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.

SAND2022-3569C

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

ALEX J. LEE, JOSHUA RACKERS, WILLIAM BRICKER

DEPARTMENT OF CHEMICAL & BIOLOGICAL ENGINEERING

UNIVERSITY OF NEW MEXICO

ACS SPRING 2022, SAN DIEGO



CHEMICAL & BIOLOGICAL ENGINEERING

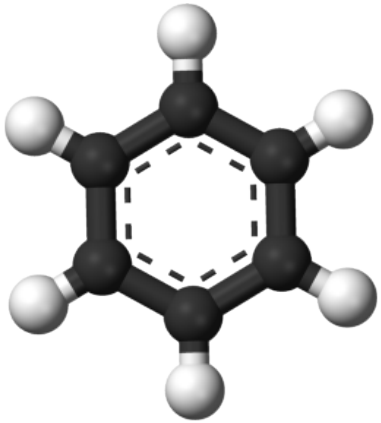


Sandia National Laboratories

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.

MOTIVATION

Quantum molecular modeling can predict a system's properties from first principles.



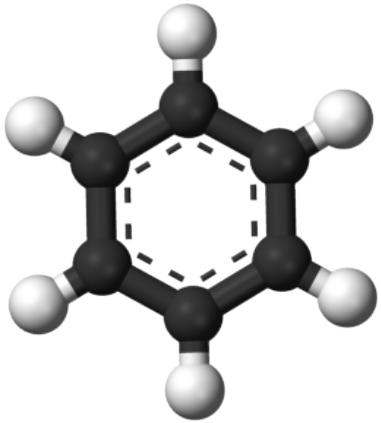
input →

$$\hat{H}\Psi = E\Psi$$

Atomic coordinates (x, y, z)

MOTIVATION

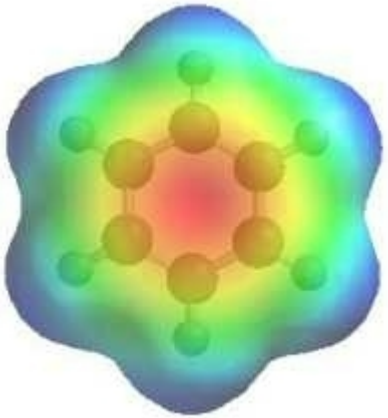
Quantum molecular modeling can predict a system's properties from first principles.



input

$$\hat{H}\Psi = E\Psi$$

output

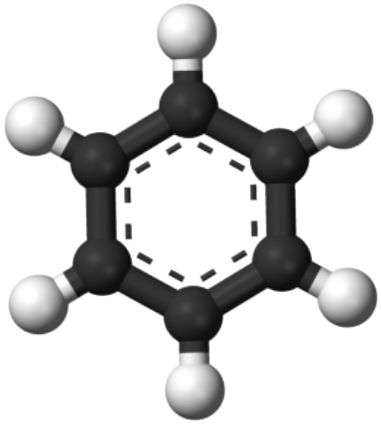


Atomic coordinates (x, y, z)

Density ρ

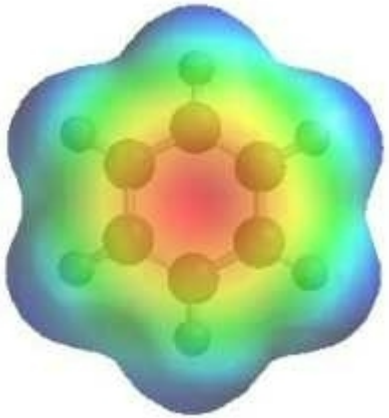
MOTIVATION

Quantum molecular modeling can predict a system's properties from first principles.



Atomic coordinates (x, y, z)

input → $\hat{H}\Psi = E\Psi$ → output

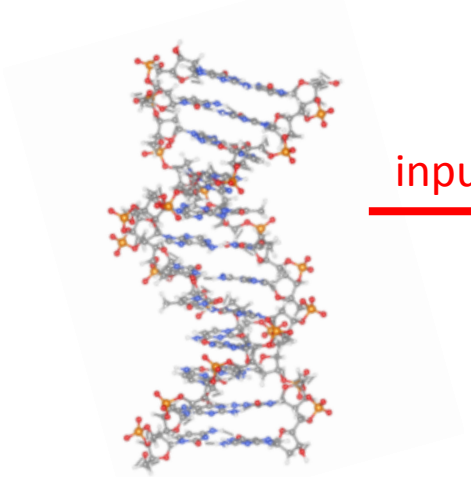


Density ρ

Computational scaling	
Density functional theory	(^{3})
Hybrid functionals	(^{4})
MP2 and beyond	(^{5} to ^{7})

MOTIVATION

Quantum molecular modeling can predict a system's properties from first principles.



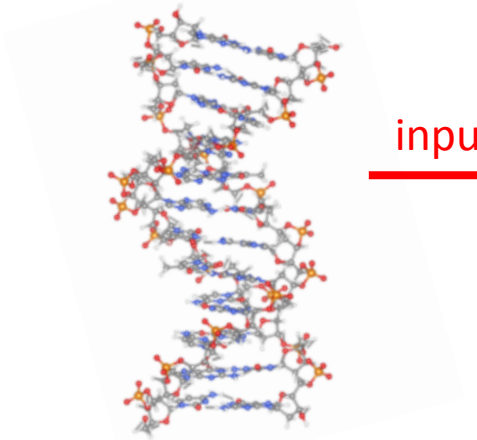
Atomic coordinates (x, y, z)

input → $\hat{H}\Psi = E\Psi$

Computational scaling	
Density functional theory	$(^{3})$
Hybrid functionals	$(^{4})$
MP2 and beyond	$(^{5}\text{ to }^{7})$

MOTIVATION

Quantum molecular modeling can predict a system's properties from first principles.



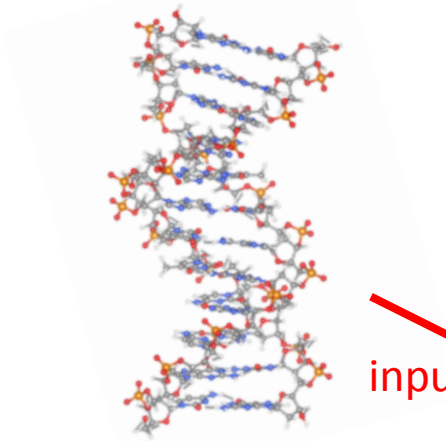
Atomic coordinates (x, y, z)

input → $\hat{H}\Psi = E\Psi$ → ~~output~~

Computational scaling	
Density functional theory	$(^{3})$
Hybrid functionals	$(^{4})$
MP2 and beyond	$(^{5}\text{ to }^{7})$

MOTIVATION

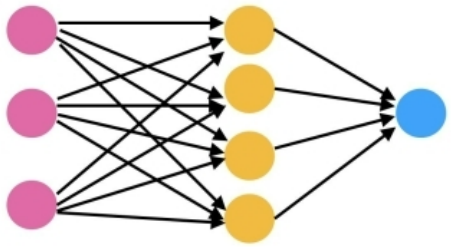
Quantum molecular modeling can predict a system's properties from first principles.



Atomic coordinates (x, y, z)

input

$$\hat{H}\Psi = E\Psi$$

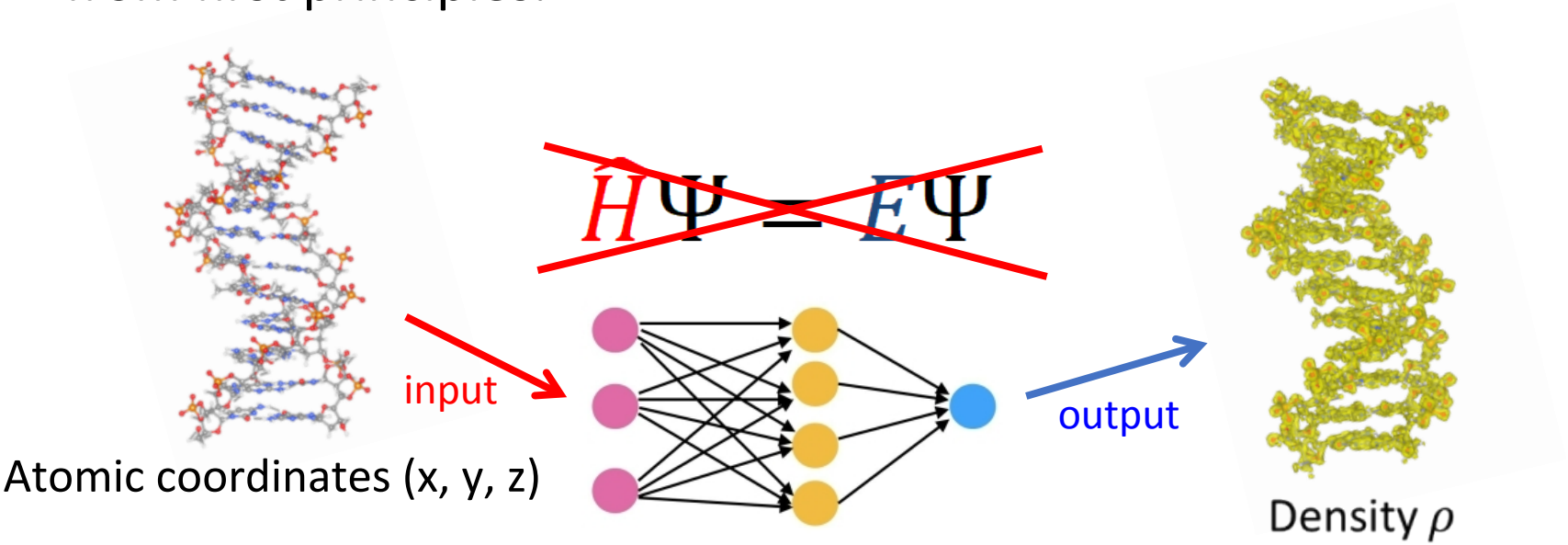


Computational scaling

Density functional theory	$(^{3})$
Hybrid functionals	$(^{4})$
MP2 and beyond	$(^{5}\text{ to }^{7})$

MOTIVATION

Quantum molecular modeling can predict a system's properties from first principles.



