

Approaching coupled cluster accuracy with a general-purpose neural network potential through transfer learning

Justin S. Smith^{1,2,3,†}, Benjamin T. Nebgen^{2,4,†}, Roman Zubatyuk^{2,5,†}, Nicholas Lubbers^{2,3}, Christian Devereux¹, Kipton Barros², Sergei Tretiak^{2,4,}, Olexandr Isayev^{6,*}, Adrian E. Roitberg^{1,*}*

¹ Department of Chemistry, University of Florida, Gainesville, FL 32611, USA

² Theoretical Division, Los Alamos National Laboratory, Los Alamos, NM 87545, USA

³ Center for Nonlinear Studies, Los Alamos National Laboratory, Los Alamos, NM 87545, USA

⁴ Center for Integrated Nanotechnologies, Los Alamos National Laboratory, Los Alamos, NM 87545, USA

⁵ Department of Chemistry, Physics, and Atmospheric Science, Jackson State University, Jackson, MS 39217, USA

⁶ UNC Eshelman School of Pharmacy, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, USA

†These authors contributed equally to this work.

* Corresponding authors; email: ST (serg@lang.gov), OI (olexandr@olexandrisayev.com) and AER (roitberg@ufl.edu)

ABSTRACT

Computational modeling of chemical and biological systems at atomic resolution is a crucial tool in the chemist's toolset. The use of computer simulations requires a balance between cost and accuracy: quantum-mechanical methods provide high accuracy but are computationally expensive and scale poorly to large systems, while classical force fields are cheap and scalable, but lack transferability to new systems. Machine learning can be used to achieve the best of both approaches. Here we train a general-purpose neural network potential (ANI-1ccx) that approaches CCSD(T)/CBS accuracy on benchmarks for reaction thermochemistry, isomerization, and drug-like molecular torsions. This is achieved by training a network to DFT data then using transfer learning techniques to retrain on a dataset of gold standard QM calculations (CCSD(T)/CBS) that optimally spans chemical space. The resulting potential is broadly applicable to materials science, biology and chemistry, and billions of times faster than CCSD(T)/CBS calculations.

Introduction

The central questions in modern chemistry relate to the identification and synthesis of molecules for useful applications. Historically, discoveries have often been serendipitous, driven by a combination of intuition and experimental trial and error.^{1,2} In the modern age, the computer revolution has brought about powerful computational methods based on quantum mechanics (QM) to create a new paradigm for chemistry research.^{3,4} At great computational expense, these methods can provide accurate chemical properties (e.g. energies, forces, structures, reactivity, etc.) for a wide range of molecular systems. Coupled cluster theory systematically approaches the exact solution to the Schrödinger equation, and is considered a gold standard for many quantum chemistry applications.^{5–7} When CCSD(T) (coupled cluster considering single, double, and perturbative triple excitations) calculations are combined with an extrapolation to the complete basis set limit (CBS)^{8,9}, even the hardest to predict non-covalent and intermolecular interactions can be computed quantitatively.¹⁰ However, coupled cluster theory at the level of CCSD(T)/CBS is computationally expensive, and often impractical for systems with more than a dozen atoms.

Since the computational cost of highly accurate QM methods can be impractical, researchers often seek to trade accuracy for speed. Density functional theory (DFT)^{11–13}, perhaps the most popular QM method, is much faster than coupled cluster theory. In practice, however, DFT requires empirical selection of a density functional, and so DFT-computed properties are not as reliable and objective as coupled cluster techniques at guiding experimental science. Even stronger approximations can be made to achieve better efficiency. For example, classical force fields are commonly employed to enable large scale dynamical simulation such as protein folding¹⁴, ligand-protein docking¹⁵ or the dynamics of dislocations in materials¹⁶. These models are often fragile; a force field fit to one system may not accurately model other systems¹⁷. An outstanding challenge is to simultaneously capture a great diversity of chemical processes with a single linear scaling model potential.

Machine learning (ML) methods have seen much success in the last decade due to increased availability of data and improved algorithms.^{18–20} Applications of ML are becoming increasingly common in experimental and computational chemistry. Recent chemistry related work reports on ML models for chemical

reactions^{21,22}, potential energy surfaces^{23–27}, forces^{28–30}, atomization energies^{31–33}, atomic partial charges^{32,34–36}, molecular dipoles^{26,37,38}, materials discovery^{39–41}, and protein-ligand complex scoring⁴².

Many of these studies represent important and continued progress toward ML models of quantum chemistry that are transferable (*i.e.* applicable to related, but new chemical processes) and extensible (*i.e.* accurate when applied to larger systems). These advances aim to revolutionize chemistry through applications to chemical and biological systems. Since molecular dynamics simulations underpin much of computational chemistry and biology, transferable, accurate, and fast prediction of molecular energies and forces is particularly important for the next generation of linear scaling model potential energy surfaces.

Transferable and extensible ML potentials often require training on very large datasets. One such approach is the ANI class of methods. The ANI-1 potential aims to work broadly for molecules in organic chemistry⁴³.

A key component of this potential is the ANI-1 dataset, which consists of DFT energies for 22M randomly selected molecular conformations from 57k distinct small molecules⁴⁴. This vast amount of data would be impractical to generate at a level of theory more accurate than DFT⁴⁵. However, advances in machine learning methodologies are greatly reducing the required dataset sizes. The ANI-1x dataset, constructed using active learning, contains DFT data for 5M conformations of molecules with an average size of 15 atoms²⁵. Active learning iteratively adds new QM calculations to the dataset for specific cases where the current ML model cannot make a good prediction. Despite the much smaller size of the ANI-1x dataset, potentials trained on it vastly outperform those trained on the ANI-1 dataset, especially on transferability and extensibility benchmarks. Even with the success of the ANI-1x potential, its true accuracy is still reliant upon the accuracy of the underlying DFT data.

A remaining challenge is to develop ML-based potentials that reach coupled cluster-level accuracy while retaining transferability and extensibility over a broad chemical space. The difficulty is that datasets with CCSD(T)-level accuracy are very expensive to construct and therefore tend to be limited in chemical diversity. Previous studies have trained on high-quality QM data for small molecules at equilibrium conformations^{46,47} and for non-equilibrium conformations of a single molecule⁴⁸. A limitation is that ML

models trained on datasets which lack chemical diversity are not expected to be transferable or extensible to new systems. The present work uses transfer learning^{49,50} to train an ML potential that is accurate, transferable, extensible, and therefore, broadly applicable. In transfer learning, one begins with a model trained on data from one task and then retrains the model on data from a different, but related task, often yielding high accuracy predictions^{51–53} even when data is sparsely available. In our application, we begin by training a neural network on a large quantity of lower-accuracy DFT data (the ANI-1x dataset with 5M non-equilibrium molecular conformations²⁵), and then we retrain to a much smaller dataset (about 500k intelligently selected conformations from ANI-1x) at the CCSD(T)/CBS level of accuracy. Such a high quality and diverse dataset is a first of its kind for training machine learning based potentials. The resulting general-purpose potential, ANI-1ccx, and dataset exceeds the accuracy of DFT in benchmarks for isomerization energies, reaction energies, molecular torsion profiles and energies and forces at non-equilibrium geometries, while being roughly nine orders of magnitude faster than DFT. The ANI-1ccx potential is available on GitHub (https://github.com/isayev/ASE_ANI) as a user-friendly Python interface integrated with the Atomic Simulation Environment⁵⁴ package (ASE; <https://wiki.fysik.dtu.dk/ase/>).

Results

Relative Conformer Energy

We compare the errors of ANI-1ccx (trained with transfer learning), ANI-1x (trained on DFT data only), and direct DFT calculations (ω B97X/6-31g*). We also compare to a model, ANI-1ccx-R, that was trained only with the CCSD(T)*/CBS data, i.e., without transfer learning from the DFT data. CCSD(T)*/CBS is a highly accurate extrapolation to high level QM. For details see the methods section. To test transferability and extensibility, we employ four benchmarks to appraise the accuracy of molecular energies and forces, reaction thermochemistry, and the computation of torsional profiles on systems consisting of CHNO. The GDB-10to13 benchmark²⁵ is designed to evaluate relative energies, atomization energies, and force calculations on a random sample of 2,996 molecules containing 10 to 13 C, N, or O atoms (with H added to saturate the molecules). The GDB-10to13 molecules are randomly perturbed along their normal modes

to produce between 12 and 24 non-equilibrium conformations per molecule. HC7/11⁵⁵ is a benchmark designed to gauge the accuracy of hydrocarbon reaction and isomerization energies. The ISOL6

Table 1 Accuracy in predicting conformer energy differences on the GDB-10to13 benchmark

	ANI-1ccx	ANI-1ccx-R	ANI-1x	ω B97X
MAD*	1.46	1.81	1.97	1.42
RMSD*	2.07	2.54	2.79	2.04

*Units are in kcal*mol^-1.

benchmark⁵⁶ (a subset of the ISOL24/11 benchmark) measures isomerization energies for organic molecules. Finally, we test on the Genentech torsion benchmark⁵⁷, which contains 62 diverse organic molecule torsion profiles (45 containing only CHNO).

