Neural Networks
Learning Quantum Chemistry

Olexandr Isayev
University of North Carolina at Chapel Hill
olexandr@unc.edu
http://olexandrisayev.com
ANI-1: An extensible DL potential with DFT accuracy at force field computational cost

*Chem. Sci.*, 2017, 8, 3192-3203

ANI-1: A data set of 20M off-equilibrium DFT calculations for organic molecules

*Scientific Data* 4, Article number: 170193 (2017) DOI: 10.1038/sdata.2017.193

https://github.com/isayev/ASE_ANI

Justin S. Smith
Christian Devereux
Kavindri Ranasinghe
Adrian Roitberg

Roman Zubatyuk
Mariya Popova
Olexandr Isayev

Mark Berger
We want to train a padawan network to become a DFT jedi master.
Molecular mechanics/Classical force field

\[ u(r^N) = \sum_{\text{bonds}} k_i (l_i - l_{i,0})^2 + \sum_{\text{angles}} k_i (\theta_i - \theta_{i,0})^2 \]

\[ + \sum_{\text{torsions}} \frac{V_n}{2} (1 + \cos(n\omega - \gamma)) \]

\[ + \sum_{i=1}^{N} \sum_{j=i+1}^{N} \left( 4e_{ij} \left( \frac{\sigma_{ij}}{r_{ij}} \right)^{12} - \left( \frac{\sigma_{ij}}{r_{ij}} \right)^6 \right) + \frac{q_1 q_i}{r_{ij}} \]

Continuum solvent model

Hydrophobic effect is roughly proportional to surface area

The Electronic Schrödinger Equation (QM)

\[ \hat{H} \Psi(r; R) = E_{el} \Psi(r; R) \]

Electronic wave function

Electronic potential energy

- \[ \frac{1}{2} \sum_i \nabla_i^2 \]

\[ - \sum_{i,j} \frac{1}{r_{ij}} \]

\[ \sum_{A,i} \frac{Z_A}{r_{Ai}} \]

\[ \sum_{A>B} \frac{Z_A Z_B}{R_{AB}} \]
General purpose ML potentials

Accuracy

Force fields

Semi-empirical QM

DFT & HF

CCSD(T)

Cost

1

10

10^3

10^5

10^7
Design Principles

Create a “Force Field” in the sense of a mapping from coordinates $R \rightarrow$ Energy (Forces) with no a-priori functional form

- Accurate and reproducible
- Fast
- Input consisting only of things that the Schrödinger equation needs. (i.e. atomic numbers and positions, plus charge and spin)
- Forces as true gradients of the energy
- Extensible in atomic elements
- Extensible to molecules of very different sizes
- Self-learning
Machine learning molecular potentials

Training Set

Angular contribution

Radial contribution

Molecular Representation

Neural Net model

Energy prediction

\[ C = \frac{1}{N} \sum_{i} \left( E_{T}^{ML} - E_{T}^{QM} \right)^2 \]

Error feedback with parameter update


Total energy correlation
ANI-1 vs. DFT
(131 molecules with 10 heavy atoms, 8200 total molecules + conformations) [units: kcal/mol]

73 total structures
10 Heavy atoms
25 Total atoms
RMSE: 1.2 kcal/mol
(0.048 kcal/mol/atom)
DFT time: 1143.11s
ANI time: 0.0032s
357,000x speedup!
Relative Energy correlation (30kcal/mol)

ANI-1 potential unrelaxed scans

Generating the ANI-1 training set from GDB-11

- Non-equilibrium conformations generated via normal mode sampling (NMS)
- NMS samples along normal mode coordinates
- Final set contains ~22M data points

<table>
<thead>
<tr>
<th>Number of heavy atoms</th>
<th>Total Molecules</th>
<th>Total training data</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>3</td>
<td>8800</td>
</tr>
<tr>
<td>2</td>
<td>13</td>
<td>39370</td>
</tr>
<tr>
<td>3</td>
<td>20</td>
<td>128,880</td>
</tr>
<tr>
<td>4</td>
<td>63</td>
<td>535,660</td>
</tr>
<tr>
<td>5</td>
<td>275</td>
<td>1,444,890</td>
</tr>
<tr>
<td>6</td>
<td>1,408</td>
<td>1,309,620</td>
</tr>
<tr>
<td>7</td>
<td>7,850</td>
<td>5,276,930</td>
</tr>
<tr>
<td>8</td>
<td>48,319</td>
<td>8,472,200</td>
</tr>
<tr>
<td>Total</td>
<td>57,951</td>
<td>17,216,350</td>
</tr>
</tbody>
</table>

Use the ANI-1 dataset:
ANI-1: A data set of 20M off-equilibrium DFT calculations for organic molecules

Available at: https://github.com/isayev/ANI1_dataset
How can we tell if the network is out of domain and should not be trusted?
Active Learning (AL) is a powerful tool!

