Electronic Supplementary Material (ESI) for Chemical Science.

Supplementary information for: *"ANI-1: An extensible neural network potential with DFT accuracy at force field computational cost"*

Justin S. Smith[1], Olexandr Isayev[2,*], Adrian E. Roitberg[1,*]

[1]*Department of Chemistry, University of Florida, Gainesville, FL 32611, USA*

[2]*UNC Eshelman School of Pharmacy, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, USA*

* Corresponding authors; email: O.I. (olexandr@olexandrisayev.com) or A.E.R. (roitberg@ufl.edu)

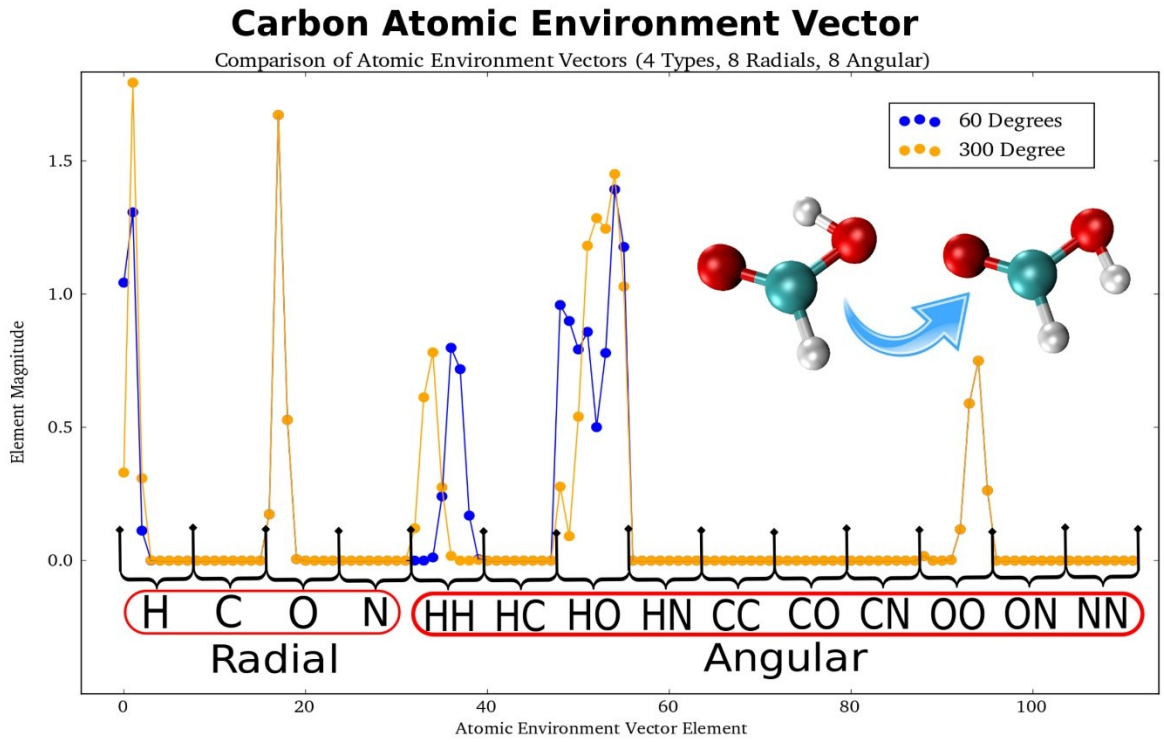

Figure S1: A visualization of atomic environment vectors for the carbon atom ($\vec{G}_1^C$) in formic acid, computed with our modified angular symmetry functions and atomic number differentiated. The figure shows two $\vec{G}_1^C$, blue and orange, of two conformations and labels each sub-vector for clarity. The two conformation only differ in the C-O-H angle depicted in the figure.

