Supporting Information

ANI Neural Network Potentials for Small Molecule pK_a Prediction

Ross James Urquhart, Alexander van Teijlingen, Tell Tuttle*

*tell.tuttle@strath.ac.uk

Table of Contents

Supporting Information

1. Generation of Conformers

1.1 CREST Conformer Generation

Table S1: Details regarding number of CREST Generated Conformers based on Avogadro UFF minimized structures.

Carbene	Number of Conformers generated			
	Gas Phase		Aqueous Phase	
	Protonated	Deprotonated	Protonated	Deprotonated
1	3	2	3	2
2	4	4	3	4
3	2	3	3	3
4	11	9	9	9
5	9	4	8	3
6	5	7	7	8
7	2	3	2	3
8	4	6	4	4
9	2	1	2	2
10	2	1	1	2
11	2	2	2	2

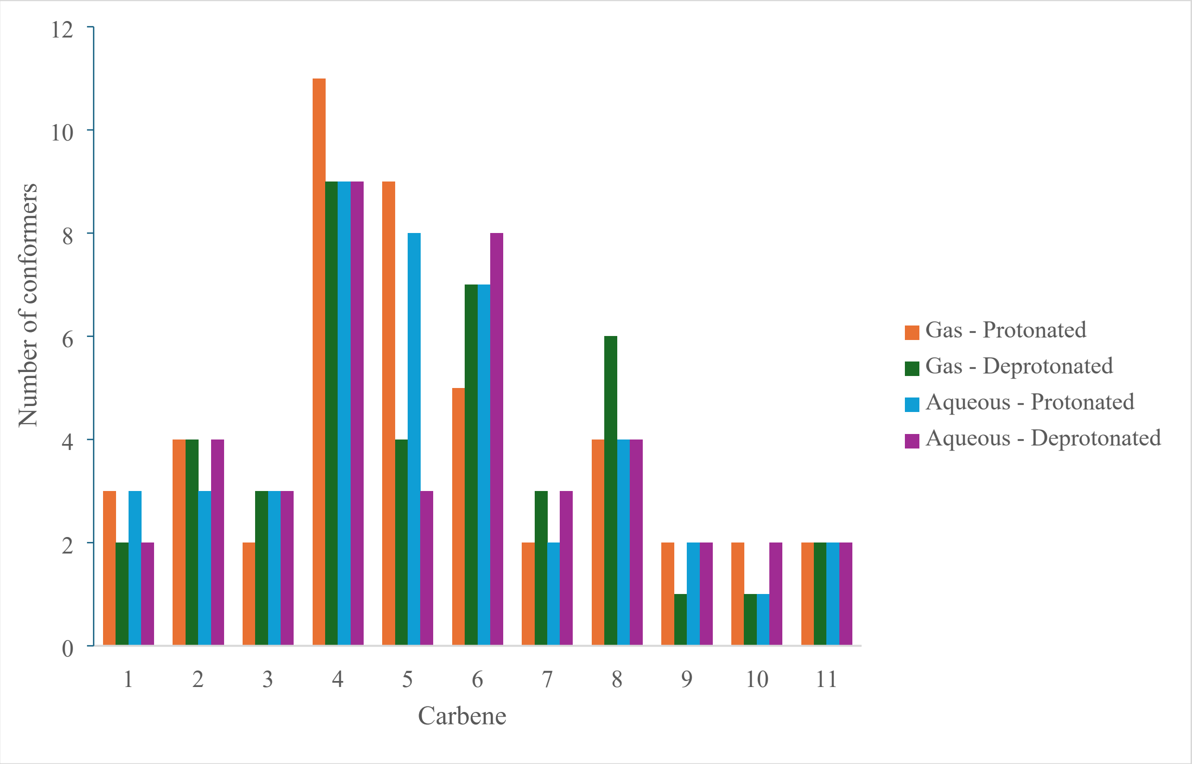

Figure S1: Bar chart detailing the number of conformers generated per phase and state for each carbene within the validation set based on Avogadro UFF minimized structures.

Table S2: Details regarding number of CREST generated conformers resulting from DFT optimised structures.