Table 1 provides mean absolute deviations (MAD) and root mean squared deviations (RMSD) for the ANI potentials and ω B97X/6-31g*, on the GDB-10to13 benchmark from the COMP6²⁵ benchmark suite. Reference values are recomputed at the CCSD(T)*/CBS level of theory. Table 1 only considers conformations within 100kcal*mol^-1 of the energy minima for each molecule. The conformational energy ΔE is the energy difference between all conformers for a given molecule in the benchmark.²⁵ Methods compared are the ANI-1ccx transfer learning potential, ANI-1ccx-R trained only on coupled cluster data, ANI-1x trained only on DFT data, and the DFT reference (ω B97X). Our analysis concludes that training a model only to the smaller CCSD(T)*/CBS dataset (ANI-1ccx-R) results in a 23% degradation in RMSD compared to the transfer learning model (ANI-1ccx). The DFT trained ANI-1x model has a 36% increase in RMSD over ANI-1ccx. ANI-1ccx performs as well as the original reference (ω B97X/6-31G*) in the 100kcal*mol^-1 energy range on the GDB-10to13 CCSD(T)*/CBS benchmark. Recall that each ANI model is an ensemble average over 8 neural networks. Without an ensemble of networks, the MAD and RMSD of ANI models degrades by about 25%.²⁵ Supplementary Table 5 provides errors for all methods within the full energy range of the GDB-10to13 benchmark. Notably, ANI-1ccx outperforms DFT with an RMSD of 3.2 kcal*mol^-1 vs. 5.0 kcal*mol^-1 for DFT, which means the ANI-1ccx model generalizes

better to high energy conformations than ω B97X/6-31G*. Supplementary Figure 3 shows correlation plots for the ANI models vs. CCSD(T)*/CBS.

Atomization Energy

Figure 1 displays a comparison of atomization energy deviation from reference CCSD(T)*/CBS for DFT (blue) and ANI-1ccx (orange) for all conformations in GDB-10to13 within 100 kcal \cdot mol $^{-1}$ of the conformational minima. Compared to the DFT functional, the ANI-1ccx potential provides a more accurate prediction of the CCSD(T)*/CBS atomization energy. The distribution for ANI-1ccx has a standard deviation of 2.3 kcal \cdot mol $^{-1}$, while the DFT distribution is much wider, with a standard deviation of 6.3 kcal \cdot mol $^{-1}$. The MAD/RMSD for DFT vs. reference CCSD(T)*/CBS is 15.9/17.1 kcal \cdot mol $^{-1}$, while for ANI-1ccx it is 1.9/2.5 kcal \cdot mol $^{-1}$. Supplementary Figure 4 shows an attempt to correct the systematic shift of the DFT model to the reference CCSD(T)*/CBS atomization energies via a linear fitting of the atomic elements in each system. Even after this non-trivial correction, ANI-1ccx is still more accurate than DFT vs. the more accurate coupled cluster atomization energies. The corrected DFT has a distribution with a standard deviation of 5.5 kcal \cdot mol $^{-1}$ with MAD/RMSD of 4.9/5.9 kcal \cdot mol $^{-1}$.

Forces

Accurate forces are important for MD simulations and geometry optimization. Therefore, we explicitly assess force accuracy as well. It is impractical to obtain forces with the CCSD(T)*/CBS extrapolation due to extreme computational expense with existing packages. However, MP2/cc-pVTZ (dubbed here as MP2/TZ) provides a high-quality alternative. Table 2 compares MP2/TZ force calculations on the GDB-10to13 benchmark to MP2/cc-pVDZ (MP2/DZ), ω B97X/6-31G*, ANI-1x, and ANI-1ccx models. ANI-1ccx provides the best prediction of MP2/TZ forces compared to all other methods. Notably, ANI-1ccx forces deviate less from the MP2/TZ target forces than the original ANI-1x DFT trained potential, providing evidence that the transfer learning process not only corrects energies but forces as well. Supplementary Figure 5 also shows a comparison between ANI-1ccx and experimental results for C-C center of mass radial distribution functions for cyclohexane.

Reaction and Isomerization Energy

The HC7/11 and ISOL6 benchmarks address the calculation of reaction and isomerization energies and are depicted in Figure 2. For each reaction, reference energies and calculated energies are provided in Supplementary Tables 7 and 8. Figure 2 shows the differences between the computed and the reference energies, for the reaction

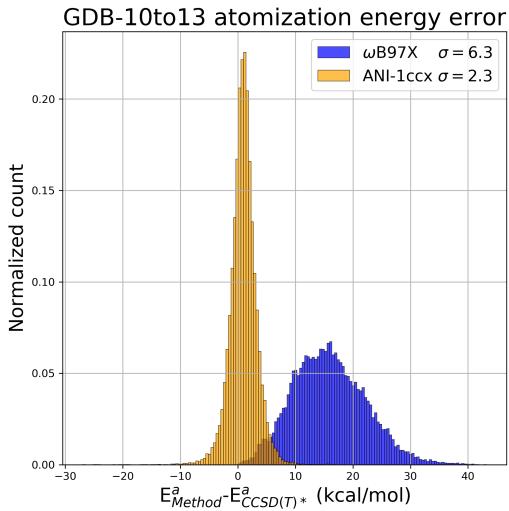


Figure 1: Accuracy in predicting atomization energies. Error of the ANI-1ccx predicted atomization energy E^a on the GDB-10to13 benchmark relative to CCSD(T)*/CBS and compared against ω B97X.

Table 2 Accuracy for calculating atomic forces on the GDB-10to13 benchmark

	ANI-1ccx	ANI-1x	ω B97X	MP2/DZ
MP2/TZ	3.4/5.3*	4.7/7.1*	3.7/5.9*	4.6/5.9*

MAE/RMSE in $\text{kcal}^\text{mol}^{-1}\text{\AA}^{-1}$

and isomerization energies individually for ω B97X/6-31g*, ANI-1x, ANI-1ccx, and our CCSD(T)*/CBS. HC7/11 used target MP2/6-311+G(2df,2p) and ISOL6 used target CCSD(T)-F12a/aug-cc-pVDZ calculations. The latter is the most accurate simulation method currently available. As was done in the original benchmarks, single point energy calculations using all ANI models, ω B97X, and CCSD(T)*/CBS were performed on the original benchmark structures. These energies were used to calculate the reaction energies. For the HC7/11 benchmark, the medium-sized basis DFT reference ω B97X/6-31g* is not sufficient for describing the chemistry represented in these complex hydro-carbon reactions. Likewise, ANI-1x, trained to data from this functional, closely mirrors the behavior of DFT. Similarly, the transfer learning-based

ANI-1ccx model tends to mirror its CCSD(T)*/CBS reference calculations and substantially outperforms DFT compared to the target reaction energies. Overall MAD/RMSD on the HC7/11 benchmark for DFT, ANI-1x, ANI-1ccx, and CCSD(T)*/CBS are 16.4/22.2, 19.1/24.6, 2.5/2.9 and 1.6/1.8 kcal^{*}mol⁻¹, respectively. These results are shown in Supplementary Table 6

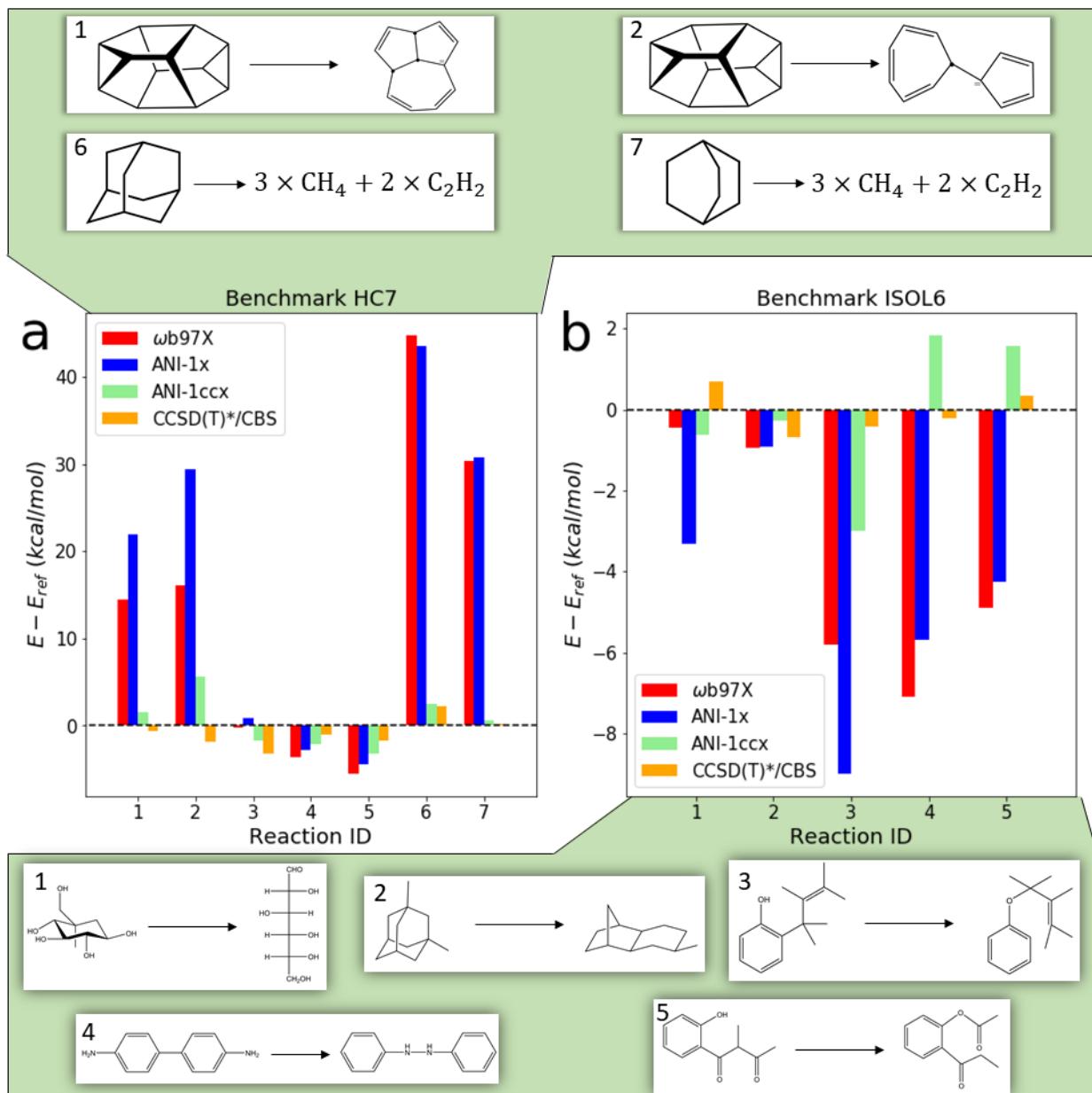


Figure 2: Accuracy in predicting reaction and isomerization energy. ANI-1ccx reaction and isomerization energy difference prediction on the **(a)** HC7/11 and **(b)** ISOL6 benchmarks, relative to CCSD(T)/CBS.