Model QBC tells us where the potential breaks in two areas:

1. where a model overfits
2. where we lack data to describe a specific interaction

<table>
<thead>
<tr>
<th>ANI Model</th>
<th>Entire data set</th>
<th>ANI Energies</th>
<th>Standard deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>ANI-CV1</td>
<td></td>
<td>$E_1$</td>
<td>$\sigma = \sqrt{\frac{1}{5} \sum_{i=1}^{5} (E_i - \bar{E})^2}$</td>
</tr>
<tr>
<td>ANI-CV2</td>
<td></td>
<td>$E_2$</td>
<td></td>
</tr>
<tr>
<td>ANI-CV3</td>
<td></td>
<td>$E_3$</td>
<td></td>
</tr>
<tr>
<td>ANI-CV4</td>
<td></td>
<td>$E_4$</td>
<td></td>
</tr>
<tr>
<td>ANI-CV5</td>
<td></td>
<td>$E_5$</td>
<td></td>
</tr>
</tbody>
</table>

- **Held out for validation**
- **Included in the training set**

Active Learning - The Big Picture
An automated and self-consistent data generation framework
Justin S Smith, Ben Nebgen, Nicholas Lubbers, Olexandr Isayev, Adrian E Roitberg

Non-equilibrium Conformational sampler

Molecule Sampling
(e.g. GDB small molecule database, small peptides, drug like molecules)

Structure Pools

The Journal of Chemical Physics 148 (24), 241733

Compute Cluster
Computations with QM

Ensemble of ANI networks

Train network ensemble

ANI-1x Dataset
(i.e. energies, forces, dipoles)

Query by committee
New test data

The Journal of Chemical Physics 148 (24), 241733
Active Learning allows us to get better accuracy while training to a much small data set.
Training and benchmark (COMP6 set)

- Training:
  - ~65k molecules
  - ~84 conformers

- GDB07to09:
  - 2000 molecules
  - 24 conformers

- GDB10to13:
  - 3000 molecules
  - 24 conformers

- DrugBank:
  - 850 molecules
  - 16 conformers

- Tripeptide:
  - 250 molecules
  - 8 conformers

- s66x8:
  - 66 molecules
  - 8 conformers
ANI-MD Benchmark

• 12 drug molecules and 2 proteins
• Mean size 75 atoms (max 312 atoms)
• 1ns of molecular dynamics (MD)
• Dynamics at 300K
• MD ran on ANI-1x potential
• 128 randomly sampled frames
Active-learning results vs. random sampling
ANI-1x forces vs. baseline semi-empirical QM

Correlation of force components for the ANI-MD Benchmark
ANI-1x optimized drug molecules

DrugBank - Bonds

Work in progress with Christian Devereux @ UF

The diagram shows box plots comparing bond errors across different methods for Drug and Tripeptides.

- **Drug**
  - ANI
  - UFF
  - MMFF
  - PM6

- **Tripeptides**
  - ANI
  - UFF
  - MMFF
  - PM6
ANI-1x optimized drug molecules

DrugBank - Angles

Work in progress with Christian Devereux @ UF
ANI-1x predicted harmonic frequencies

Work in progress with Christian Devereux @ UF

DrugBank Benchmark (cm⁻¹)

Tripeptide Benchmark (cm⁻¹)
Diels-Alder reaction

Substituents (42)

Diagram showing the reaction between a diene and a dienophile, resulting in a cyclohexane structure.

Substitution sites

Active-learning reactions: Diels-Alder reaction

<table>
<thead>
<tr>
<th></th>
<th>New data points</th>
<th>Avg. RMSE* (kcal/mol)</th>
<th>Avg. R/P ∆∆E (kcal/mol)</th>
<th>Barrier Height (kcal/mol)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Original ANI network</td>
<td>0</td>
<td>16.67</td>
<td>3.49</td>
<td>30.27</td>
</tr>
<tr>
<td>60 random Diels-Alder reactions in sample (interpolation)</td>
<td>0.70</td>
<td>0.88</td>
<td>0.70</td>
<td></td>
</tr>
<tr>
<td>60 random Diels-Alder reactions out of sample/ held out test set reactions (extrapolation)</td>
<td>~30k</td>
<td>1.28</td>
<td>0.81</td>
<td>2.00</td>
</tr>
</tbody>
</table>
Active-learning reactions: Diels-Alder reaction
Claisen rearrangement

Auto-TS (Transition state generator):

Substitution sites

Active-learning reactions: Claisen rearrangement
Cope rearrangement

Substituents (6)

Auto-TS (Transition state generator):

QM data

Substitution sites

Active-learning reactions: Cope rearrangement

**Graphical Data:***

- Energy (kcal/mol) vs. Reaction Coordinate (angstroms) for different reactions.
- Comparison between DFT and ANI methods for energy calculations.
- Table of energy values for different reactions.