Carbene	Number of Conformers generated			
	Gas Phase		Aqueous	
	Protonated	Deprotonated	Protonated	Deprotonated
1	3	3	2	3
2	3	2	2	3
3	2	2	3	2
4	9	10	10	10
5	3	5	5	8
6	7	6	5	6
7	3	3	2	2
8	3	3	4	6
9	1	2	1	1
10	1	1	1	2
11	2	2	3	3

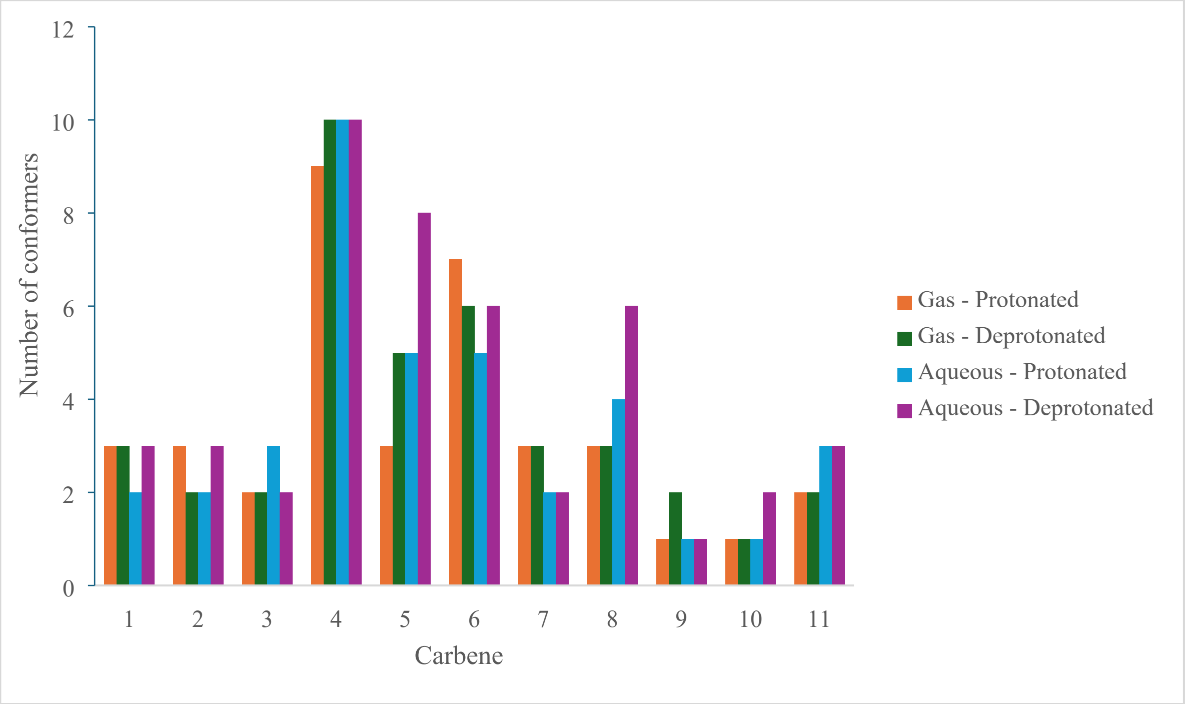

Figure S2: Bar chart detailing the number of conformers generated per phase and state for each carbene within the validation set based on Avogadro UFF minimized structures.