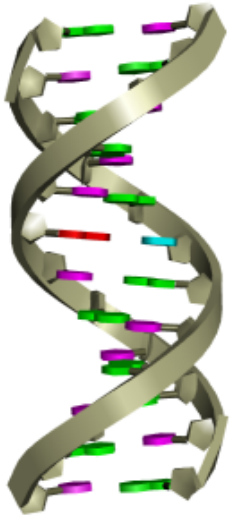
Computational scaling

Density functional theory	$O(N^3)$
Hybrid functionals	$O(N^4)$
MP2 and beyond	$O(N^5)$ to $O(N^7)$

METHODS

Training a DNA machine learning model

Molecular dynamics

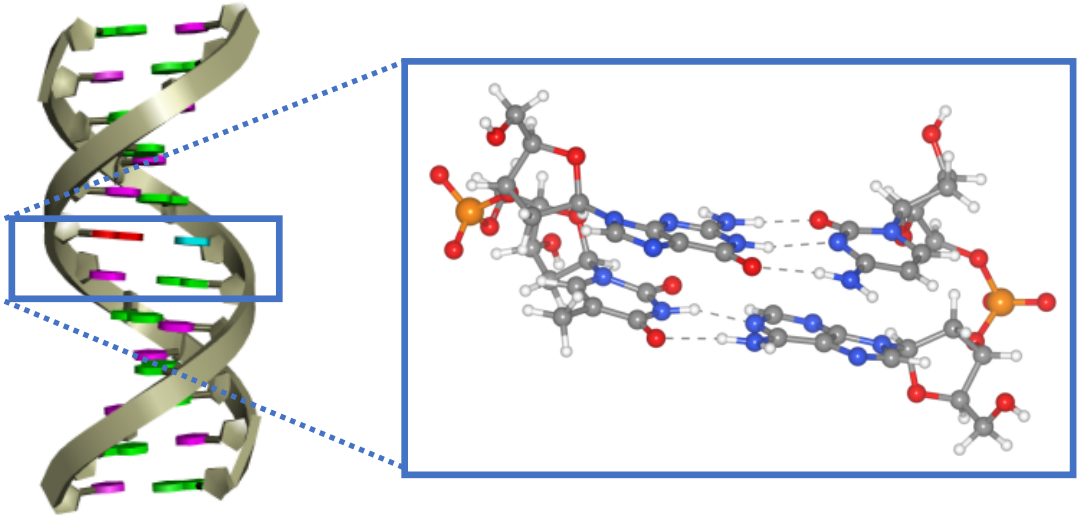


METHODS

Training a DNA machine learning model

Molecular dynamics

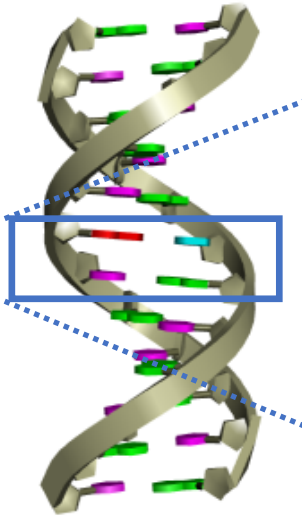
Quantum DFT



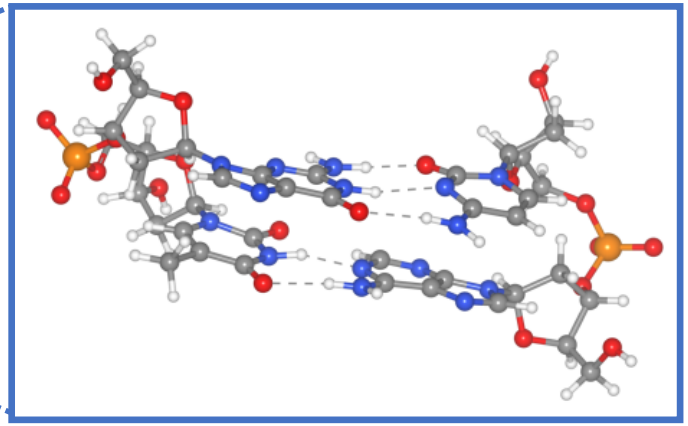
METHODS

Training a DNA machine learning model

Molecular dynamics



Quantum DFT



Training density

$$\rho(\mathbf{r}) = \sum_K C_K \phi_K$$

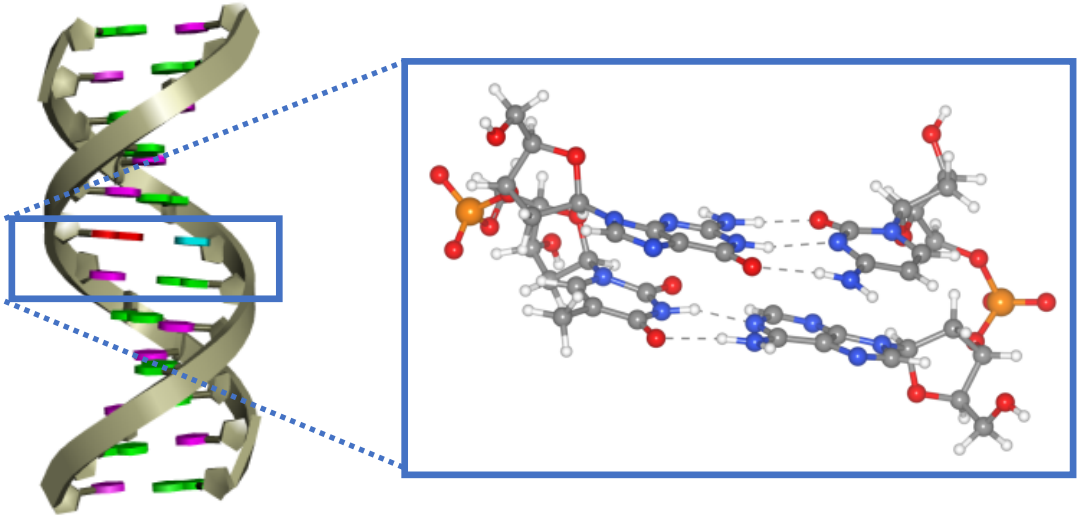
METHODS

Training a DNA machine learning model

Molecular dynamics

Quantum DFT

Training density



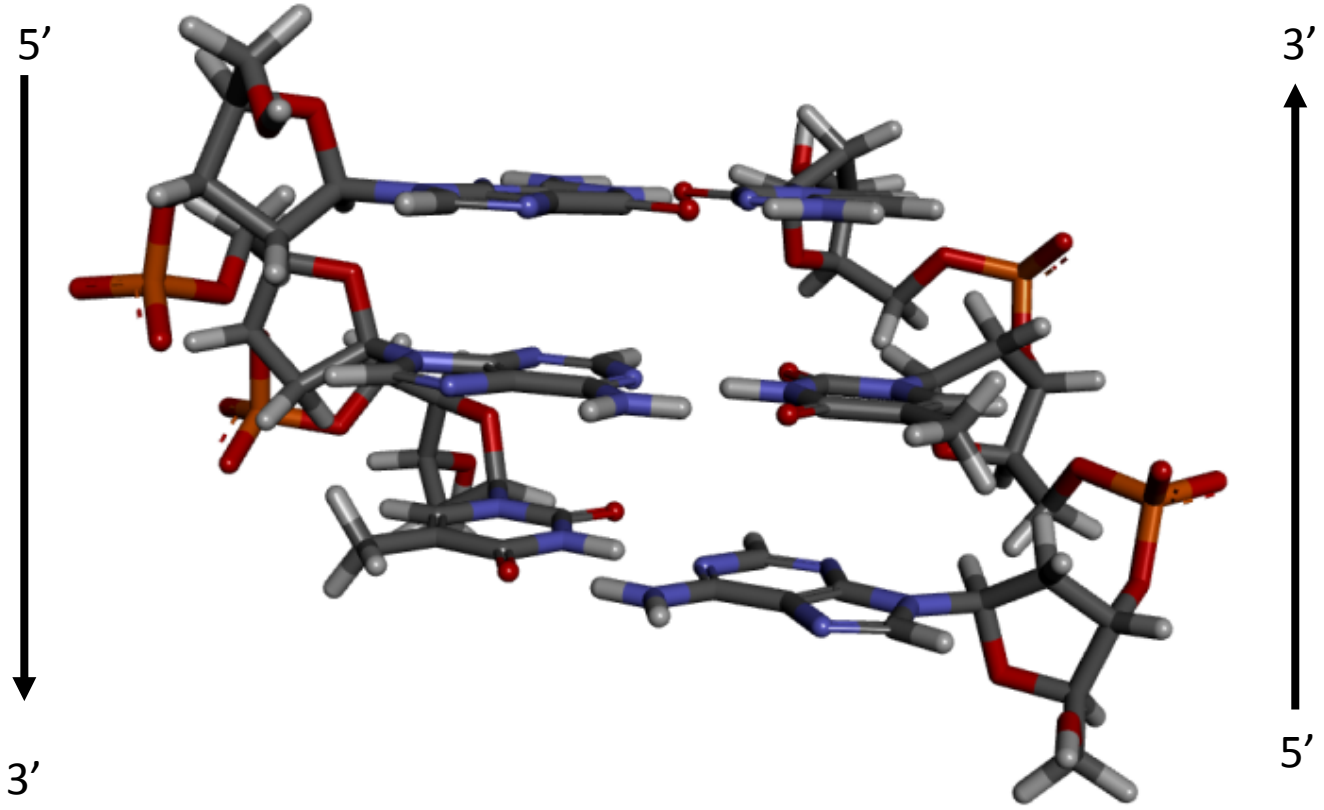
$$\rho(\mathbf{r}) = \sum_K C_K \phi_K$$

- Configurations sampled from molecular dynamics
- Smallest meaningful unit – base pair step
- Machine learning target is the electron density $\rho(\mathbf{r})$