Methods compared are the ANI-1ccx transfer learning potential, ANI-1x trained only on DFT data, the DFT reference (ω B97X), and our coupled cluster extrapolation scheme CCSD(T)*/CBS. The top panel provides the HC7 reactions numbered 1, 2, 6, and 7 and bottom panel shows the ISOL6 reactions numbered 1-5.

Figure 2b displays a similar comparison for the five medium sized organic C, H, N, O containing molecules of the ISOL6⁵⁶ isomerization energy benchmark. A similar trend is seen in this case as with the HC7/11 benchmark, where ANI-1x deviations tend to correlate with the large deviations of its reference DFT. The prediction error for the transfer learning-based model is greatly reduced compared to the DFT trained model and DFT itself. For the ISOL6 reactions shown in Figure 2b, overall MAD/RMSD for DFT, ANI-1x, ANI-1ccx, and CCSD(T)*/CBS are 3.8/4.7, 4.6/5.3, 1.5/1.8 and 0.5/0.5 kcal*mol⁻¹, respectively.

Molecular Torsions

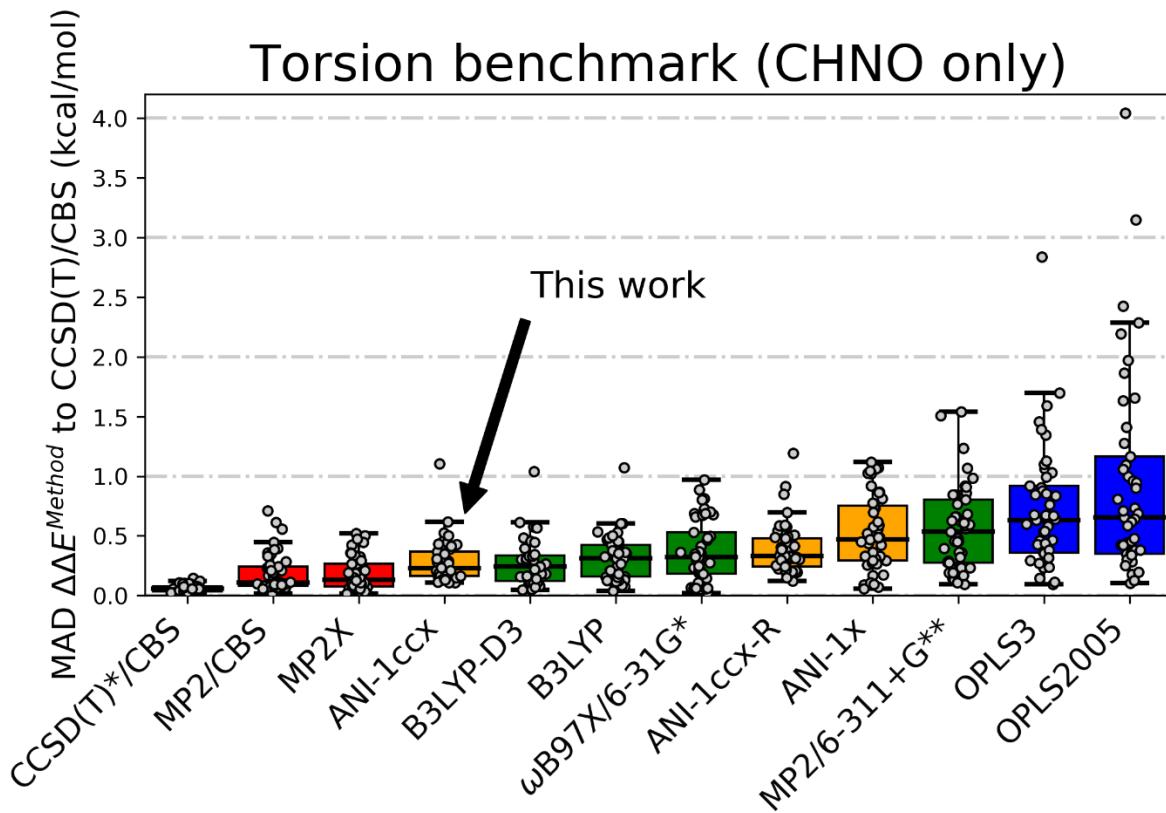


Figure 3: Accuracy in predicting torsional energies relevant to drug discovery. Methods compared are QM (red and green), molecular mechanics (blue) and ANI (orange) performance on 45 torsion profiles containing C, H, N and O atomic elements. The grey dots represent the MAD of a given torsion scan vs.

gold standard CCSD(T)/CBS. The box extends from the upper to lower quartile and the black horizontal line in the box is the median. The upper “whisker” extends to the last datum less than the third quartile plus 1.5 times the interquartile range while the lower “whisker” extends to the first datum greater than the first quartile minus 1.5 times the interquartile range.

Molecular torsions play an import role in computational drug discovery (e.g. in screening ligands for favorable protein binding) and in modeling the assembly of soft materials. Therefore, we compare the new ANI-1ccx transfer learning-based potential against various QM and molecular mechanics (MM) based methods from the molecular torsion benchmark of Sellers et. al.⁵⁷ This benchmark provides a measure of accuracy for a model at reproducing potential energy profiles from a diverse set of molecular torsions of small organic molecules containing the atoms C, H, N and O. These torsions are representative torsions typically found in small drug like molecules.

Figure 3 provides a comparison of results for three highly accurate but computationally expensive QM methods, four moderately computationally expensive QM methods, and two commonly used small molecule force fields. These data were obtained from Sellers et. al.⁵⁷ We also add the ANI potentials (ANI-1ccx, ANI-1ccx-R, and ANI-1x) used in this work, as well as CCSD(T)*/CBS reference energy calculations. Other semi-empirical QM and MM methods studied in Sellers et al. are left out of this comparison since each one performed worse than OPLS2005 on the benchmark. Each torsion in the benchmark is generated through a restrained optimization, where the torsional degree of freedom is fixed every 10 degrees, and the remaining degrees of freedom are relaxed through an optimization process. The red boxes in Figure 3 represent QM methods (first three from the left) that are so computationally intense the MP2 restrained optimized structures were used and single point calculations were performed for that method. The green QM methods were all optimized using their own forces. In an ideal setting, all methods would provide their own structures for the final energy calculation, though, such an exercise would be prohibitively computationally expensive for the more rigorous QM methods. The ANI and MM models also carried out restrained optimizations using their own forces. The MP2 structures are not used here

because the usefulness of such efficient methods can only be gauged without the assistance of less efficient QM methods. The ANI-1x potential, trained to the ANI-1x DFT dataset plus active learning-based dihedral corrections, obtains a median MAD of 0.47 kcal^{*}mol⁻¹ on the benchmark. The ANI-1x potential performs similarly to MP2/6-311+G** and to the ANI-1ccx-R potential. The DFT trained ANI-1x also outperforms OPLS3, one of the most accurate and widely used small molecule force-fields available. Further, the transfer learning-based ANI-1ccx potential achieves a median MAD of 0.23 kcal^{*}mol⁻¹, a 51% reduction in error over ANI-1x vs. the CCSD(T)/CBS target. ANI-1ccx exceeds the performance of all DFT (B3LYP-D3/6-311+G**, B3LYP/6-311+G**, and ωB97X/6-31g*) methods utilized in this study, approaching the accuracy of higher-level, and costlier, ab initio QM methods (MP2/CBS and MP2.X/CBS). The ANI-1ccx potential achieves these prediction accuracies without an increase in computational cost over the original ANI-1x potential. Results for ANI-1x and ANI-1ccx before and after active learning for dihedral reparameterization can be found in Supplementary Figure 1. The dihedral scans used to compute ANI-1ccx's error can be found in Supplementary Figure 6. Each ANI-1ccx restrained optimization (averaged over the 36 angles for each of the 45 torsions) took approximately 0.58s on a single NVIDIA V100 GPU. A similar timing comparison was reported in Sellers et al.⁵⁷ for the QM and MM methods. Comparing to this literature result, the ANI model on a single GPU is (on average) as fast as OPLS3 on a CPU and 6200 times faster than B3LYP-D3 on a CPU. While a GPU to CPU comparison with the use of different optimization methods is not exactly a fair comparison, it does provide a sense of the computational affordability of the ANI potential. Moreover, the ANI potential scales more easily than QM, exhibiting linear scaling (compare to O(N³) for the most efficient QM) on large systems and a smaller pre-factor and better memory scaling.

Discussion

Great progress has been made in creating faster and more accurate QM methods, but even in modern computer architectures the cost involved in the improved accuracy becomes prohibitive very quickly. With the advent of machine learning, we can and must make the leap to modern statistical and data-driven approaches, which have the potential to drive rapid progress in drug and materials design as well as

applications to natural systems such as proteins. The ANI-1ccx potential (available at https://github.com/isayev/ASE_ANI) presented in this work is an attractive alternative to density functional theory approaches and standard force fields for conformational searches, molecular dynamics, and the calculation of reaction energies. The availability of high-quality QM reference data, produced with a new extrapolation scheme to CCSD(T)/CBS, allowed us to use transfer learning techniques to build a chemically accurate universal ANI potential. Accuracy benchmarks show that the transfer learning-based ANI-1ccx outperforms DFT on test cases where DFT fails to accurately describe reaction thermochemistry and on small molecule torsion benchmarks. After extensive benchmarking, we conclude that ANI-1ccx captures a broad range of organic chemistry, with accuracy comparable to QM calculations at the coupled-cluster level of theory. Comparisons between transfer learning and naïve training to only the small dataset of high quality QM calculations show that transfer learning is a superior approach. As such this work offers a computationally efficient and accurate ML-based molecular potential for general use across a broad range of chemical systems.

