Q. Rev. Biophys.*, 2012, **45**, 427–491.

[54] S. Jones, H. P. Shanahan, H. M. Berman and J. M. Thornton, *Nucleic Acids Res.*, 2003, **31**, 7189–7198.

[55] R. Joshi, J. M. Passner, R. Rohs, R. Jain, A. Sosinsky, M. A. Crickmore, V. Jacob, A. K. Aggarwal, B. Honig and R. S. Mann, *Cell*, 2007, **131**, 530–543.

[56] C. Oguey, N. Foloppe and B. Hartmann, *PLoS One*, 2010, **5**, e15931.

[57] Z. Deng, Q. Wang, Z. Liu, M. Zhang, A. C. D. Machado, T.-P. Chiu, C. Feng, Q. Zhang, L. Yu, L. Qi, J. Zheng, X. Wang, X. Huo, X. Qi, X. Li, W. Wu, R. Rohs, Y. Li and Z. Chen, *Nat. Commun.*, 2015, **6**, 7642.

[58] S. Y. Ponomarev, K. M. Thayer and D. L. Beveridge, *Proc. Natl. Acad. Sci. U. S. A.*, 2004, **101**, 14771–14775.

[59] P. Grochowski and J. Trylska, *Biopolymers*, 2008, **89**, 93–113.

[60] T. E. Exner and P. G. Mezey, *J. Phys. Chem. A*, 2002, **106**, 11791–11800.

[61] BIOVIA, Dassault Systèmes, *Discovery Studio Visualizer v21.1.0.20298*, Dassault Systèmes, San Diego, CA, 2017.

[62] K. Momma and F. Izumi, *J. Appl. Crystallogr.*, 2011, **44**, 1272–1276.

[63] D. Case, H. Aktulga, K. Belfon, I. Ben-Shalom, S. Brozell, D. Cerutti, I. T.E. Cheatham, G. Cisneros, V. Cruzeiro, T. Darden, R. Duke, G. Giambasu, M. Gilson, H. Gohlke, A. Goetz, R. Harris, S. Izadi, S. Izmailov, C. Jin, K. Kasavajhala, M. Kaymak, E. King, A. Kovalenko, T. Kurtzman, T. Lee, S. LeGrand, P. Li, C. Lin, J. Liu, T. Luchko, R. Luo, M. Machado, V. Man, M. Manathunga, K. Merz, Y. Miao, O. Mikhailovskii, G. Monard, H. Nguyen, K. O'Hearn, A. Onufriev, F. Pan, S. Pantano, R. Qi, A. Rahnamoun, D. Roe, A. Roitberg, C. Sagui, S. Schott-Verdugo, J. Shen, C. Simmerling, N. Skrynnikov, J. Smith, J. Swails, R. Walker, J. Wang, H. Wei, R. Wolf, X. Wu, Y. Xue, D. York, S. Zhao and P. Kollman, *Amber 2021*, University of California, San Francisco, 2021.

[64] W. L. Jorgensen, J. Chandrasekhar and J. D. Madura, *J. Chem. Phys.*, 1983, **79**, 926.

[65] J. M. Turney, A. C. Simmonett, R. M. Parrish, E. G. Hohenstein, F. Evangelista, J. T. Fermann, B. J. Mintz, L. A. Burns, J. J. Wilke, M. L. Abrams, N. J. Russ, M. L. Leininger, C. L. Janssen, E. T. Seidl, W. D. Allen, H. F. Schaefer, R. A. King, E. F. Valeev, C. D. Sherrill and T. D. Crawford, *WIREs Comput. Mol. Sci.*, 2012, **2**, 556.

[66] J. P. Perdew and M. Ernzerhof, *J. Chem. Phys.*, 1996, **105**, 9982.

[67] T. H. Dunning Jr. and P. J. Hay, in *Methods of Electronic Structure Theory. Modern Theoretical Chemistry*, ed. H. F. Schaefer, Springer, Boston, MA, 1977, vol. 3, ch. Gaussian Basis Sets for Molecular Calculations.