METHODS

Training a DNA machine learning model

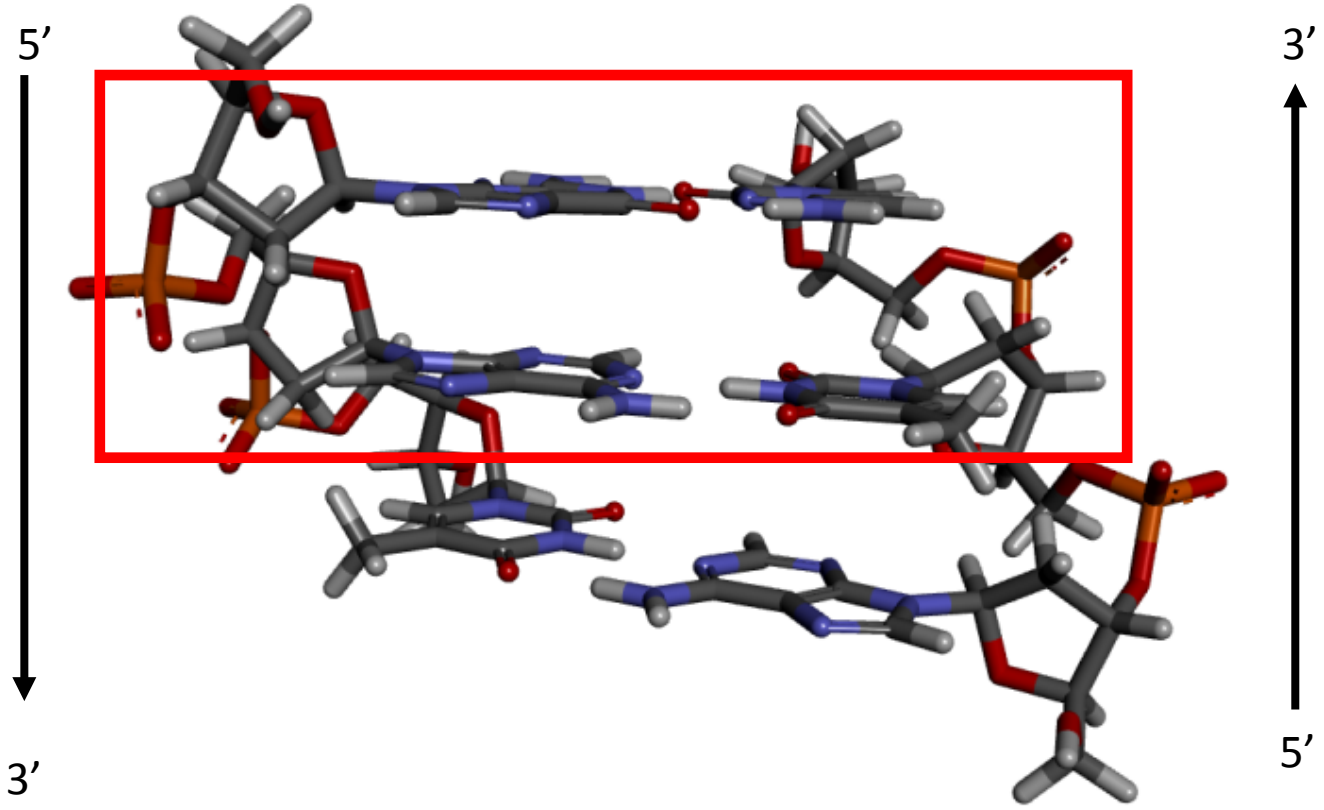
5’ GAT 3’
3’ CTA 5’



METHODS

Training a DNA machine learning model

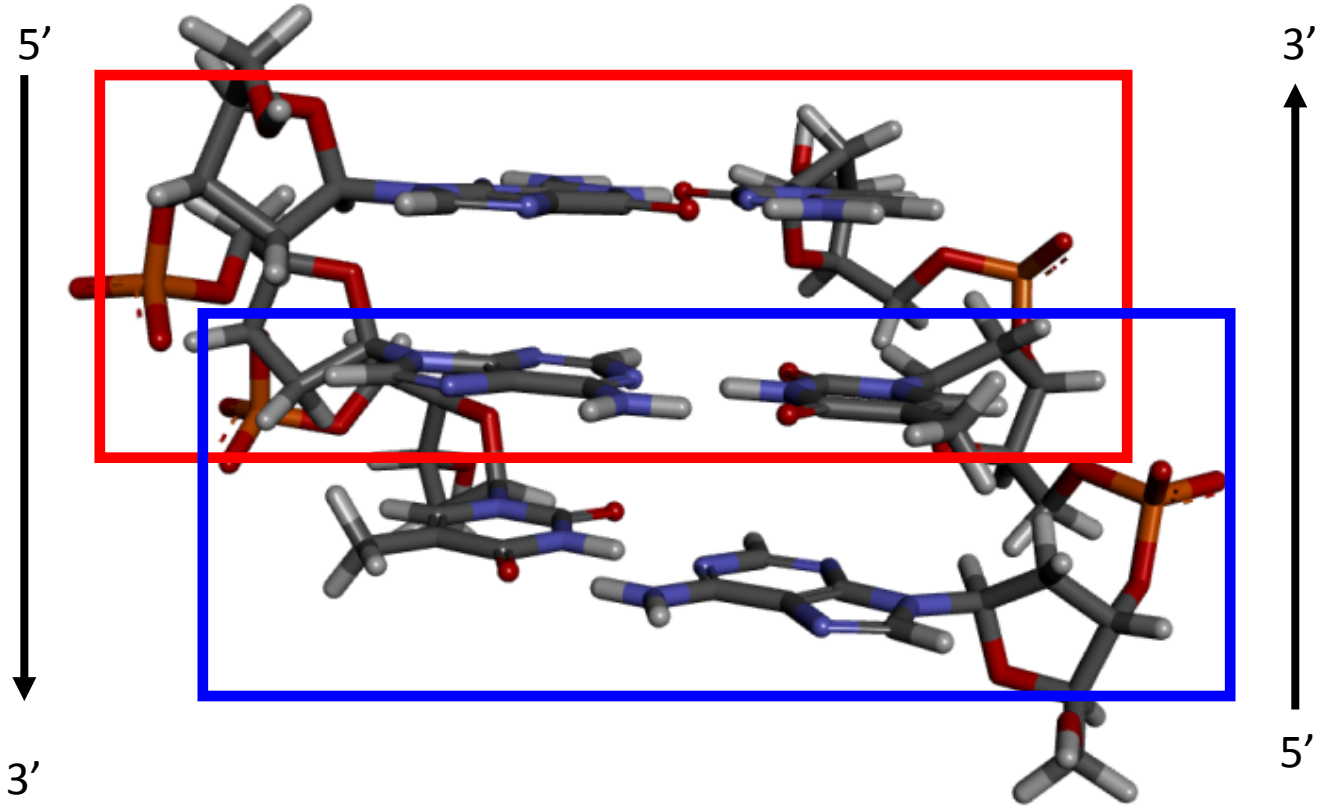
5’ GAT 3’
3’ CTA 5’



METHODS

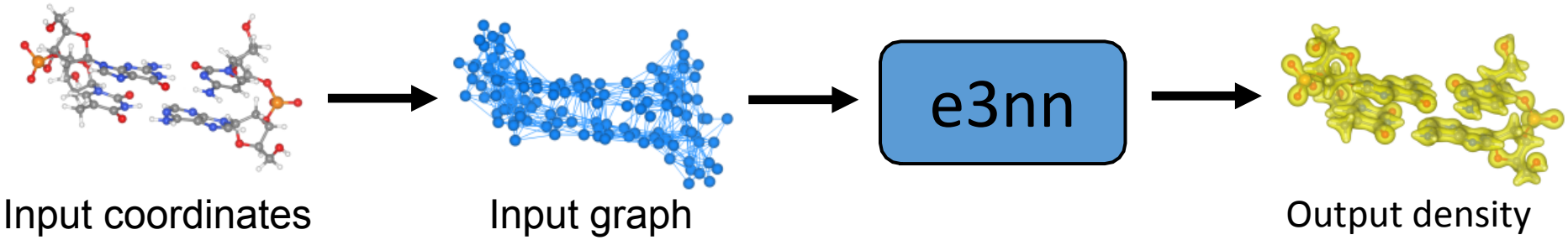
Training a DNA machine learning model

5’ G A T 3’
3’ C T A 5’



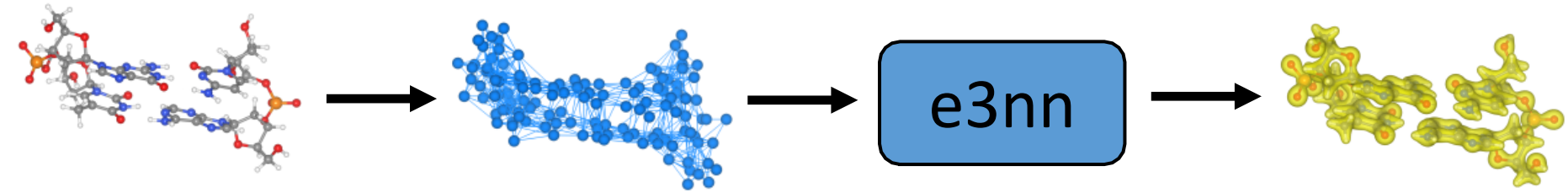
METHODS

e3nn is a type of **graph convolutional neural network** that is equivariant to translations, rotations, and reflections.

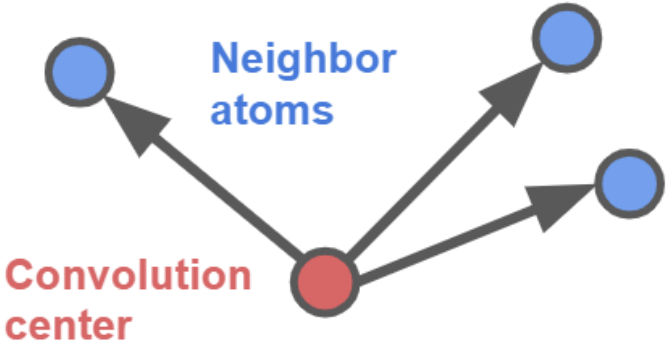


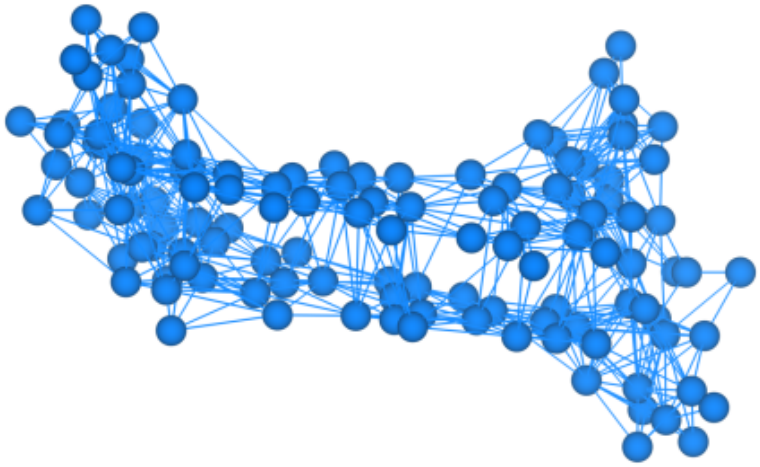
METHODS

e3nn is a type of **graph convolutional neural network** that is equivariant to translations, rotations, and reflections.