Future work will aim to validate and retrain (if necessary) the ANI-1ccx potential for applications in condensed phase simulation. For smaller molecular systems, the ANI-1ccx potential is an accurate and efficient alternative to expensive QM methods and might find indirect ways to become applicable in such condensed phase simulation, e.g. using ANI-1ccx to parametrize force fields for condensed phase simulation. As with any model, ani-1ccx has limitations. Some of them can be overcome by adding more data and through active learning methods and retraining. This category includes new and different chemical environments, intermolecular interactions, ions, new atomic elements and reactions. There is a set of limitations that would require the development of new theory and methods, for instance for recovering long range interactions through the addition of coulomb interactions, to treat multiple electronic excited states or radicals.

Methods

An efficient and accurate CCSD(T)/CBS approximation

Recalculating even 10% of the ANI-1x dataset (*i.e.*, 500k molecules) with conventional CCSD(T)/CBS would require enormous computational resources. Therefore, we developed an approximation scheme (herein referred to as CCSD(T)*/CBS) that allows highly accurate energy calculations in a high-throughput fashion.

Our CCSD(T)*/CBS method is a computationally efficient approximation of CCSD(T)/CBS energies that takes advantage of the linear-scaling domain-localized DPLNO-CCSD(T) method developed by Neese *et al*⁵⁸ which is implemented in the ORCA software package⁵⁹. It provides an affordable alternative capable of achieving near CCSD(T) accuracy at a fraction of the computational cost. The DLPNO approximation relies on the MP2 method to estimate energy contributions from interacting electron pairs and effectively reduce the active orbital space. Table 1 provides accuracy and timing benchmarks, clearly

Table 3 Computational cost and accuracy of our coupled cluster approximation

	CPU-core hours*		Mean absolute deviation from CCSD(T)-F12 (kcal*mol- 1kcal*mol^-1)	
	Alanine (13 atoms)	Aspirin (21 atoms)	S66	W4-11
CCSD(T)/CBS	9.13	427.00	0.03	1.31
CCSD(T)*/CBS (this work)	1.44	7.44	0.09	1.46

* All calculations are performed on an Intel Xeon E5-2630 v3 @ 2.40GHz CPU

showing our CCSD(T)*/CBS approximation provides accurate energies vs. the CCSD(T)-F12 level of theory⁶⁰ in a computationally efficient way. S66 and W4-11 are standard benchmarks for interaction and atomization energies of small molecules^{61,62}. See Supplementary Table 1 for a more detailed comparison. Details of our CCSD(T)*/CBS scheme plus additional benchmarks are given in supplemental information Section S1.1.

Using active learning for CCSD(T)*/CBS dataset curation

The existing ANI-1x active learning generated dataset²⁵ is used to train an initial DFT (to the ωB97X/6-31G* model chemistry⁶³) potential, likewise dubbed ANI-1x. The ANI-1x dataset consists of 5M

conformations from 64k small molecules and complexes of molecules containing only CHNO atoms. All model and training procedures are detailed in the ANI-1 work.⁴³ Section S1.2 provides details of the architecture, selection of hyperparameters, and held out test set errors. To reduce variance and increase accuracy, all ANI results presented in this work are the ensemble prediction of 8 ANI neural networks, i.e. the ANI-1x potential used in this work is an ensemble of 8 ANI-1x neural networks trained to different splits of the ANI-1x dataset and the ANI-1ccx network is built by transfer learning from the eight ANI-1x networks.²⁵ The disagreement between predictions of ensemble members can be used as a proxy to the prediction error, enabling rapid identification of molecular conformations where the current ANI model fails.

Despite the efficiency of our CCSD(T)*/CBS extrapolation scheme, optimal curation of the coupled cluster dataset is still essential since we can only perform a limited number of these calculations. As a source of structures for CCSD(T)*/CBS data generation, we choose to systematically subsample the existing ANI-

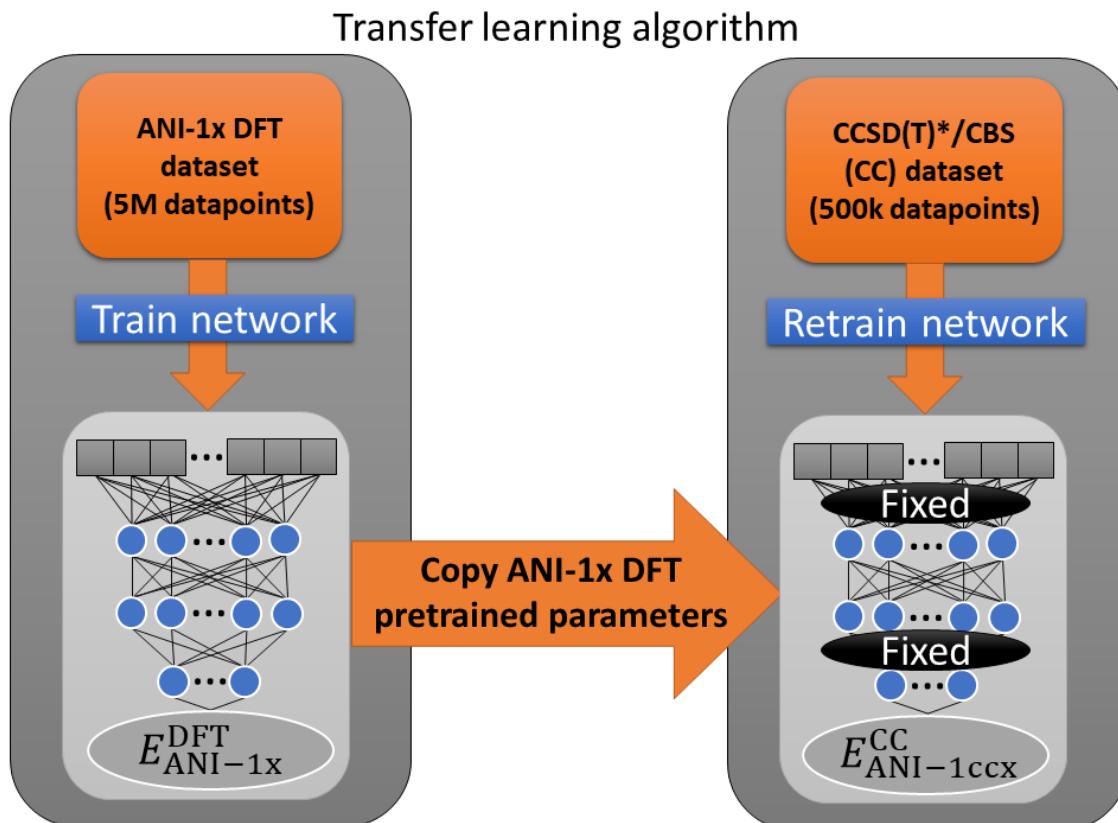


Figure 4: Diagram of the transfer learning technique evaluated in this work. Transfer learning starts from a pretrained ANI-1x DFT model, then retrains to higher accuracy CCSD(T)*/CBS data with some parameters fixed during training.

1x dataset with 5M molecules, since this dataset already provides a pool of highly diverse molecular configurations and conformations. We begin with an initial random subsample of 200k data points, then iteratively we select new data for coupled cluster calculations according to maximal ensemble disagreement (*i.e.*, query by committee⁶⁴). Through 3 iterations of coupled cluster data generation using active learning, we grow the coupled cluster dataset to about 480k molecules. To further improve the ANI potential’s description of torsion profiles, we also perform 20 iterations of active learning²⁵ on random molecular torsions from small and druglike molecules to enhance ANI-1x with about 200k new DFT calculations. The ANI driven torsion sampling technique is detailed in Section S1.3. Of these torsion conformations, we randomly select 10% for CCSD(T)*/CBS calculations. The result is an enhanced ANI-1x DFT dataset containing 5.2M datapoints and a high-accuracy CCSD(T)*/CBS dataset containing about 500k datapoints.

Training to high-accuracy data using transfer learning

Here we describe the transfer learning methodology (depicted schematically in Figure 4) used to create ANI-1ccx. First, an ANI potential is trained to the DFT dataset with the new active learning torsion data added, yielding a potential equivalent to the ANI-1x potential²⁵. Note, a single ANI potential is composed of multiple ANI neural network models. We then retrain each ANI-1x model to the CCSD(T)*/CBS data with 65,280 of the 325,248 optimizable neural network parameters held constant for each ANI model in the ensemble. Training a single ANI model to the original 5.2 million molecule dataset takes approximately 4 hours on a NVIDIA Titan V GPU, while retraining to the 500k molecule CCSD(T)*/CBS dataset takes around 30 minutes. Neural network parameters are organized into a set of hidden layers. The ANI models trained in this work contain 4 hidden layers; we leave two hidden layers to be optimized during the transfer learning process, while the other two layers are left fixed to reduce the number of optimizable parameters during the training process and thus avoid overfitting to the smaller CCSD(T)*/CBS dataset. Details of

ANI-1ccx's performance on its test set are given in Supplementary Table 3. An alternative to transfer learning is Δ -learning⁴⁶. With Δ -learning, one trains a new model to correct for the difference between CCSD(T)/CBS and the existing model pretrained on DFT data. Although Δ -learning yields similar accuracy to transfer learning, it needs to evaluate the neural networks twice to make inferences. More information on Δ -learning and its accuracy is provided in Supplementary Figure 2 and Supplementary Table 4.

Data availability

All relevant data are available from the authors upon reasonable requests.