[68] B. P. Pritchard, D. Altarawy, B. Didier, T. D. Gibson and T. L. Windus, *J. Chem. Inf. Model.*, 2019, **59**, 4814–4820.

[69] F. Weigend, *Phys. Chem. Chem. Phys.*, 2006, **8**, 1057–1065.

[70] M. Geiger, T. Smidt, A. Musaelian, B. K. Miller, W. Boomsma, B. Dice, K. Lapchevskyi, M. Weiler, M. Tyszkiewicz, S. Batzner, M. Uhrin, J. Frellsen, N. Jung, S. Sanborn, J. Rackers and M. Bailey, *Euclidean neural networks: e3nn*, Zenodo, 2020, `https://doi.org/10.5281/zenodo.5292912`.

[71] T. E. Smidt, *Trends Chem.*, 2021, **3**, 82–85.

[72] A. S. Christensen and O. A. von Lilienfeld, *Mach. Learn.: Sci. Technol.*, 2020, **1**, 045018.

[73] B. Jayaram, K. A. Sharp and B. Honig, *Biopolymers*, 1989, **28**, 975–93.

[74] R. C. Harris, T. Mackoy, A. C. D. Machado, D. Xu, R. Rohs and M. O. Fenley, *RSC Biomol. Sci.*, 2012, **2**, 53–80.

[75] M. C. Vega, I. G. Sáez, J. Aymamí, R. Eritja, G. A. Van der Marel, J. H. Van Boom, A. Rich and M. Coll, *Eur. J. Biochem.*, 1994, **222**, 721–726.

[76] Plotly Technologies Inc., *Collaborative data science*, Plotly Technologies Inc., Montreal, QC, 2015, `https://plot.ly`.

[77] M. J. Van Vleet, A. J. Misquitta, A. J. Stone and J. R. Schmidt, *J. Chem. Theory Comput.*, 2016, **12**, 3851–3870.

[78] C. I. Bayly, P. Cieplak, W. Cornell and P. A. Kollman, *J. Phys. Chem.*, 1993, **97**, 10269–10280.

[79] A. Laio, J. VandeVondele and U. Rothlisberger, *J. Phys. Chem. B*, 2002, **106**, 7300–7307.

[80] R. P. Feynman, *Phys. Rev.*, 1939, **56**, 340–343.

[81] J. Rico, R. Lopez, I. Ema and G. Ramirez, *Int. J. Quantum Chem.*, 2004, **100**, 221 – 230.

[82] J. Rico, R. Lopez, I. Ema and G. Ramirez, *J Comput. Chem.*, 2007, **28**, 748–58.

[83] P. V. Afonine, R. W. Grosse-Kunstleve, P. D. Adams, V. Y. Lunin and A. Urzhumtsev, *Acta Crystallogr.*, 2007, **D63**, 1194–1197.

[84] M. C. Wahl, S. T. Rao and M. Sundaralingam, *Biophys. J.*, 1996, **70**, 2857–2866.

[85] D. MacDonald, K. Herbert, X. Zhang, T. Polgruto and P. Lu, *J. Mol. Biol.*, 2001, **306**, 1081–1098.

[86] Y.-G. Gao, H. Robinson and A. H.-J. Wang, *Eur. J. Biochem.*, 1999, **261**, 413–420.

[87] T. Chatake and T. Sunami, *J. Inorg. Biochem.*, 2013, **124**, 15–25.

[88] L. Lercher, M. A. McDonough, A. H. El-Sagheer, A. Thalhammer, S. Kriaucionis, T. Brown and C. J. Schofield, *Chem. Commun.*, 2014, **50**, 1794–1796.

[89] B. F. Eichman, J. M. Vargason, B. H. M. Mooers and P. S. Ho, *Proc. Natl. Acad. Sci. U. S. A.*, 2000, **97**, 3971–3976.

[90] C. A. Davey, D. F. Sargent, K. Luger, A. W. Maeder and T. J. Richmond, *J. Mol. Biol.*, 2002, **319**, 1097–1113.