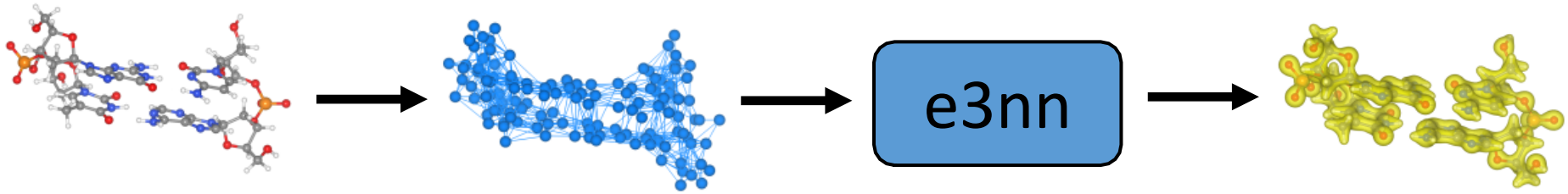
e3nn learns geometric features such as bonding interactions about its environment.



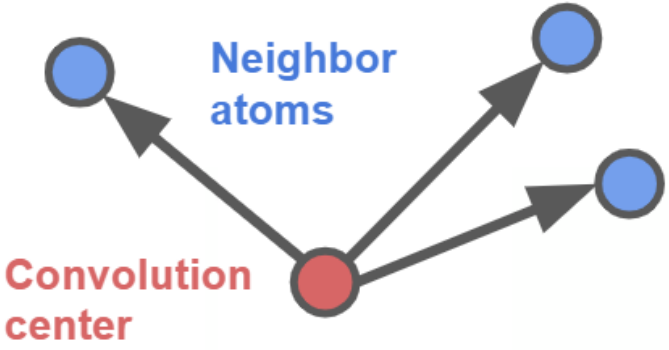


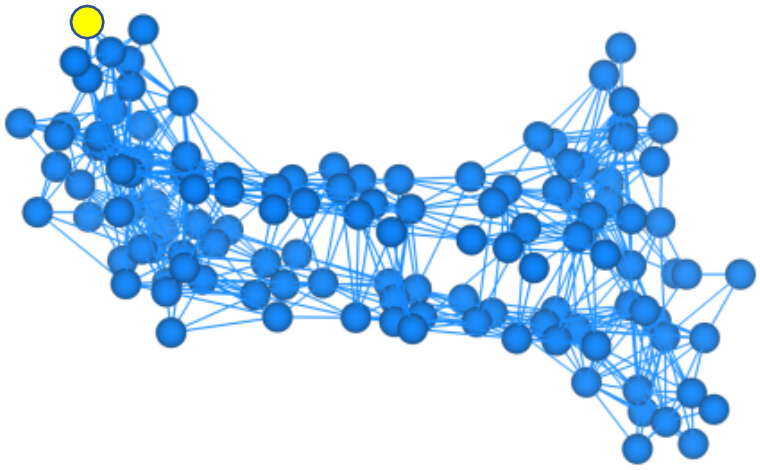
METHODS

e3nn is a type of **graph convolutional neural network** that is equivariant to translations, rotations, and reflections.



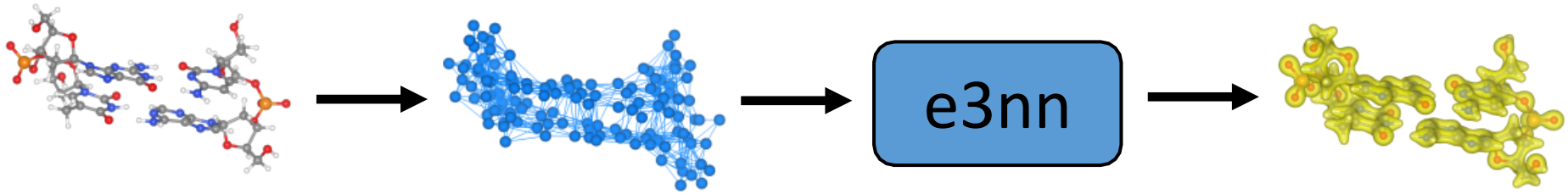
e3nn learns geometric features such as bonding interactions about its environment.



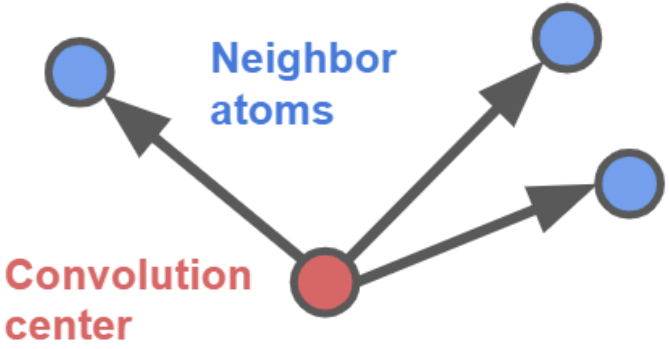


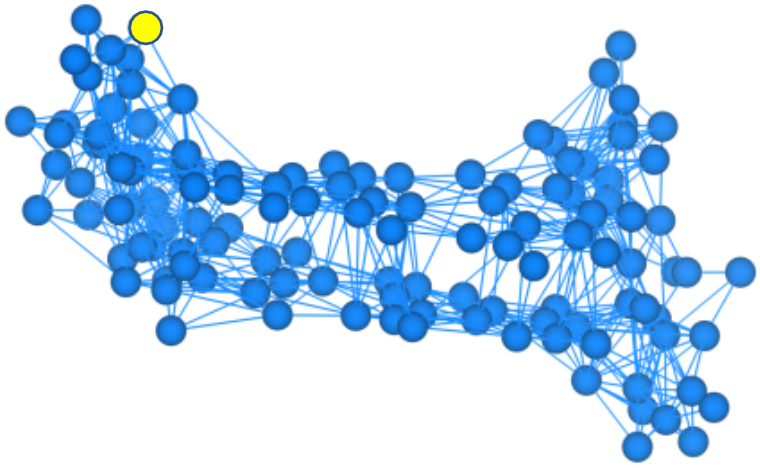
METHODS

e3nn is a type of **graph convolutional neural network** that is equivariant to translations, rotations, and reflections.



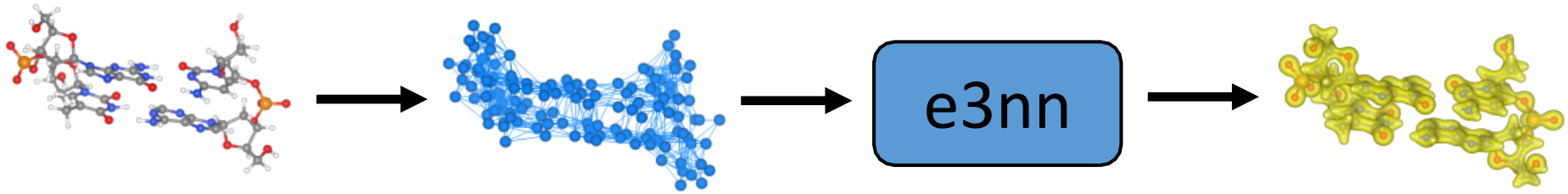
e3nn learns geometric features such as bonding interactions about its environment.



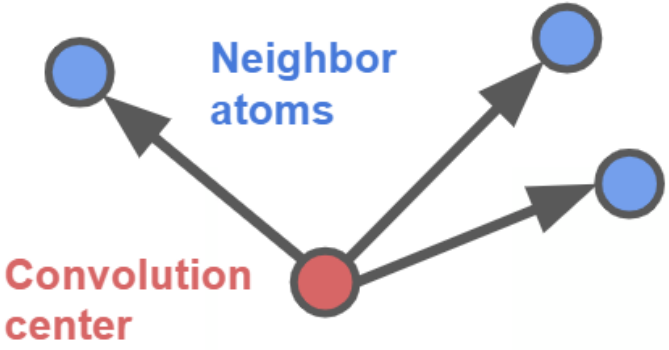


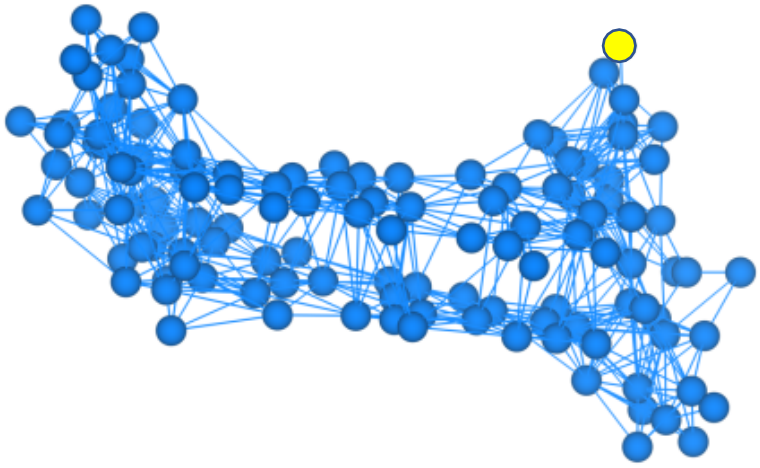
METHODS

e3nn is a type of **graph convolutional neural network** that is equivariant to translations, rotations, and reflections.



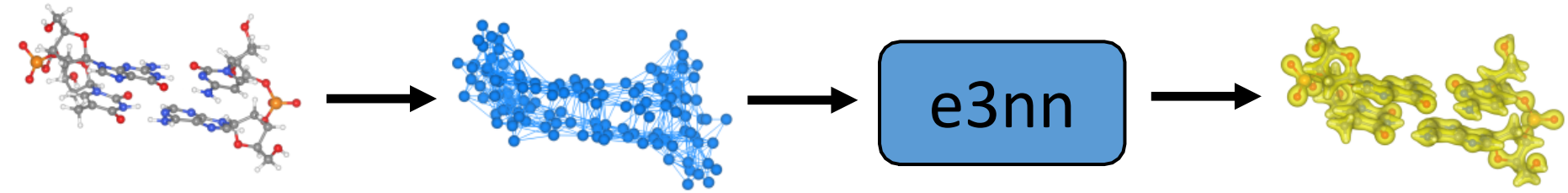
e3nn learns geometric features such as bonding interactions about its environment.