Code availability

All code needed to run this model can be found at https://github.com/isayev/ASE_ANI

References

1. Ramsay, O. B. Serendipity: Accidental Discoveries in Science (Roberts, Royston M.). *J. Chem. Educ.* **67**, A311 (2009).
2. Berson, J. A. Discoveries missed, discoveries made: creativity, influence, and fame in chemistry. *Tetrahedron* **48**, 3–17 (1992).
3. Pople, J. A. Quantum Chemical Models (Nobel Lecture). *Angew. Chemie Int. Ed.* **38**, 1894–1902 (1999).
4. Kohn, W. Nobel Lecture: Electronic structure of matter—wave functions and density functionals. *Rev. Mod. Phys.* **71**, 1253–1266 (1999).
5. Purvis, G. D. & Bartlett, R. J. A full coupled-cluster singles and doubles model: The inclusion of disconnected triples. *J. Chem. Phys.* **76**, 1910–1918 (1982).
6. Bartlett, R. J. & Musiał, M. Coupled-cluster theory in quantum chemistry. *Rev. Mod. Phys.* **79**, 291–352 (2007).
7. Daniel Crawford, T. & F. Schaefer III, H. An Introduction to Coupled Cluster Theory for Computational Chemists. in *Rev Comp Chem* **14**, 33–136 (2007).
8. Hobza, P. & Šponer, J. Toward true DNA base-stacking energies: MP2, CCSD(T), and complete basis set calculations. *J. Am. Chem. Soc.* **124**, 11802–11808 (2002).
9. Feller, D., Peterson, K. A. & Crawford, T. D. Sources of error in electronic structure calculations on small chemical systems. *J. Chem. Phys.* **124**, 054107 (2006).
10. Řezáč, J., Riley, K. E. & Hobza, P. Extensions of the S66 data set: More accurate interaction energies and angular-displaced nonequilibrium geometries. *J. Chem. Theory Comput.* **7**, 3466–3470 (2011).

11. Grimme, S. Density functional theory with London dispersion corrections. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **1**, 211–228 (2011).
12. Thanthiriwatte, K. S., Hohenstein, E. G., Burns, L. A. & Sherrill, C. D. Assessment of the performance of DFT and DFT-D methods for describing distance dependence of hydrogen-bonded interactions. *J. Chem. Theory Comput.* **7**, 88–96 (2011).
13. Mardirossian, N. & Head-Gordon, M. Thirty years of density functional theory in computational chemistry: An overview and extensive assessment of 200 density functionals. *Mol. Phys.* **115**, 2315–2372 (2017).
14. Dill, K. A. & MacCallum, J. L. The protein-folding problem, 50 years on. *Science* **338**, 1042–1046 (2012).
15. Dror, R. O., Dirks, R. M., Grossman, J. P., Xu, H. & Shaw, D. E. Biomolecular Simulation: A Computational Microscope for Molecular Biology. *Annu. Rev. Biophys.* **41**, 429–452 (2012).
16. Meyers, M. A., Mishra, A. & Benson, D. J. Mechanical properties of nanocrystalline materials. *Progress in Materials Science* **51**, 427–556 (2006).
17. Rauscher, S. *et al.* Structural ensembles of intrinsically disordered proteins depend strongly on force field: A comparison to experiment. *J. Chem. Theory Comput.* **11**, 5513–5524 (2015).
18. Jordan, M. I. & Mitchell, T. M. Machine learning: Trends, perspectives, and prospects. *Science (80-.)* **349**, 255–260 (2015).
19. Gil, Y., Greaves, M., Hendler, J. & Hirsh, H. Amplify scientific discovery with artificial intelligence. *Science (80-.)* **346**, 171–172 (2014).
20. LeCun, Y. A., Bengio, Y. & Hinton, G. E. Deep learning. *Nature* **521**, 436–444 (2015).
21. Ahneman, D. T., Estrada, J. G., Lin, S., Dreher, S. D. & Doyle, A. G. Predicting reaction performance in C–N cross-coupling using machine learning. *Science (80-.)* **360**, 186–190 (2018).
22. Klucznik, T. *et al.* Efficient Syntheses of Diverse, Medicinally Relevant Targets Planned by Computer and Executed in the Laboratory. *Chem* **4**, 522–532 (2018).
23. Schütt, K. T., Arbabzadah, F., Chmiela, S., Müller, K. R. & Tkatchenko, A. Quantum-Chemical Insights from Deep Tensor Neural Networks. *Nat. Commun.* **8**, 13890 (2017).
24. Chmiela, S. *et al.* Machine learning of accurate energy-conserving molecular force fields. *Sci. Adv.* **3**, e1603015 (2017).
25. Smith, J. S., Nebgen, B., Lubbers, N., Isayev, O. & Roitberg, A. E. Less is more: Sampling chemical space with active learning. *J. Chem. Phys.* **148**, (2018).
26. Yao, K., Herr, J. E., Toth, D. W., McIntyre, R. & Parkhill, J. The TensorMol-0.1 Model Chemistry: a Neural Network Augmented with Long-Range Physics. *Chem. Sci.* **9**, 2261–2269 (2017).
27. Behler, J. First Principles Neural Network Potentials for Reactive Simulations of Large Molecular and Condensed Systems. *Angew. Chemie Int. Ed.* **56**, 12828–12840 (2017).
28. Li, Z., Kermode, J. R. & De Vita, A. Molecular dynamics with on-the-fly machine learning of quantum-mechanical forces. *Phys. Rev. Lett.* **114**, 096405 (2015).
29. Glielmo, A., Sollich, P. & De Vita, A. Accurate interatomic force fields via machine learning with

- covariant kernels. *Phys. Rev. B* **95**, (2017).
- 30. Kruglov, I., Sergeev, O., Yanilkin, A. & Oganov, A. R. Energy-free machine learning force field for aluminum. *Sci. Rep.* **7**, (2017).
 - 31. Rupp, M., Tkatchenko, A., Muller, K.-R. & von Lilienfeld, O. A. Fast and accurate modeling of molecular atomization energies with machine learning. *Phys. Rev. Lett.* **108**, 58301 (2012).
 - 32. Faber, F. A. *et al.* Prediction Errors of Molecular Machine Learning Models Lower than Hybrid DFT Error. *J. Chem. Theory Comput.* **13** (11), 5255–5264 (2017).
 - 33. Lubbers, N., Smith, J. S. & Barros, K. Hierarchical modeling of molecular energies using a deep neural network. *J. Chem. Phys.* **148**, (2018).
 - 34. Bleiziffer, P., Schaller, K. & Riniker, S. Machine Learning of Partial Charges Derived from High-Quality Quantum-Mechanical Calculations. *J. Chem. Inf. Model.* **58**, 579–590 (2018).
 - 35. Hermann, J., DiStasio, R. A. & Tkatchenko, A. First-Principles Models for van der Waals Interactions in Molecules and Materials: Concepts, Theory, and Applications. *Chemical Reviews* **117**, 4714–4758 (2017).
 - 36. Nebgen, B. *et al.* Transferable Dynamic Molecular Charge Assignment Using Deep Neural Networks. *J. Chem. Theory Comput.* **14**, 4687–4698 (2018).
 - 37. Sifain, A. E. *et al.* Discovering a Transferable Charge Assignment Model Using Machine Learning. *J. Phys. Chem. Lett.* **9**, 4495–4501 (2018).
 - 38. Gastegger, M., Behler, J. & Marquetand, P. Machine Learning Molecular Dynamics for the Simulation of Infrared Spectra. *Chem. Sci.* **8**, 6924–6935 (2017).
 - 39. Ramprasad, R., Batra, R., Pilania, G., Mannodi-Kanakkithodi, A. & Kim, C. Machine learning in materials informatics: Recent applications and prospects. *npj Comput. Mater.* **3**, 54 (2017).
 - 40. Raccuglia, P. *et al.* Machine-learning-assisted materials discovery using failed experiments. *Nature* **533**, 73–76 (2016).
 - 41. Isayev, O. *et al.* Universal fragment descriptors for predicting properties of inorganic crystals. *Nature Communications* **8**, 15679 (2017).
 - 42. Ragoza, M., Hochuli, J., Idrobo, E., Sunseri, J. & Koes, D. R. Protein-Ligand Scoring with Convolutional Neural Networks. *J. Chem. Inf. Model.* **57**, 942–957 (2017).
 - 43. 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.* **8**, 3192–3203 (2017).
 - 44. Smith, J. S., Isayev, O. & Roitberg, A. E. Data Descriptor: ANI-1, A data set of 20 million calculated off-equilibrium conformations for organic molecules. *Sci. Data* **4**, 170193 (2017).
 - 45. Kranz, J. J., Kubillus, M., Ramakrishnan, R., Von Lilienfeld, O. A. & Elstner, M. Generalized Density-Functional Tight-Binding Repulsive Potentials from Unsupervised Machine Learning. *J. Chem. Theory Comput.* **14**, 2341–2352 (2018).
 - 46. Ramakrishnan, R., Dral, P. O., Rupp, M. & Von Lilienfeld, O. A. Big data meets quantum chemistry approximations: The Δ -machine learning approach. *J. Chem. Theory Comput.* **11**, 2087–2096 (2015).
 - 47. Bartók, A. P. *et al.* Machine learning unifies the modeling of materials and molecules. *Sci. Adv.* **3**,

e1701816 (2017).