1.2 Conformer Analysis

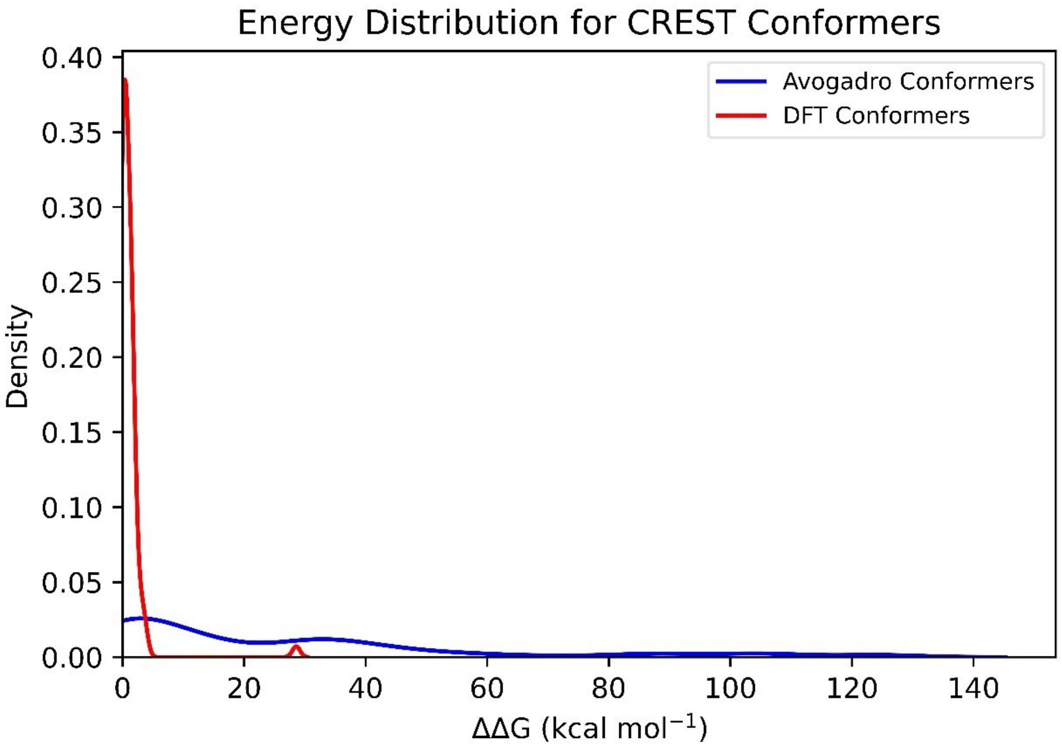

Figure S3: Energy Distribution detailing ΔΔG value compared to the lowest energy conformer for both DFT and Avogadro generated conformers of each carbene. DFT generated conformers show a smaller distribution with most conformers located close in energy to the minimum energy conformer. Conversely, the Avogadro generated conformers show a larger distribution and thus shows that relatively higher energy conformers are generated compared to with DFT.

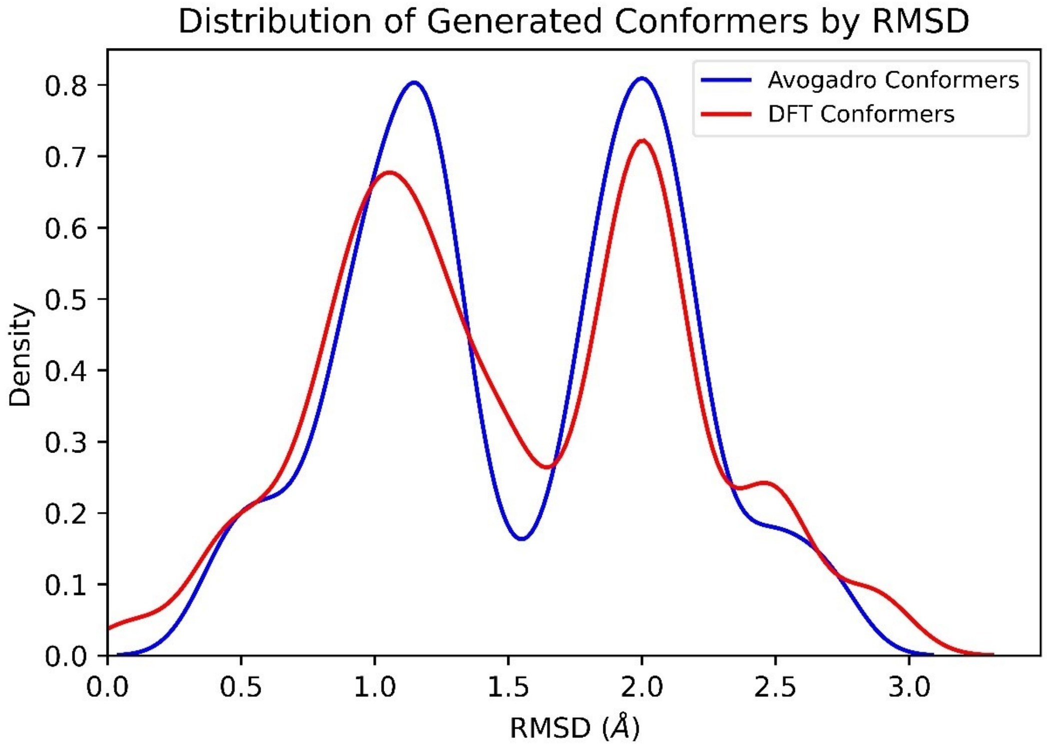

Figure S4: RMSD of conformers generated *via* DFT inputs and Avogadro inputs. Conformers are compared to the lowest energy conformer for each carbene. The figure shows that a wide range of conformers were generated and that conformers generated between DFT and Avogadro inputs show some similarity.