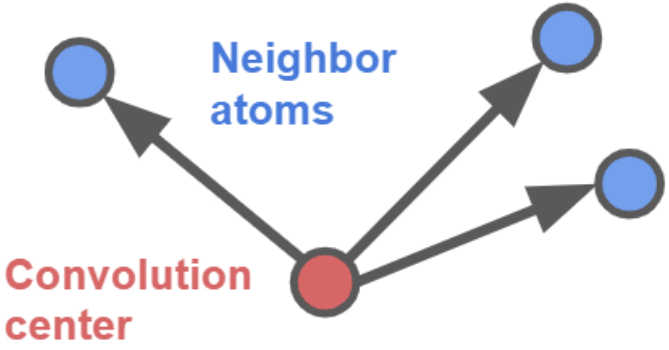


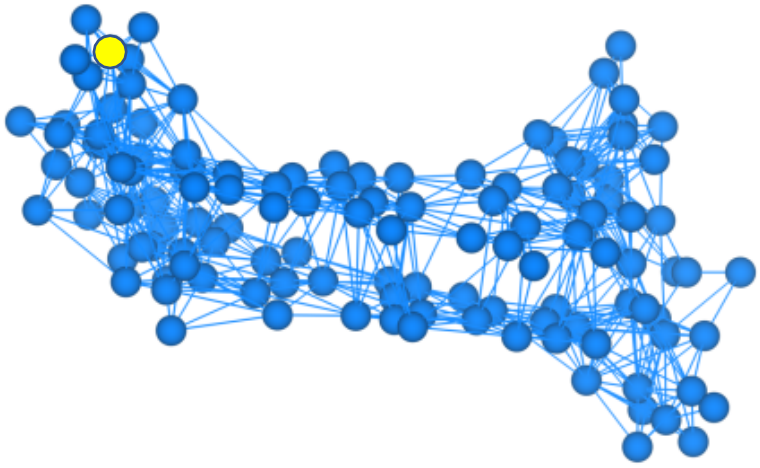
METHODS

e3nn is a type of **graph convolutional neural network** that is equivariant to translations, rotations, and reflections.



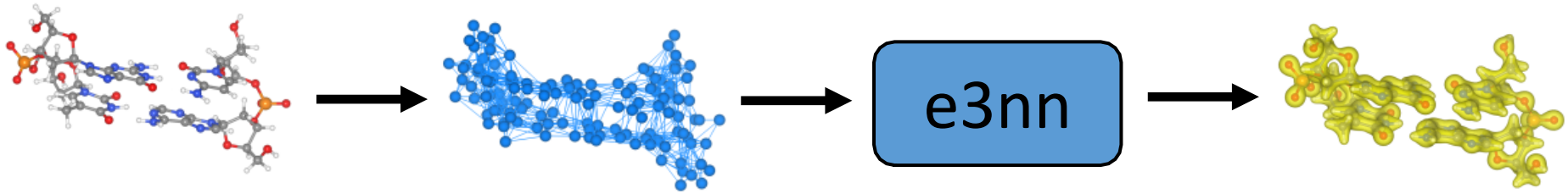
e3nn learns geometric features such as bonding interactions about its environment.



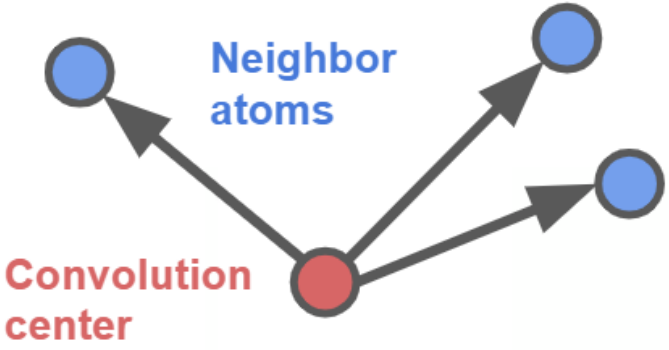


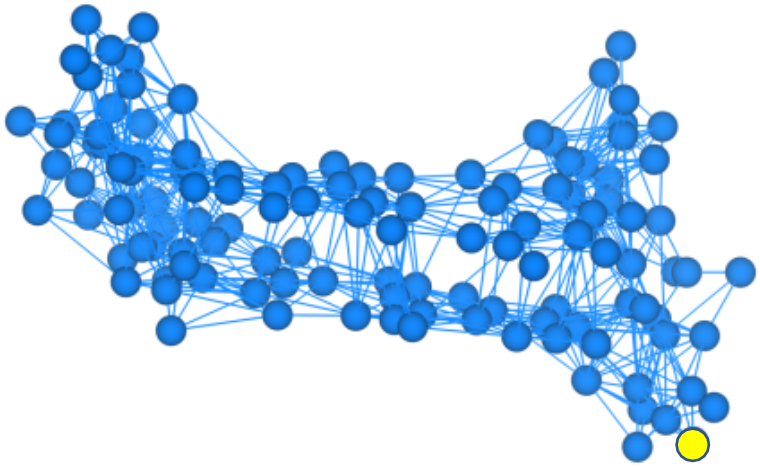
METHODS

e3nn is a type of **graph convolutional neural network** that is equivariant to translations, rotations, and reflections.



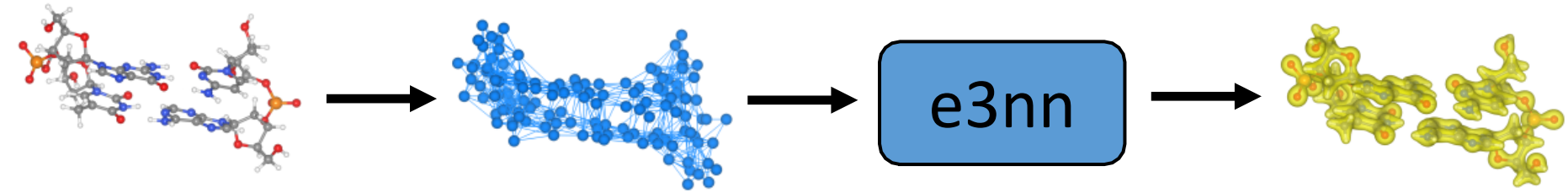
e3nn learns geometric features such as bonding interactions about its environment.





METHODS

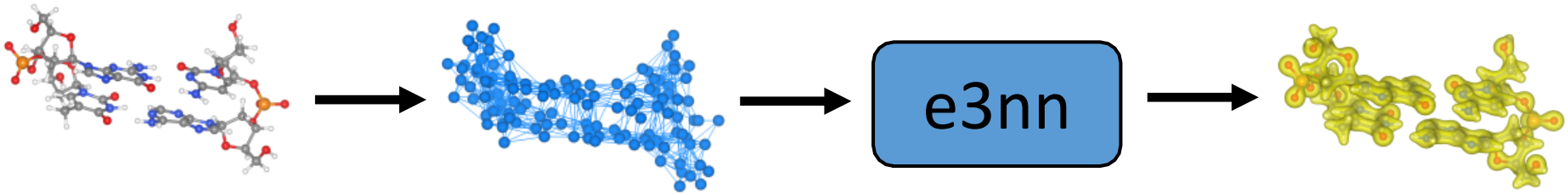
e3nn is a type of **graph convolutional neural network** that is equivariant to translations, rotations, and reflections.



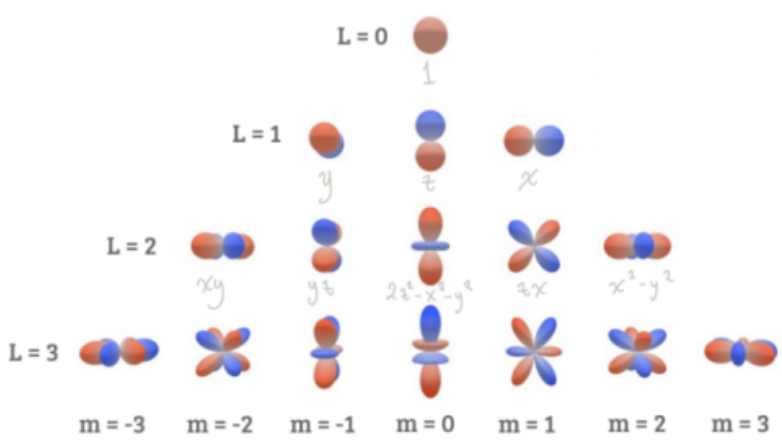
Output density: $\rho(\mathbf{r}) = \sum_{i=0}^{N_{atoms}} \sum_{k=0}^{N_{basis}} \sum_{l=0}^{l_{max}} \sum_{m=-l}^{+l} C_{iklm} Y_{lm} e^{-\alpha_{ikl}(r-r_i)^2}$

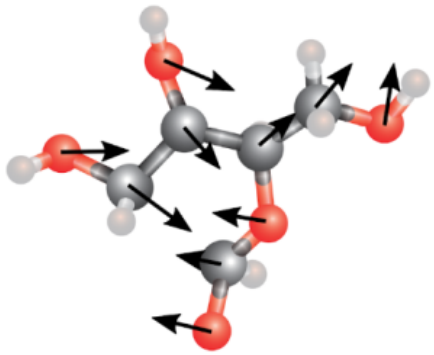
METHODS

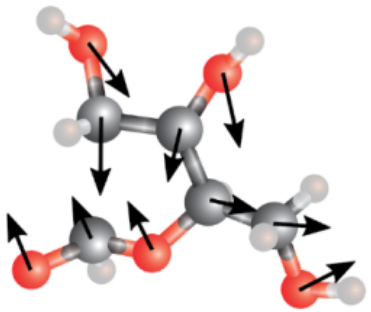
e3nn is a type of **graph convolutional neural network** that is equivariant to translations, rotations, and reflections.