48. Chmiela, S., Sauceda, H. E., Müller, K. R. & Tkatchenko, A. Towards exact molecular dynamics simulations with machine-learned force fields. *Nat. Commun.* **9**, Preprint at <https://arxiv.org/abs/1802.09238> (2018).
49. Taylor, M. E. & Stone, P. Transfer Learning for Reinforcement Learning Domains : A Survey. *J. Mach. Learn. Res.* **10**, 1633–1685 (2009).
50. Pan, S. J. & Yang, Q. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering* **22**, 1345–1359 (2010).
51. Turki, T., Wei, Z. & Wang, J. T. L. Transfer Learning Approaches to Improve Drug Sensitivity Prediction in Multiple Myeloma Patients. *IEEE Access* **5**, 7381–7393 (2017).
52. Rosenbaum, L., Dörr, A., Bauer, M. R., Frankmboeckler & Zell, A. Inferring multi-target qsar models with taxonomy-based multi-task learning. *J. Cheminform.* **5**, 33 (2013).
53. Dai, W., Yang, Q., Xue, G.-R. & Yu, Y. Boosting for transfer learning. in *Proceedings of the 24th international conference on Machine learning - ICML '07* 193–200 (ACM Press, 2008). doi:10.1145/1273496.1273521
54. Hjorth Larsen, A. *et al.* The atomic simulation environment - A Python library for working with atoms. *Journal of Physics Condensed Matter* **29**, 273002 (2017).
55. Peverati, R., Zhao, Y. & Truhlar, D. G. Generalized gradient approximation that recovers the second-order density-gradient expansion with optimized across-the-board performance. *J. Phys. Chem. Lett.* **2**, 1991–1997 (2011).
56. Luo, S., Zhao, Y. & Truhlar, D. G. Validation of electronic structure methods for isomerization reactions of large organic molecules. *Phys. Chem. Chem. Phys.* **13**, 13683 (2011).
57. Sellers, B. D., James, N. C. & Gobbi, A. A Comparison of Quantum and Molecular Mechanical Methods to Estimate Strain Energy in Druglike Fragments. *J. Chem. Inf. Model.* **57**, 1265–1275 (2017).
58. Ripplinger, C., Pinski, P., Becker, U., Valeev, E. F. & Neese, F. Sparse maps - A systematic infrastructure for reduced-scaling electronic structure methods. II. Linear scaling domain based pair natural orbital coupled cluster theory. *J. Chem. Phys.* **144**, (2016).
59. Neese, F. The ORCA program system. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* (2012). doi:10.1002/wcms.81
60. Adler, T. B., Knizia, G. & Werner, H. J. A simple and efficient CCSD(T)-F12 approximation. *J. Chem. Phys.* **127**, 221106 (2007).
61. Kesharwani, M. K., Karton, A., Sylvetsky, N. & Martin, J. M. L. The S66 Non-Covalent Interactions Benchmark Reconsidered Using Explicitly Correlated Methods Near the Basis Set Limit*. *Aust. J. Chem.* 238–248 (2018). doi:10.1071/CH17588
62. Karton, A., Daon, S. & Martin, J. M. L. W4-11: A high-confidence benchmark dataset for computational thermochemistry derived from first-principles W4 data. *Chem. Phys. Lett.* **510**, 165–178 (2011).
63. Chai, J. Da & Head-Gordon, M. Systematic optimization of long-range corrected hybrid density functionals. *J. Chem. Phys.* **128**, 084106 (2008).

64. Seung, H. S., Opper, M. & Sompolinsky, H. Query by committee. in *Proceedings of the fifth annual workshop on Computational learning theory - COLT '92* 287–294 (ACM Press, 1992). doi:10.1145/130385.130417
65. Pordes, R. *et al.* The open science grid. in *Journal of Physics: Conference Series* **78**, 012057 (IOP Publishing, 2007).
66. Sfiligoi, I. *et al.* The pilot way to Grid resources using glideinWMS. in *2009 WRI World Congress on Computer Science and Information Engineering, CSIE 2009* **2**, 428–432 (IEEE, 2009).

Acknowledgements

J.S.S. thanks the University of Florida for the graduate student fellowship and the Los Alamos National Laboratory (LANL) Center for Non-linear Studies for resources and hospitality. R.Z. acknowledges support by National Science Foundation (NSF) grants 1456638 and 1338192. We gratefully acknowledge the support and hardware donation from NVIDIA Corporation and express our special gratitude to Mark Berger. The authors acknowledge support of the U.S. Department of Energy (DOE) through the LANL LDRD Program. This work was performed, in part, at the Center for Integrated Nanotechnologies, an Office of Science User Facility operated for the U.S. DOE Office of Science. We also acknowledge the LANL Institutional Computing (IC) program and ACL data team for providing computational resources. O.I. acknowledges support from DOD-ONR (N00014-16-1-2311) and Eshelman Institute for Innovation award. The authors acknowledge Extreme Science and Engineering Discovery Environment (XSEDE) award DMR110088, which is supported by National Science Foundation grant number ACI-1053575. This research in part was done using resources provided by the Open Science Grid^{65,66} which is supported by the National Science Foundation award 1148698, and the U.S. DOE Office of Science. AER thanks NSF CHE-1802831 and OI thanks NSF CHE-1802789.

Author Contribution

J.S.S. modified codes for transfer learning, assisted in data selection, analyzed results, formulated methods, and wrote the manuscript. B.T.N. developed scripts to generate CCSD(T)*/CBS data, ran all calculations, and edited the manuscript. R.Z. developed the CCSD(T)*/CBS extrapolation scheme, assisted in the analysis of results, and wrote the manuscript. N.L. helped formulate transfer learning methodology and edited the manuscript. C.D. wrote scripts for relaxed torsion scans using ANI models, assisted in benchmarks, and edited the manuscript. K.B. helped formulate methods and edited the manuscript. S.T., O.I., and A.E.R formulated methods, analyzed results, and edited the manuscript.

Competing Interests

The authors declare no competing interests.

Supplementary information

Approaching coupled cluster accuracy with a general-purpose neural network potential through transfer learning

Smith et al.

Supplementary Methods

S1.1 Detailed description of the CCSD(T)*/CBS scheme

The linear-scaling domain-localized DPLNO approximation, and its accuracy, are primarily controlled by an electron pair cutoff threshold. There are three pre-set thresholds implemented in ORCA: LoosePNO, NormalPNO, and TightPNO, which are recommended for rapid estimation, general thermochemistry and kinetics, and non-covalent interactions and molecular conformations, respectively.¹ They were benchmarked against two high-quality and distinct datasets: S66² for intermolecular interaction energies relevant to biomolecules and HF51³ for reaction energies of medium-sized organic molecules. The effect of tightening the DPLNO threshold is very significant for these datasets: TightPNO setting reduces energy RMSD from 0.36 to 0.06 kcal*mol⁻¹ for S66 and from 0.38 to 0.16 kcal*mol⁻¹ for HF51, compared to NormalPNO.¹ Therefore, in order to obtain a high-quality approximation to CCSD(T) energies TightPNO thresholds must be properly chosen.

Another key component of our efficient computational method is the complete basis set (CBS) extrapolation. We have applied the widely used two-point “EP1”⁴ extrapolation scheme originally proposed by Hobza and coworkers⁵. The total energy is computed as the sum of the MP2/CBS extrapolated energy and the difference between CCSD(T) and MP2 energies calculated with a smaller basis set. For the sake of computational efficiency we have used cc-pVTZ and cc-pVQZ basis sets for extrapolation of HF and MP2 energies using the formulas of Halkier⁶ and Helgaker⁷ ($\alpha_{34} = 5.46$, $\beta_{34} = 3.05$ ⁸). The difference between CCSD(T) and MP2 energies was estimated with the cc-pVTZ basis set as shown in Supplementary Equation 1.

$$E_{total}^{CBS} \approx E_{HF}^{CBS} + E_{MP2}^{CBS} + (E_{CCSD(T)}^{cc-pVTZ} - E_{MP2}^{cc-pVTZ}) \quad 1$$

The most computationally demanding term in the Equation 1 is $E_{CCSD(T)}^{cc-pVTZ}$, which we need to calculate with at least TightPNO level of accuracy. To reduce computational cost further, we approximated this term using an idea similar to the CBS extrapolation approach above: TightPNO and NormalPNO-CCSD(T) energies were calculated with a smaller basis set and the difference was added to the NormalPNO-CCSD(T)/cc-pVTZ energy. Summarizing, the CCSD(T)/cc-pVTZ part in Supplementary Eq. 1 was estimated with Supplementary Equation 2.

$$E_{CCSD(T)}^{cc-pVTZ} \approx E_{Normal-DPLNO-CCSD(T)}^{cc-pVTZ} + (E_{Tight-DPLNO-CCSD(T)}^{cc-pVDZ} - E_{Normal-DPLNO-CCSD(T)}^{cc-pVDZ}) \quad 2$$

Such a composite scheme for approximating CCSD(T)/CBS energies is computationally efficient, as it requires computation of Tight-DPLNO-CCSD(T) with at most a double-zeta basis set. The robustness of this suggested CCSD(T)*/CBS scheme is evaluated against two distinct datasets with high-quality CCSD(T)-F12 data available: S66 benchmark⁹ for weak intermolecular interactions and the extensive W4-11 benchmark¹⁰ for thermochemistry. CCSD(T)*/CBS performs excellent on both benchmarks (See Supplementary Table 1).

In Supplementary Table 1, the extrapolation schemes are noted with the “small” basis set, which is used to calculate $\Delta CCSD(T)$ with respect to MP2, the MP2/CBS energy obtained using extrapolation with basis set with higher cardinal number. CCSD(T)*/CBS is equal to the TightPNO-CCSD(T)*/CBS(TZ) scheme, which was applied for obtaining our reference training dataset. TightPNO-CCSD(T)* denotes that the Δ TightPNO term is the difference between TightPNO and NormalPNO energy using a basis set with lower cardinal number. The data in Supplementary Table 1 shows that CCSD(T)*/CBS provides excellent balance between cost and performance with errors no worse than the canonical CCSD(T)/CBS(aDZ) method. It also should be noted that incorporating the Δ TightPNO approach helps to achieve much lower error, compared to NormalPNO, without sacrificing much computational efficiency.

S1.2 Neural network model details

S1.2.1 Neural network architecture

The models trained in this work all utilize variable size networks for different atomic species. This is done since the molecules in the ANI-1x datasets are proportioned 48% H, 30% C, 13% N, and 9% O. Supplementary Table 2 provides details of each model architecture used in this work.

S1.2.2 Atomic environment vector parameters

The ANI model atomic environment vector (AEV) is computed using the in-house NeuroChem software suite. These AEVs are computed identically to those published in the ANI-1 work.¹¹ In this work the atomic elements C, H, N, and O are described by the AEVs (using the parameters below) yielding a total of 384 AEV elements per atom. The AEV parameters used to train each model are supplied below.