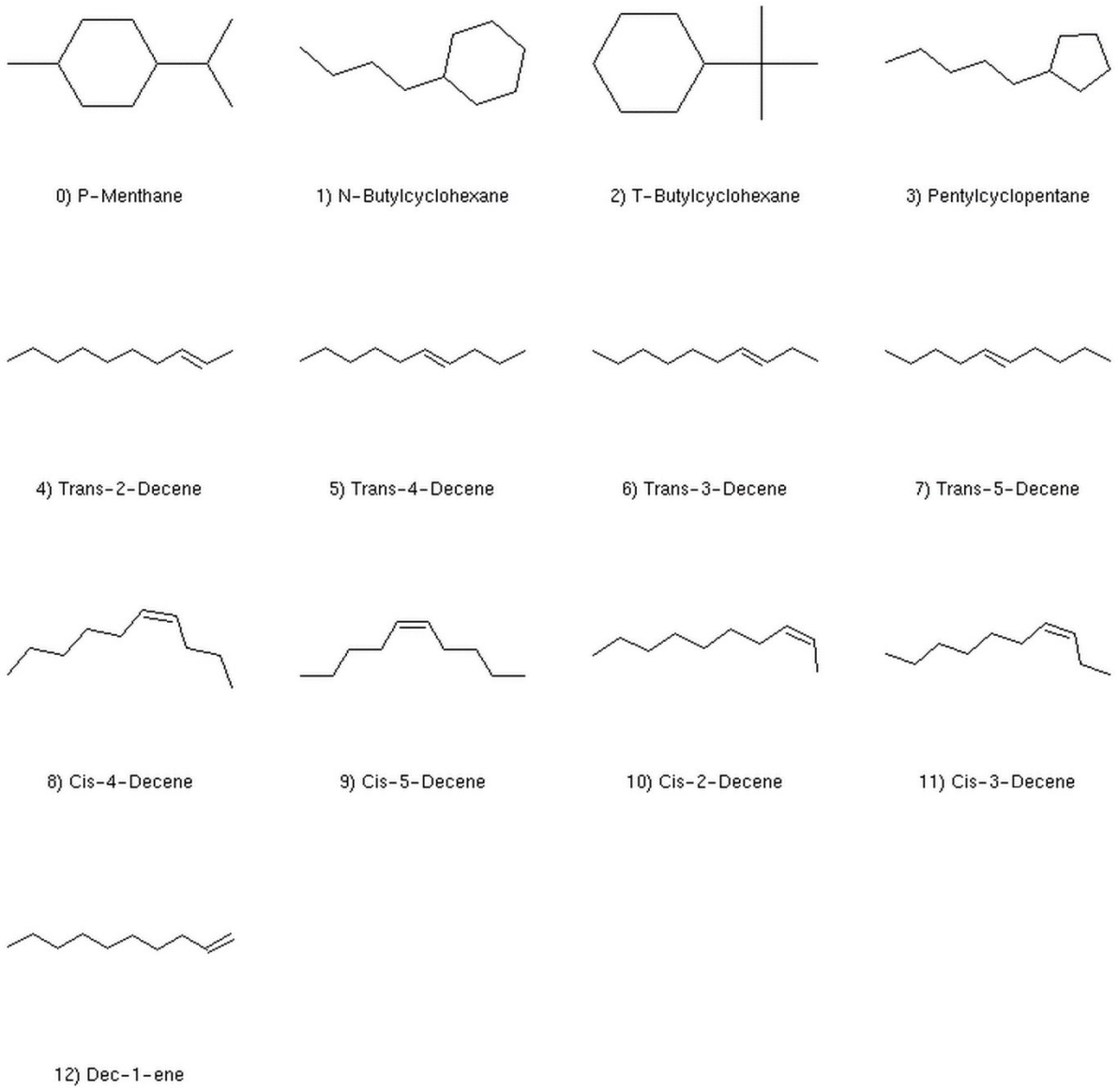

Figure S2: All structural and geometric isomers used to generate the data for the isomer case study in section 4.2. The molecular indices map to the isomer index (x-axis) of Figure 4 in Section 4.2.

Number of heavy atoms	Total Molecules	Max Temperature	S value	Total data points	ANI-1 test set RMSE per atom (kcal/mol/atom)
1	3	2,000.0	500	8800	7.33×10^{-2}
2	13	1,500.0	450	39370	5.96×10^{-2}
3	20	1,000.0	425	128,880	4.16×10^{-2}
4	63	600.0	400	535,660	3.41×10^{-2}
5	275	600.0	200	1,444,890	3.71×10^{-2}
6	1,408	600.0	30	1,309,620	4.36×10^{-2}
7	7,850	600.0	20	5,276,930	6.65×10^{-2}
8	48,319	450.0	5	8,472,200	7.43×10^{-2}
Total	57,951	-	-	17,216,350	6.66×10^{-2}

Table S1: List of information and parameters used to generate the ANI-1 data set. The first column represents the number of heavy atoms per molecule in the test set. Total represents a combination of all test sets. The molecules are obtained from the GDB-11 database.