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

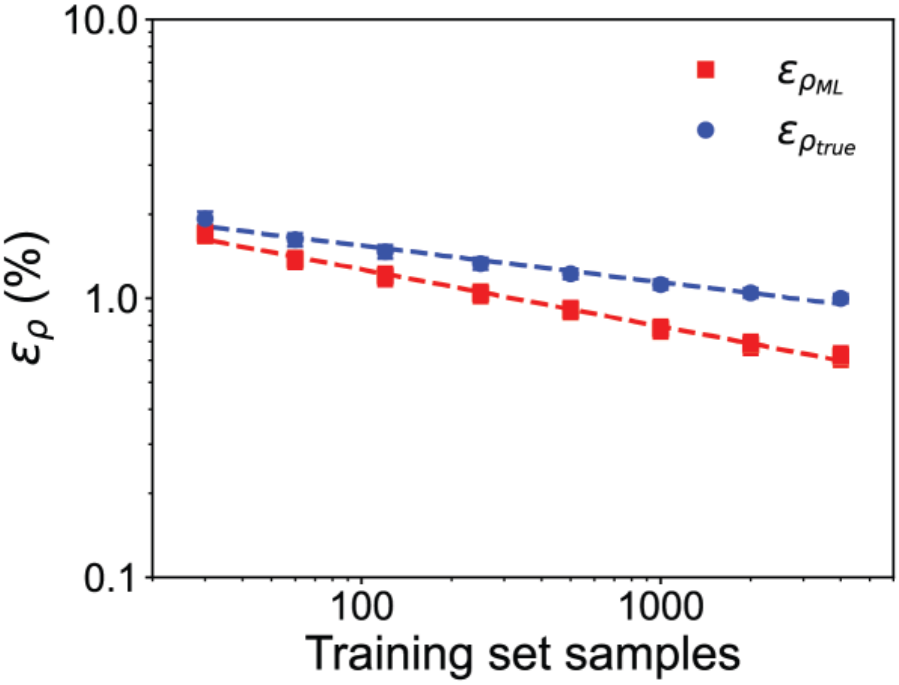


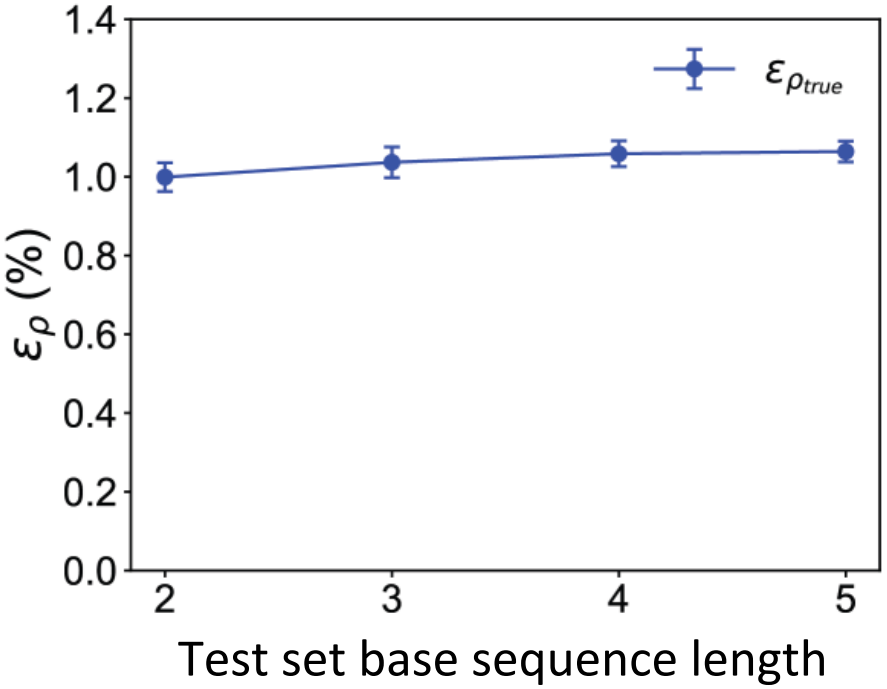




RESULTS

Density prediction error: $\epsilon_\rho(\%) = 100 \times \frac{\int d\boldsymbol{r}|\rho_{\mathrm{QM}}(\boldsymbol{r}) - \rho_{\mathrm{ML}}(\boldsymbol{r})|}{\int d\boldsymbol{r}\rho_{\mathrm{QM}}(\boldsymbol{r})}$

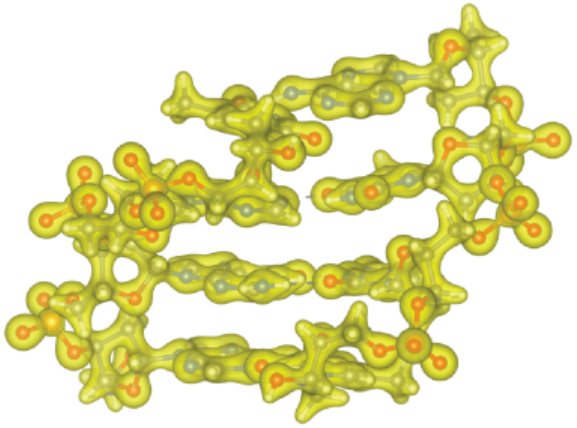




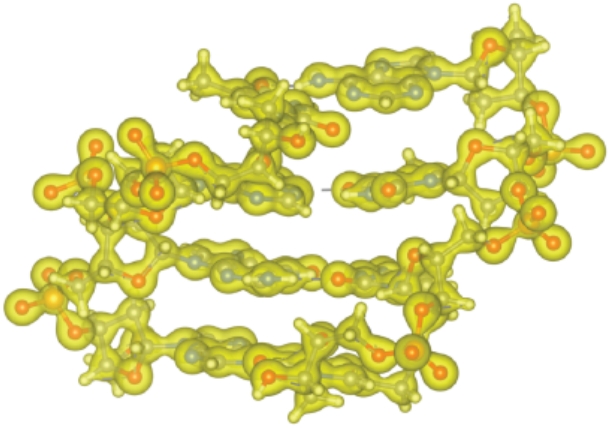
- The trained machine learning model predicts 2-mer test set densities with an error of $\epsilon_{\rho_{ML}}$ = 0.62%.
- ϵ_ρ does not significantly increase for longer DNA base sequences.

RESULTS

Example: 4-mer model predicted densities



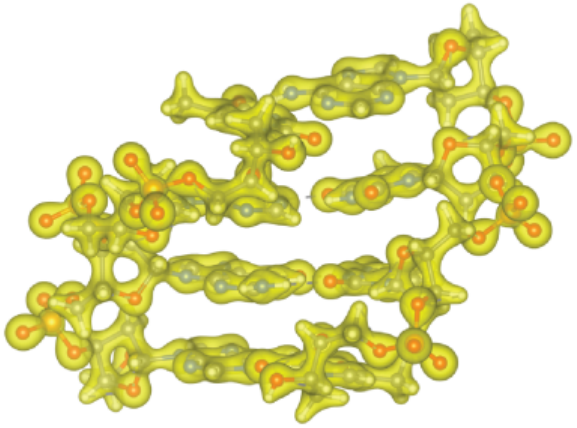
ρ_{ML} (ϵ_{true}= 1.02%)



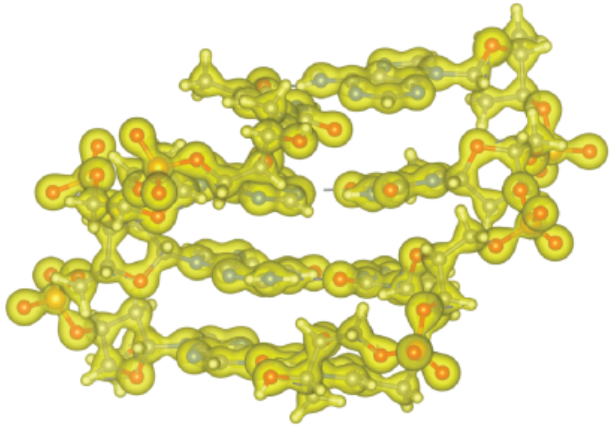
$\rho_{isolated\ atoms}$ (ϵ_{true}= 9.68%)

RESULTS

Example: 4-mer model predicted densities

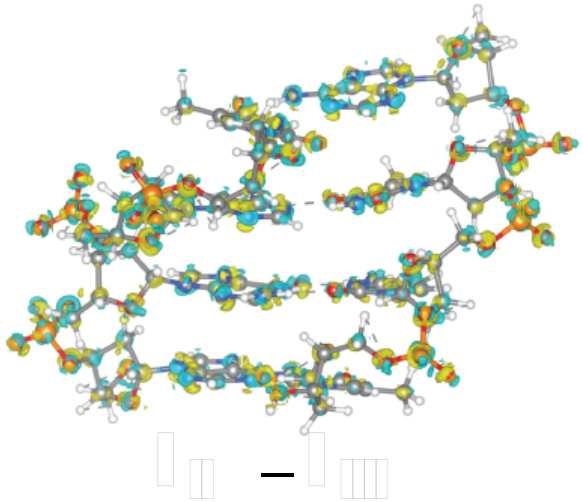


ρ_{ML} (ϵ_{true}= 1.02%)

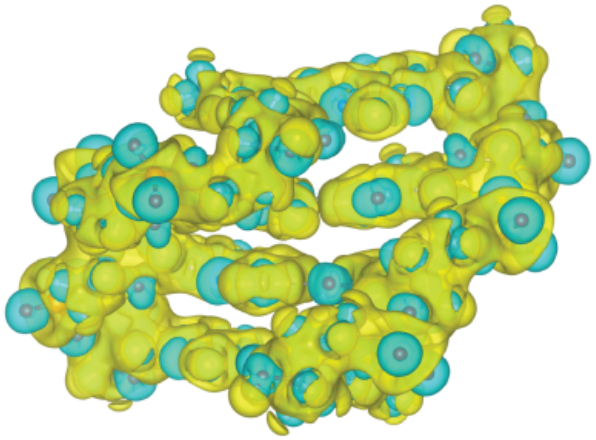


$\rho_{isolated\ atoms}$ (ϵ_{true}= 9.68%)