Radial Parameters:

- Radial cutoff = 5.2 Å

- $\eta^{Radial} = [16]$
- $R_s^{Radial} = [0.900000, 1.168750, 1.437500, 1.706250, 1.975000, 2.243750, 2.51250, 2.781250, 3.050000, 3.318750, 3.587500, 3.856250, 4.125000, 4.39375, 4.662500, 4.931250] \text{ \AA}$

Angular Parameters:

- Angular cutoff = 3.5 Å
- $\eta^{Angular} = [8]$
- $R_s^{Angular} = [0.900000, 1.550000, 2.200000, 2.850000] \text{ \AA}$
- $\zeta = [32.00000]$
- $\theta_s = [0.19634954, 0.58904862, 0.9817477, 1.3744468, 1.7671459, 2.1598449, 2.552544, 2.945243]$

S1.2.3 Training details

Prior to training the ANI models, a linear fitting to the energy per atomic species is performed—essentially, an empirical self-energy term for each element. This linear fitting is over the entire dataset and the fit is performed with respect to the number of each atomic element in a given molecule as the input. The ANI models are then trained to the QM calculated energy minus the linear fitted prediction. The energy obtained from this process is roughly analogous to the process of computing an atomization energy, but without any per-atom bias. The linear fitting parameters (in Hartrees) used in this work are provided below.

ANI-1x DFT Linear fitting parameters:

- H = -0.600952980000
- C = -38.08316124000
- N = -54.70775770000
- O = -75.19446356000

ANI-1x CCSD(T)*/CBS Linear fitting parameters:

- H = -0.5991501324919538
- C = -38.03750806057356
- N = -54.67448347695333
- O = -75.16043537275567

ANI-1x DFT to CCSD(T)*/CBS Δ Linear fitting parameters:

- H = -0.003172990955487249
- C = 0.04396089482092749
- N = 0.03789128635905942
- O = 0.029194038876402876

The following hyperparameters are used during training for all ANI models. These parameters have been determined through rigorous hyperparameter searches in prior ANI potential^{11,12} development.

- Mini-batch size: 2560 molecules
- Initial learning rate: 1.0E-3
- Patience for annealing learning rate: 100
- Multiplier for annealing learning rate: 0.5
- Learning rate for training termination: 1.0E-5
- ADAM¹³ stochastic optimization is used with default parameters

A single epoch of training the ANI-1x DFT models takes 12.5s while a single epoch for training the ANI-1x CCSD(T)*/CBS model takes 1.25s on a single Titan V. Approximately 1450 epochs of training are carried out on all models before convergence. This leads to a total training time of 5 hours for the DFT models, 5.5 hours for training the DFT plus transfer and Δ -learning models, and 0.5 hours for training the model which was only trained to the CCSD(T)*/CBS data.

S1.2.4 Ensemble held out test set results

Supplementary Table 3 provides the 1/8th held out test set mean absolute error (MAE) and root mean squared error (RMSE) for the DFT trained ANI-1x and CCSD(T)*/CBS trained ANI-1ccx and ANI-1ccx-R single models from the ensemble. The ANI-1ccx models were trained to a small dataset (500k) still fits to its reference as well as the ANI-1x models trained to a larger dataset (5M). However, the ANI-1ccx-R models, which were only trained to the coupled cluster data with no transfer learning, performs significantly worse on the held-out test set. From this we conclude that transfer learning is performing as designed by providing the performance of a model trained to 5M datapoints with only 500k datapoints.

S1.3 Active learning molecular torsions

Since molecular dynamics simulations and normal mode perturbation is used for sampling the ANI-1x dataset, molecular torsion barriers can be poorly described by the ANI-1x potential. In other words, ANI torsion barriers tend to have higher error than near equilibrium conformations due to the sampling methods used to generate training data. To reduce this error and improve barrier sampling, we develop an iterative and entirely ML driven active learning technique for automatically sampling molecular torsions. We begin with a dataset of SMILES [opensmiles.org] strings; in this work we start with a portion of the ChEMBL^{14–16} database containing less than 20 total atoms. We also include the SMILES strings for molecules in the Genentech torsion benchmark¹⁷ since these molecules represent a diverse set of dihedrals in the simplest chemical environment that they can be found. From the resulting set of SMILES strings we carry out active learning iterations as follows:

1. Randomly select N smiles from the database
2. Embed the N molecules in 3D space

3. Randomly select a rotatable bond, and direction of rotation
4. Conduct relaxed scan every 10 degrees (36 points) of each selected torsion using the current ANI model
5. Carry out an ensemble disagreement test (see our work on active learning for details; a selection criterion of $\hat{\rho} = 0.2 \text{ kcal}^*\text{mol}^{-1}$ is used in this work)¹² on the $36 * N$ generated molecular conformations
6. For all M conformations that fail the ensemble disagreement test, compute ANI normal modes
7. Randomly perturb the M conformations along the normal modes to a maximum distance of 0.2\AA along each mode. Generate 4 normal mode sampled (NMS) points for each M
8. Generate DFT energies and forces for all NMS points
9. Add resulting data to the training dataset and retrain ANI model
10. Go back to 1 and iterate

We complete 20 iterations of the above scheme resulting in the generation of 202k extra DFT datapoints. We subsample 19k points from this dataset generation using our CCSD(T)/CBS extrapolation scheme. Supplementary Figure 1 compares ANI-1ccx and ANI-1x with and without the active learning generated dihedral corrections.

Supplementary Tables

Supplementary Table 1: Computational efficiency for example small molecules and performance of the CCSD(T)/CBS approximation (CCSD(T)*/CBS) evaluated against CCSD(T)-F12 reference data and compared to other composite schemes.

Method	CPU-core hours ^a		MAE / RMSD, kcal $^*\text{mol}^{-1}$	
	Alanine	Aspirin	S66 ^b	W4-11 ^c
CCSD(T)/CBS(aDZ)	1.53	42.79	0.08 / 0.10	1.58 / 1.85
CCSD(T)/CBS(haTZ)	9.13	427.00	0.03 / 0.04	1.31 / 1.53
NormalPNO-CCSD(T)/CBS(aDZ)	0.78	4.63	0.31 / 0.39	2.35 / 2.59
NormalPNO-CCSD(T)/CBS(haTZ)	1.85	16.83	0.27 / 0.36	1.91 / 1.66
TightPNO-CCSD(T)/CBS(TZ)	1.56	16.70	0.16 / 0.10	1.40 / 1.50
CCSD(T)*/CBS (our reference)	1.44	7.44	0.09 / 0.10	1.46 / 1.55

^a Calculations performed on an Intel(R) Xeon(R) E5-2630 v3 @ 2.40GHz CPU

^b Reference data from ⁹

^c Reference data from ¹⁰, averaged for all subsets of W4-11.

^d Original reference data from ²

^e Revised S66 data from ¹⁸

Supplementary Table 2: Fully connected neural network architectures for each atom type.

Hydrogen Network Architecture				
	Layer1	Layer2	Layer3	Layer4
Nodes	160	128	96	1
Activation	CELU ¹⁹	CELU	CELU	Linear
Regularization	MAX NORM (1.0E-4)	MAX NORM (1.0E-5)	MAX NORM (1.0E-6)	None
Carbon Network Architecture				
	Layer1	Layer2	Layer3	Layer4
Nodes	144	112	96	1
Activation	CELU	CELU	CELU	Linear
Regularization	MAX NORM (1.0E-4)	MAX NORM (1.0E-5)	MAX NORM (1.0E-6)	None
Oxygen and Nitrogen Network Architecture				
	Layer1	Layer2	Layer3	Layer4
Nodes	128	112	96	1
Activation	CELU	CELU	CELU	Linear
Regularization	MAX NORM (1.0E-4)	MAX NORM (1.0E-5)	MAX NORM (1.0E-6)	None

Supplementary Table 3: ANI held out test set performance. In the case of ANI-1x the errors are with respect to the reference DFT, while for ANI-1ccx and ANI-1ccx-R the errors are with respect to the CCSD(T)*/CBS reference.

Model ID (from 8x ensemble)	ANI-1x test set performance		ANI-1ccx test set performance		ANI-1ccx-R test set performance	
	MAE	RMSE	MAE	RMSE	MAE	RMSE
1	1.75	2.55	1.78	2.54	2.30	3.35
2	1.78	2.60	1.82	2.80	2.26	3.30
3	1.77	2.58	1.78	2.63	2.26	3.26
4	1.73	2.53	1.77	2.55	2.24	3.23
5	1.76	2.66	1.76	2.51	2.24	3.25
6	1.73	2.54	1.75	2.50	2.26	3.26
7	1.75	2.62	1.76	2.47	2.24	3.34
8	1.76	2.61	1.79	2.59	2.21	3.20
Mean	1.75	2.59	1.78	2.57	2.25	3.27

Supplementary Table 4. Computed conformer ΔE on the GDB-10to13 benchmark (all conformers within 100 kcal*mol⁻¹ of the global minimum) for the transfer learning-based ANI-1ccx and the Δ -learning based ANI-1ccx- Δ potential. Uncertainties are the standard deviation of each ANI model's error from the ANI ensemble used for the mean prediction. Energy units are kcal*mol⁻¹.

ANI-1ccx	ANI-1ccx- Δ
----------	--------------------

MAD	1.46±0.02	1.44±0.04
RMSD	2.07±0.04	2.04±0.15

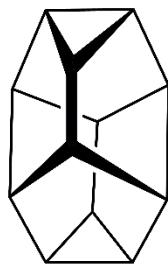
Supplementary Table 5. GDB-10to13 benchmark results comparing various ANI potentials and DFT. Errors for conformer energy differences (ΔE) and potential energies (E) for all ANI potentials. ω B97X/6-31G* errors are also provided. The blank cells are for values which cannot be compared because absolute energy differences between DFT and CCSD(T)*/CBS are arbitrary. μ and σ are the arithmetic mean and standard deviation, respectively. M and R are the MAE and RMSE, respectively. Units of energy are kcal*mol^-1.