Statistic **(Energy units of kcal/mol)**	**ANI-1 Performance**
MAE	1.316
% MAE	1.084×10^{-3}
RMSE	1.915
% RMSE	1.578×10^{-3}
MAPE (%)	4.484×10^{-4}
RMSE (kcal/mol/atom)	7.996×10^{-2}
Slope	1.000
Intercept	-1.493
R squared	1.000
Compute time (ms)	286.4
Data points	8245
Time per data point (μs)	34.74

Table S2: Statistics comparing the absolute energies of ANI-1 and DFT for a test set of 62 conformations of each 134 randomly selected molecules with 10 heavy atoms. Since this is a comparison of absolute energies, the range of energies is very large: from -365,343 to -243,973 kcal/mol.

134 molecules from GDB-10 NMS generated test set						
E^{cap} (kcal/mol)	RMSE	MAE	RMSE/atom	Max $\|\Delta E\|$	Relative RMSE	Data points
500	5.626	1.987	1.86E-01	135.966	5.589	9171
400	2.818	1.531	1.09E-01	78.449	2.708	8819
300	1.915	1.316	8.00E-02	23.876	1.768	8245
200	1.616	1.164	6.76E-02	12.722	1.367	7032
100	1.363	0.999	5.50E-02	8.226	0.977	4485
75	1.270	0.936	5.06E-02	8.226	0.843	3530
50	1.179	0.867	4.61E-02	8.226	0.694	2493
30	1.126	0.831	4.23E-02	4.551	0.566	1555
20	1.092	0.809	4.06E-02	4.332	0.454	1084
10	1.019	0.773	3.75E-02	3.953	0.363	621
Min	1.034	0.778	3.56E-02	3.634	N/A	134

Table S3: The ANI-1 potentials performance on 9171 normal mode sampling (NMS) generated conformers of 134 randomly selected molecules from the GDB-10 database. E^{cap} is imposed on a per molecules basis by throwing out any conformers that have energies E^{cap} higher than the minimum energy for that molecule's set of conformers. This leaves only conformers closer to the minimized energy structure as E^{cap} is reduced, until only the minimum energy (min) for each molecule is considered. Columns 2 through 4 show various errors to the total energies from DFT reference calculations. Column 5 shows the maximum $|\Delta E|$ over the entire data set. Column 6 shows the RMSE of energies relative to the minimum energy for each molecule's set of structures.

ANI method Network performance vs data set size				
	(Error: RMSE kcal/mol)			
	Fractional Data		Full Data	
Percent	Train	Valid	Test	GDB-10 Test
5.00%	1.49	2.07	2.10	3.21
5.00%	1.56	2.07	2.13	3.16
5.00%	1.44	2.02	2.09	3.02
5.00%	1.60	2.06	2.14	3.11
10.00%	1.39	1.73	1.80	2.68
10.00%	1.29	1.68	1.77	2.83
10.00%	1.44	1.80	1.83	2.81
25.00%	1.18	1.43	1.45	2.28
25.00%	1.17	1.42	1.45	2.41
25.00%	1.15	1.40	1.44	2.46
25.00%	1.20	1.42	1.46	2.37
50.00%	1.17	1.32	1.34	2.22
50.00%	1.20	1.33	1.36	2.22
75.00%	1.09	1.20	1.21	2.06
100.00%	1.16	1.28	1.28	1.91
Baseline - No type differentiation				
100.00%	3.61	3.78	3.84	6.55
Baseline – CM/MLP				
5.00%	42.17	46.61	48.07	1047.84
10.00%	45.49	45.77	47.14	1457.68
25.00%	35.44	38.03	38.15	503.57
50.00%	35.33	39.28	38.63	1422.11
75.00%	34.56	36.61	36.71	460.87
100.00%	33.79	35.96	36.09	493.70

Table S4: Shows how the ANAKIN-ME method scales with the size of the training set as well as information about two baseline methods trained on the same data set. The "Percent" column shows what percentage of the 17.2 million data points was used to train, validate, and test the model. The train and validate columns show the RMSE of the actual training and validation set, fractional data, used to train the model while the test sets are always full sets. The first baseline method shows how the ANAKIN-ME method performs without differentiating atomic numbers within the AEVs. The second baseline shows the performance of a sorted coulomb matrix with a multilayer perceptron (CM/MLP) neural network model on the ANI-1 data set with training set size scaling.