**Table:**

<table>
<thead>
<tr>
<th>Reaction</th>
<th>DFT Energy</th>
<th>ANI-1 Energy</th>
<th>ANI-rxn Energy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reaction 1</td>
<td>6.3 kcal/mol</td>
<td>6.1 kcal/mol</td>
<td></td>
</tr>
<tr>
<td>Reaction 2</td>
<td>6.3 kcal/mol</td>
<td>4.5 kcal/mol</td>
<td></td>
</tr>
<tr>
<td>Reaction 3</td>
<td>49.6 kcal/mol</td>
<td>3.2 kcal/mol</td>
<td></td>
</tr>
<tr>
<td>Reaction 4</td>
<td>26.7 kcal/mol</td>
<td>1.5 kcal/mol</td>
<td></td>
</tr>
<tr>
<td>Reaction 5</td>
<td>28.7 kcal/mol</td>
<td>0.3 kcal/mol</td>
<td></td>
</tr>
</tbody>
</table>
Can we do more?
AIMNet & Multimodal Learning

Work in progress by Roman Zubatyuk

Deep NN network: 45 hidden layers, 940k parameters
Performance of AIMNet network

$E_{\text{form}}$

$\Delta E$

$F$

$q$

$Q$

$C_6$

Work in progress by Roman Zubatyuk
Simulated IR spectra

NVT molecular dynamics, 300 K, 500ps
35 ms/step using Python ASE interface.

Work in progress by Roman Zubatyuk
Transferring knowledge of CCSD(T)/CBS

- Regenerate 10% of ANI-1x training data (0.5M of 5M)
- For high-level reference we use CCSD(T)/CBS accurate QM model
- We only retrain 60k of 400k neural network parameters
- Results show clear improvement over DFT trained model
- New models are exceeding the DFT in accuracy

\[ C = \frac{1}{M} \sum_{j}^{M} \left( E_{ANI,j}^{CC} - E_{j}^{CC} \right)^2 \]

\[ E_{ANI} = \sum_{i}^{N} E_{ANI,i} \]

\[ E_{DFT} = \sum_{i}^{N} E_{DFT,i} \]

Transfer learning algorithm

Work in progress by Justin Smith & Roman Zubatyuk
Toward CCSD(T)/CBS accuracy for a general purpose neural network potential

Work in progress by Justin Smith & Roman Zubatyuk
Genentech Dihedral Benchmark

Can we do even more?
Mycobacterium tuberculosis (5MXV) in explicit water
Simulated with ANI (CHNOSFCl)

- ~35K atoms
- Explicit water
- No ions
- S, F and Cl in ligand

https://www.rcsb.org/GSK1107112A

C₁₂ H₈ Cl F N₂ O S₂
Timings for a 5x ensemble prediction for ANI-2x

<table>
<thead>
<tr>
<th>GPU</th>
<th>ANI-2x time per step</th>
<th>Total time per step</th>
<th>Steps per day</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tesla V100</td>
<td>297ms</td>
<td>317ms</td>
<td>272k</td>
</tr>
</tbody>
</table>

0.5ns simulation time
Simulation of Complex Chemical Reactions

Carbon nanoparticles/sheets nucleation [4000 atoms in 60A box at 2500K, 5ns MD simulation ]

https://youtu.be/DRVMH5u8EA0
Rapid Progress of ANI-1 Capabilities

ANI-1 paper was published Feb 2017

Features:
- CHNOSFCl
- Analytical forces
- Thermochemistry
- Molecular Dynamics
- Reactions
- Molecular crystals, PBC
- Single Network for all elements
- Automated Active Learning

Users:
- Used in academic labs:
  - Stanford (Vijay Pande)
  - U Pitt (Geoff Hutchison)
  - USF (Lee Woodcock)
  - CMU
  - NCSU

Government labs, companies etc.

Summer 2018
Use the ANI-1x potential:

ANI-1: An extensible DL potential with DFT accuracy at force field computational cost

Less is more: sampling chemical space with active learning (ANI-1x).

ANI-1x interfaced to ASE Python library
Available at: https://github.com/isayev/ASE_ANI

Use the ANI-1 dataset:
ANI-1: A data set of 20M off-equilibrium DFT calculations for organic molecules

ANI Data set Python library
Available at: https://github.com/isayev/ANI1_dataset