Density difference with true density



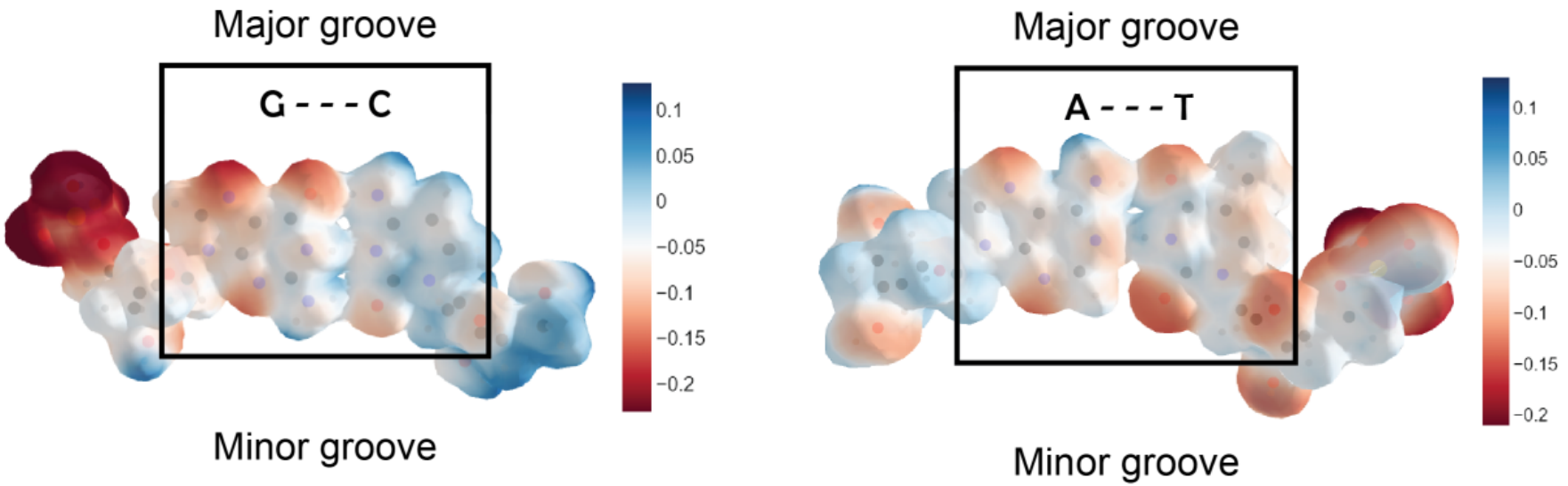
−

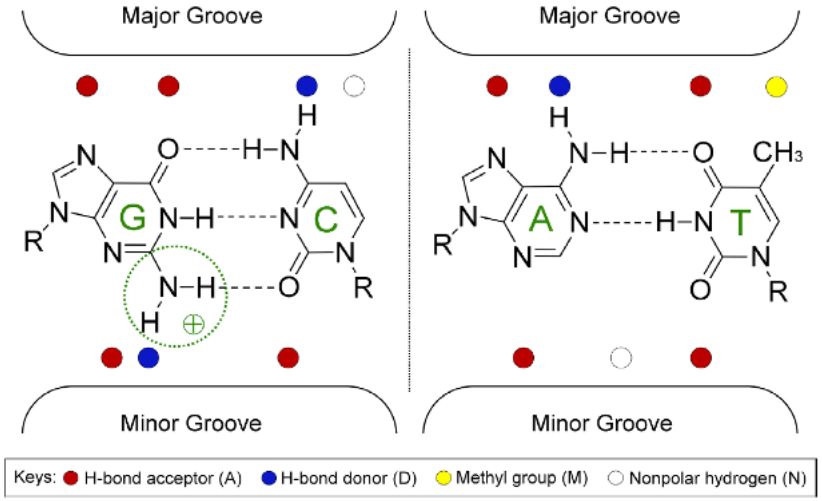


$\rho_{isolated\ atoms} - \rho_{true}$

RESULTS

Machine-learned electrostatic potentials

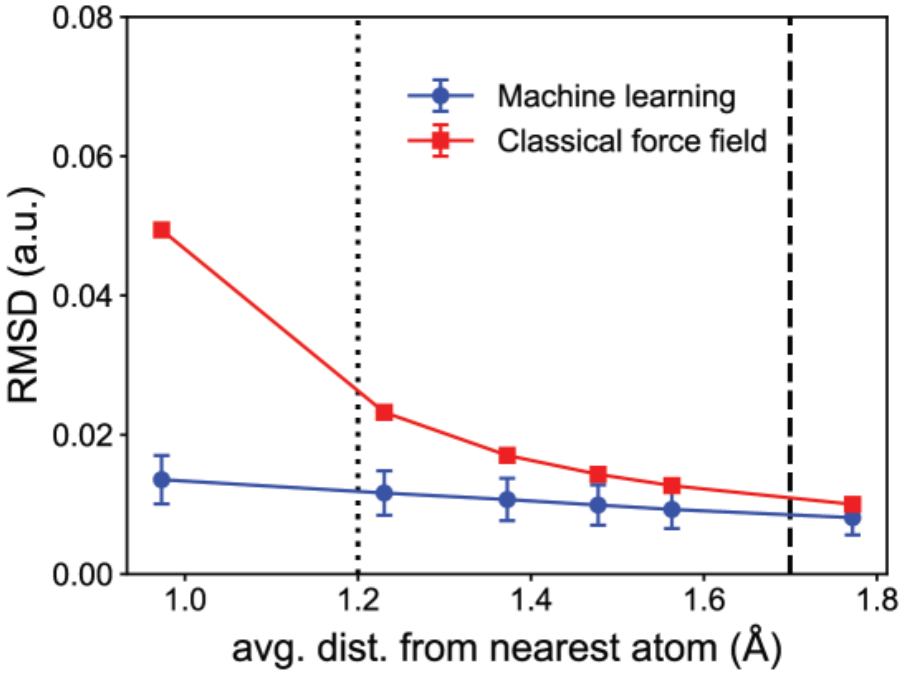


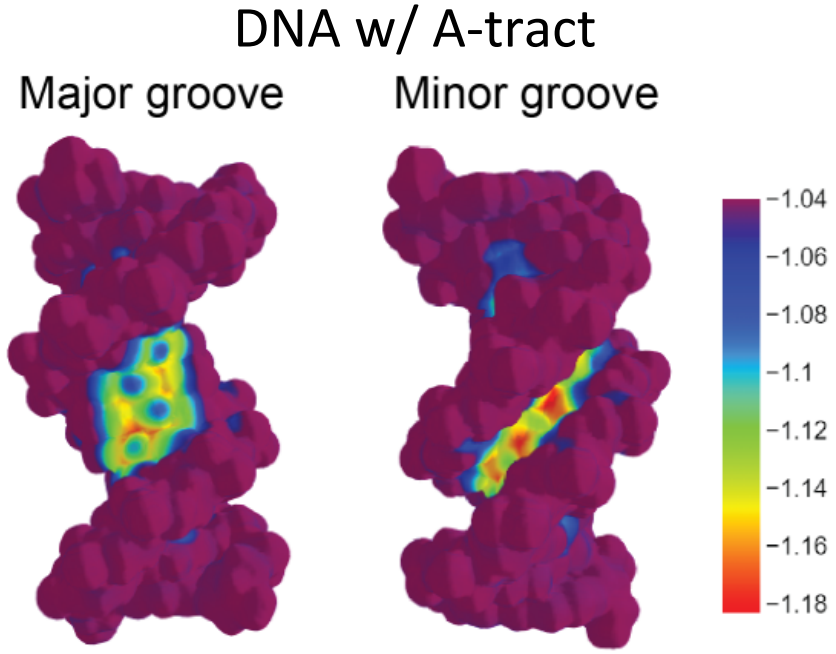


Chiu, *Nucleic Acids Research* (2017)

RESULTS

Machine-learned electrostatic potentials



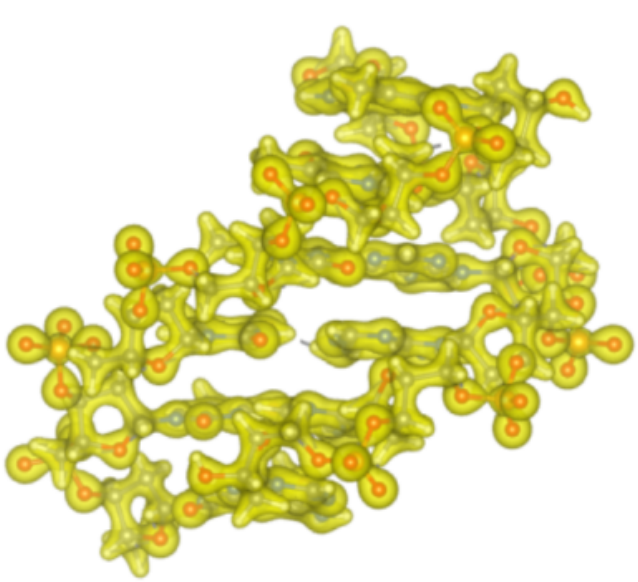


- Machine-learned electrostatic potentials are stable at short ranges compared to classical potentials.
- Application: Electrostatic potentials to study site-specific DNA-protein binding behavior

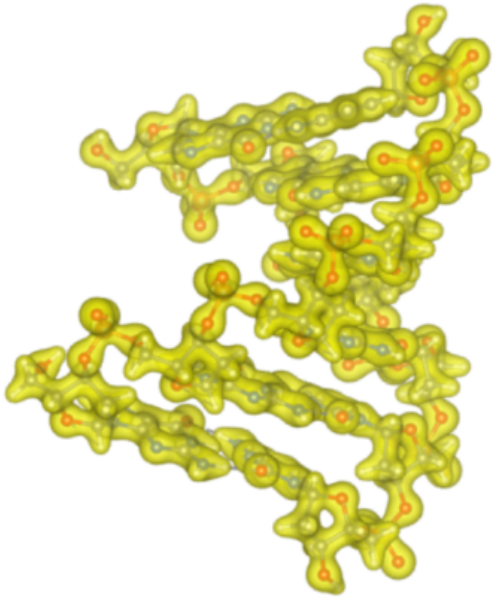
RESULTS

Extrapolating to DNA structural motifs beyond B-DNA

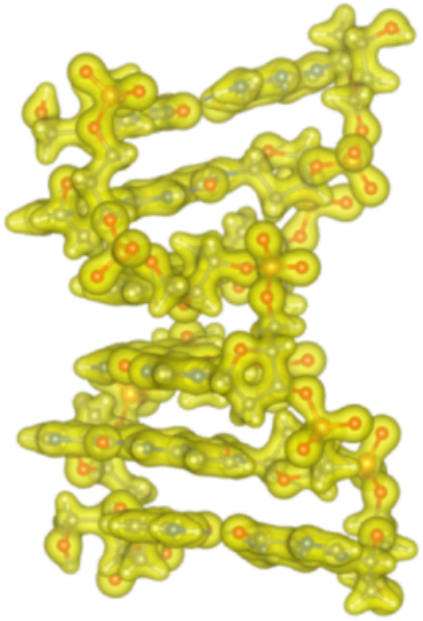
PDB code	Description	$\epsilon_{\rho true}$ (%)
251d [84]	B-DNA	1.15
1fzx [85]	B-DNA w/ A-tract	1.24
440d [86]	A-DNA	1.73
4fs5 [87]	Z-DNA	1.77