2. Reference DFT Calculations – Methods and Input

All calculations were carried out in ORCA Version 5.0.4

Functional: ωB97X

Basis Set: Def2-SVP

Auxiliary Basis Sets: Def2/J

Dispersion Correction: D4

Implicit Solvation: CPCM(Water)

Keywords: SlowConv TightSCF DefGrid2 RIJCOSX

3. Dataset sizes

Table S3: Table detailing the number of distinct data points within each dataset. Each datapoint refers to a single energy for a single structure. Gas datasets were obtained with charged and uncharged structures (Z=1 and Z=0 respectively).

Dataset	Number of Data Points
Gas_Z=0	306951
Gas_Z=1	488132
Aq_Z=0	301886
Aq_Z=1	403330

4. Trained Models

Trained models for structures in the gaseous/aqueous phase and for charged/uncharged species at varying batch sizes. Loaded through code available at https://github.com/Tuttlelab/DeepSolv

5. ANAKIN-ME (ANI) Neural Network Potential Architecture

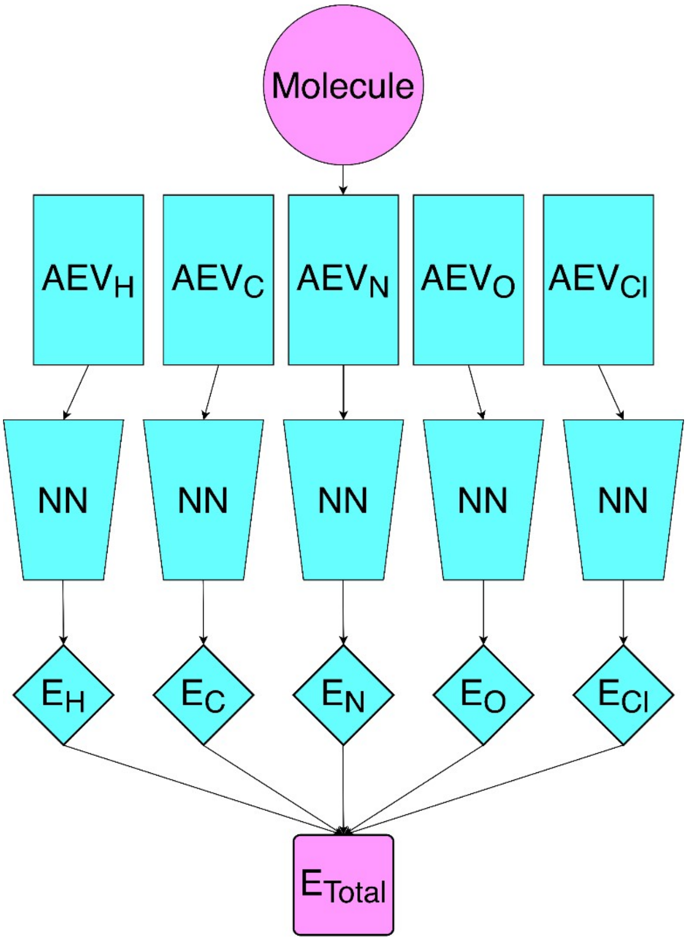

Figure S5: Diagram depicting an ANI family Neural Network Potential for atoms H, C, N, O, and Cl.

ANI models first take an input molecule and split it up into it’s constituent atoms. Each atom is then sent for Atomic Environment Vector (AEV_X in Figure 1 where X is an element) processing based on the element. AEVs encode an atom’s local environment up to a cutoff point (typically around 5 angstroms) by numerically encoding the radial component of each atom and the angular component of each atomic number pair. This captures the atom’s explicit local environment including other atoms. Nothing past the cutoff point is considered and as such ANI is unable to capture long range interaction like the Coulomb potential or non-local effects within the molecule. A more in depth look at AEV construction can be found in the original ANAKIN-ME paper.[1]

The AEVs are the passed through neural networks. There is an individual network for each element. The output of each network is an atomic energy (Figure 1, E_X, where X is the element type) of which one is produced for each atom in the input molecule. The molecular energy is then obtained by summing the atomic energies.