B-DNA w/ A-tract



A-DNA

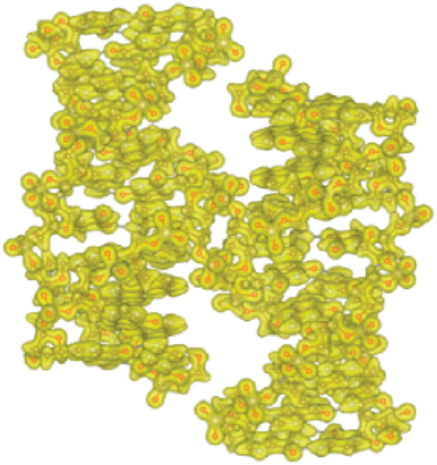


Z-DNA

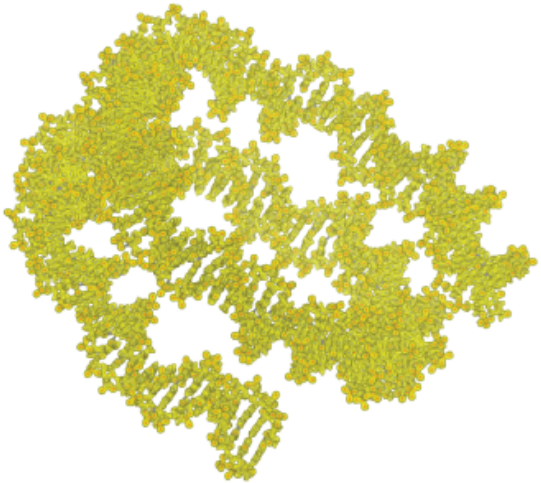
RESULTS

Examples of machine learning densities for large DNA structures

Stacked 4-way junction
1260 atoms, 6280 electrons
(PDB code 1dcw)



3.5 nm



10 nm

Nucleosome core particle
9346 atoms, 46980 electrons
(PDB code 1kx5)

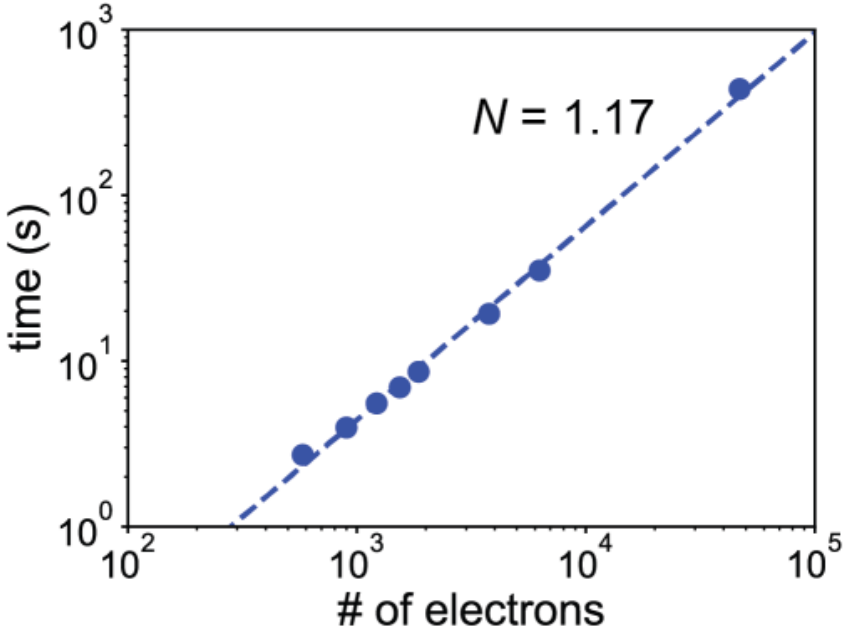
DNA origami structure
21658 atoms, 108654 electrons



25 nm

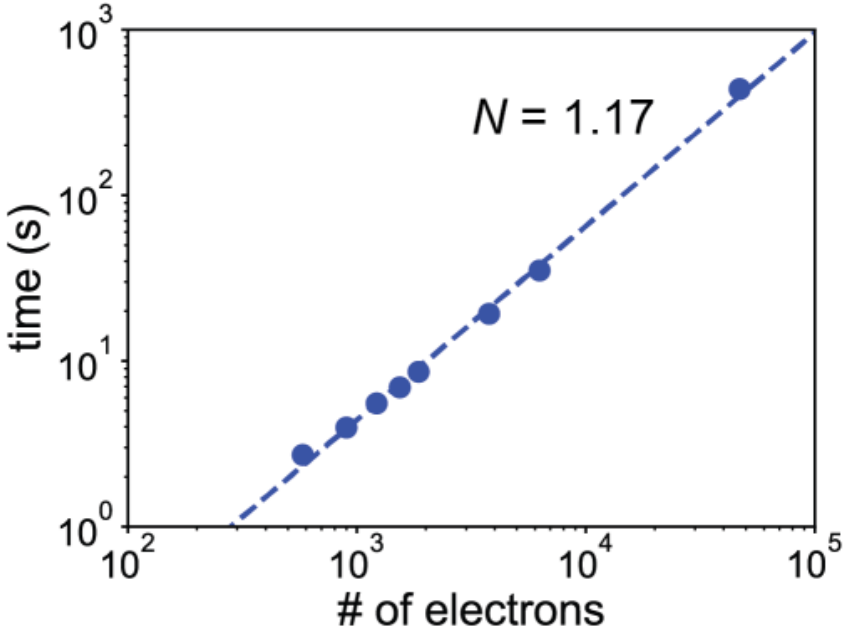
RESULTS

Examples of machine learning densities for large DNA structures



RESULTS

Examples of machine learning densities for large DNA structures



Computational scaling

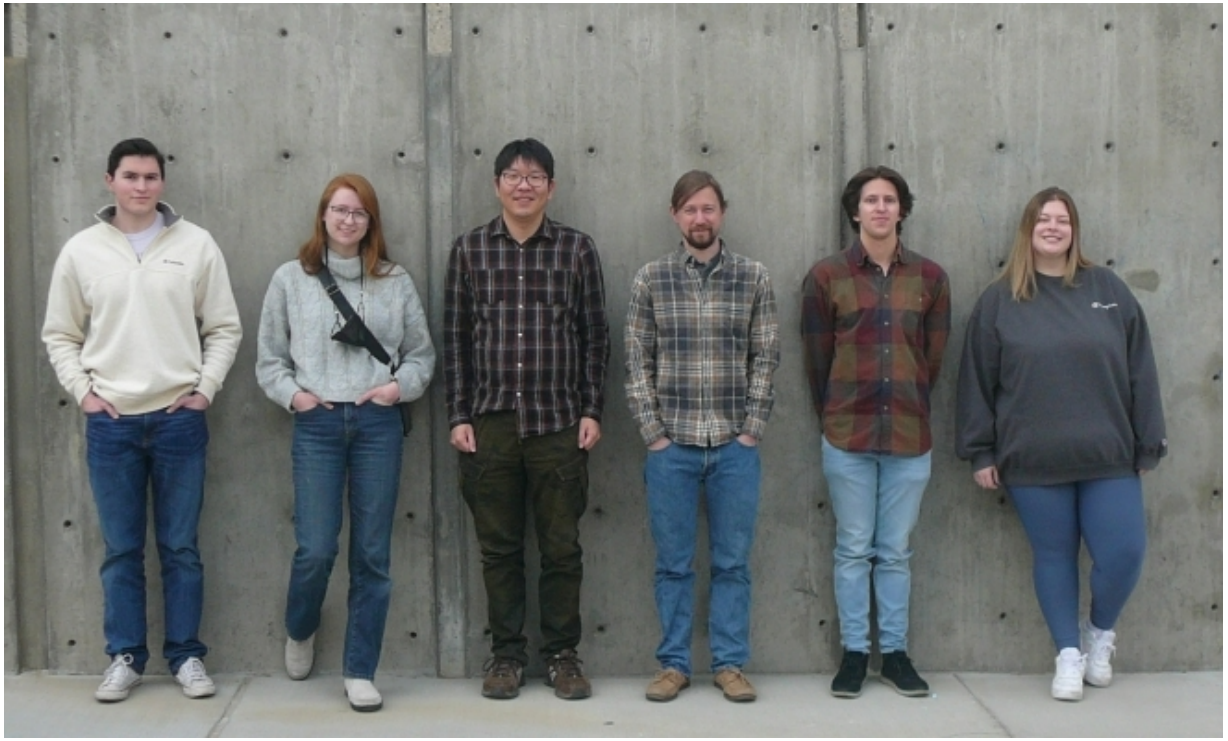
Density functional theory	(^{3})
Hybrid functionals	(^{4})
MP2 and beyond	(^{5} to ^{7})
Machine learning	( )

SUMMARY

- The machine learning model is based on an equivariant Euclidean neural network framework (e3nn) and is trained on representative B-DNA base pair steps.
- The model predicts DNA electron densities with a low error ($\epsilon_{\rho_{ML}}$ = 0.62%) that does not significantly increase for longer base sequences.
- Applications: Electrostatic potentials, forces for *ab initio* molecular dynamics, X-ray crystal structure refinement
- Paper on arXiv: xxxx.xxxxx

ACKNOWLEDGEMENTS

Bricker research group



Joshua Rackers
(Sandia Truman Fellow)



CHEMICAL & BIOLOGICAL ENGINEERING

U.S. DEPARTMENT OF ENERGY

LDRD LABORATORY DIRECTED RESEARCH & DEVELOPMENT

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.