6. Scanning ANI2x Dataset for Carbenes

Scanned ANI-2x ωB97X/6-31G(d) subset for carbenes (available at https://zenodo.org/records/10108942)

Full subset = 9,651,712 structures

Conditions used to narrow search (pseudo code):

```
If carbon not in molecule:     (Filtered down to 9274087 structures)
      Continue
If carbon has > 2 bonds:     (Filtered down to 36299 structures)
      Continue
For angle on carbon:     (Filtered down to 1524 structures)
If (angle < 145 and angle > 115) or (angle < 92 and angle > 112):
      Write molecule to XYZ
```

The final 1524 XYZ files that were checked by hand for carbenes.

7. File Archive

Relevant files and data underpinning this publication are openly available from the University of Strathclyde KnowledgeBase at:

All code relevant to this publication can be accessed openly at https://github.com/Tuttlelab/DeepSolv.

Table S4: Files included as supplementary information to complement publication. Contains details of file types and the data included.

Entry	Description	Data Type	File Path
1	Output files resulting from optimization and frequency DFT calculations at the ωB97X/Def2-SVP level of theory. Outputs are organized based on phase and charge state.	Orca output	Reference_DFT_Level_Data.zip/Reference_DFT_Level_Data_SVP/**/
2	Output files resulting from optimization and frequency DFT calculations at the ωB97X/Def2-TZVP level of theory. Outputs are organized based on phase and charge state.	Orca output	Reference_DFT_Level_Data.zip/Reference_DFT_Level_Data_TZVP/**
3	XYZ Coordinate files for 11 test carbenes in the protonated and deprotonated state. These were drawn in Avogadro and optimised with the UFF	XYZ file	Reference_Inputs_and_Structures.zip/Avogadro_molecules/
4	Orca output files as a result of optimization and frequency jobs completed on each conformer generated by Crest using geometries obtained from optimised DFT calculations	Orca output	Reference_Inputs_and_Structures.zip/DFT_Crest_mols_outputs/
5	Crest output conformers. These	XYZ file	Reference_Inputs_and_Structures.zip/Crest_Molecules/

	are split by method, either Avogadro or DFT. These are further sorted by test carbene, charge state and solvent phase		
6	Structure used as input for crest conformer generation. These are split by method, Avogadro or DFT. They are then further split by test carbene, charge state and solvent phase.	XYZ file	Reference_Inputs_and_Structures.zip/Crest_Molecules/
7	Trained model checkpoint used in publication. This file represents the best checkpoint during training according to MSE loss function. These are split by Solvent Phase, Charge state and basis set size.	.pt file	Trained_Models.zip/Models/**/best.pt
8	Trained model checkpoint used in publication. This file represents the best checkpoint during training according to MAE loss function. These are organized by Solvent Phase, Charge state and batch size.	.pt file	Trained_Models.zip/Models/**/best_L1.pt
9	Atomic self energies as determined by each individual model prior to training. These are organized by solvent phase, charge state and batch size.	csv file (comma separated)	Trained_Models.zip/Models/**/Self_Energies.csv
10	JSON file containing all hyperparameters	.json file	Trained_Models.zip/Models/**/training_config.json

	for training of ML model. These are organized by solvent phase, charge state and batch size		
11	HDF5 format file containing the training data for each model type.	.h5 file	Datasets.zip/Datasets/
12	Orca output files as a result of optimization and frequency jobs completed on butylamine conformers generated by Crest using geometries obtained from optimised DFT calculations	Orca output	Additional_Data.zip/Additional_Data/Butylamine/DFT_Crest_mols_outputs/*.out
13	XYZ Coordinate files for butylamine in the protonated and deprotonated state. The structure was drawn in Avogadro and optimised with the UFF	XYZ file	Additional_Data.zip/Additional_Data/Butylamine/Crest_Avogadro_Confs/**/*.xyz
14	Output files resulting from optimization and frequency DFT calculations at the ωB97X/Def2-SVP level of theory. Outputs are organized based on phase and charge state.	Orca output	Additional_Data.zip/Additional_Data/Butylamine/DFT_GM_Calcs/**/*.out

References

(1) Smith, J. S.; Isayev, O.; Roitberg, A. E. ANI-1: an extensible neural network potential with DFT accuracy at force field computational cost. *Chem Sci* **2017**, *8* (4), 3192-3203. DOI: 10.1039/c6sc05720a From NLM PubMed-not-MEDLINE.