Model	ΔE_M^μ	ΔE_M^σ	ΔE_R^μ	ΔE_R^σ	E_M^μ	E_M^σ	E_R^μ	E_R^σ
ANI-1x	3.84	0.07	5.82	0.12	--	--	--	--
ANI-1ccx	2.24	0.02	3.24	0.04	2.22	0.05	3.02	0.04
ANI-1ccx- Δ	2.21	0.07	3.18	0.10	--	--	--	--
ANI-1ccx-R	2.77	0.05	3.97	0.08	2.72	0.08	3.65	0.08
ω B97X/6-31G*	3.37	--	4.99	--	--	--	--	--

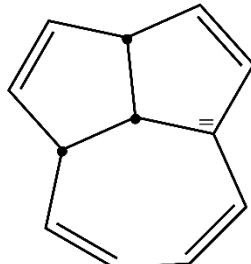
Supplementary Table 6: Performance of the two target methods and two ANI potentials used in this work. The performance comparison is on the HC7 hydrocarbon reaction energy benchmark and the ISOL6 organic molecule isomerization energy benchmark.

	<i>CCSD(T)*/CBS</i>	<i>ANI-1ccx</i>	<i>ωB97X/6-31g(d)</i>	<i>ANI-1x</i>
HC7	1.6/1.8	2.5/2.9	16.4/22.2	19.1/24.6
ISOL6	0.5/0.5	1.5/1.8	3.8/4.7	4.6/5.3

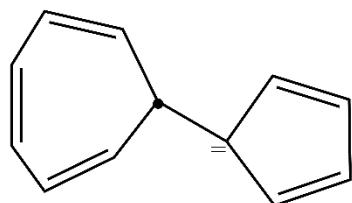
Supplementary Table 7: Reference and calculated energies (kcal^{*}mol⁻¹) for the HC7/11 benchmark. Reference calculations are obtained from Peverati, Zhao, and Truhlar^a. The rows of the table correspond to the index given in the HC7/11 portion of Figure 3 in the main article.



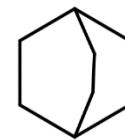
E1 (1)



E2 (22)



E3 (31)

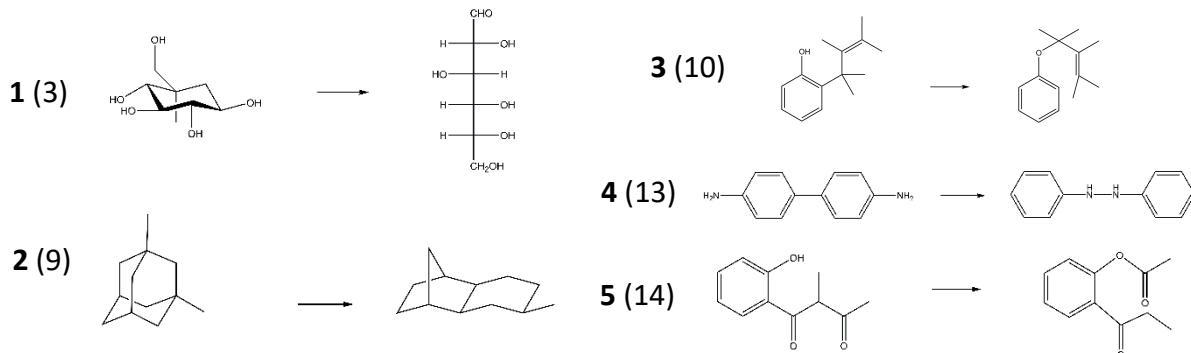


E4
(Bicyclo[2.2.2]octane)

Reaction details	Reference ^a	ANI-1ccx	CCSD(T)*/CBS	ANI-1x	ω B97X/6-31g(d)
E1 → E2	14.34	15.84	13.72	36.31	28.77
E1 → E3	25.02	30.71	23.23	54.43	41.09
Octane-a → Octane-b	1.90	0.22	-1.36	2.79	1.75
$4CH_4 + C_6H_{14} \rightarrow 5C_2H_6$	9.81	7.72	8.74	7.00	6.26
$6CH_4 + C_8H_{18} \rightarrow 7C_2H_6$	14.84	11.67	13.09	10.49	9.30
Adamantane → $3CH_4 + 2C_2H_2$	193.99	196.48	196.27	237.51	238.83
E4 → $3CH_4 + 2C_2H_2$	127.22	127.86	127.43	158.05	157.65

^a Reference data from²⁰

Supplementary Table 8: Reference and calculated energies (kcal^{*}mol⁻¹) for the ISOL6^a benchmark. Reference calculations are obtained from Peverati, Zhao, and Truhlar^b. The rows of the table correspond to the index give in the ISOL6 portion of Figure 3 in the main article. One reaction involving the atomic element fluorine (F) was left out since the ANI potential in this work is only fit to CHNO.

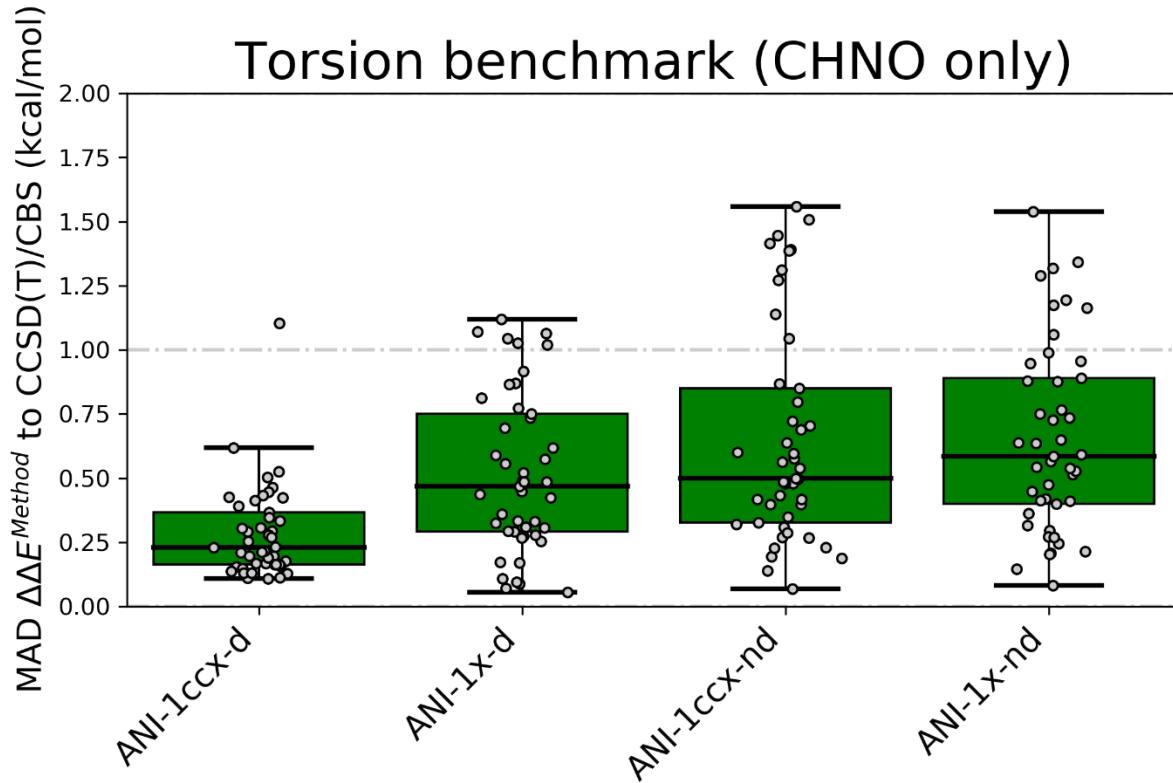


Reaction	Reference ^a	ANI-1ccx	CCSD(T) [*] /CBS	ANI-1x	ω B97X/6-31g(d)
1	9.77	9.16	10.45	6.45	9.32
2	21.76	21.48	21.07	20.85	20.80
3	6.82	3.82	6.39	-2.18	1.03
4	33.52	35.36	33.30	27.84	26.43
5	5.30	6.87	5.62	1.06	0.40

^a Reference data from²¹

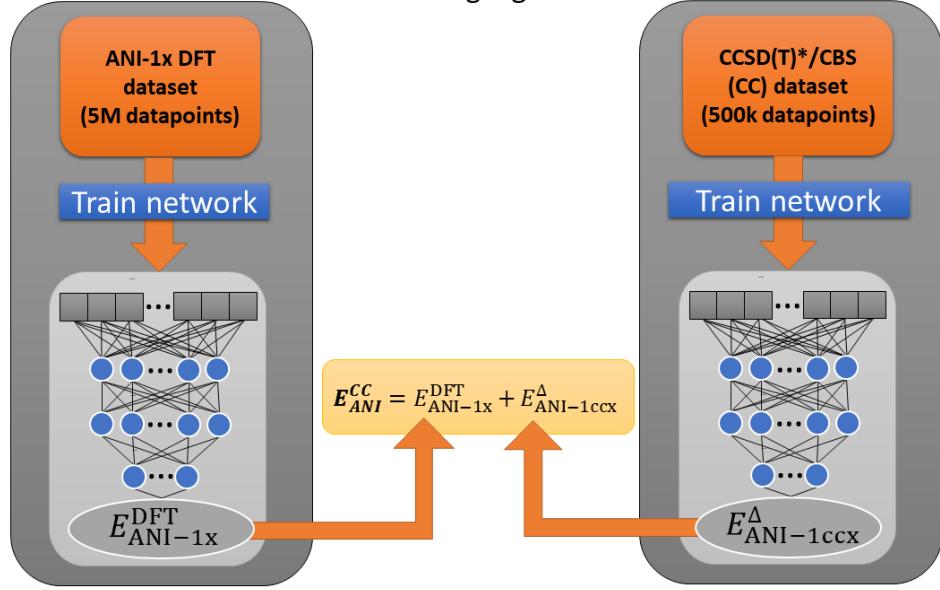
^b Reference data from²⁰

Supplementary Figures

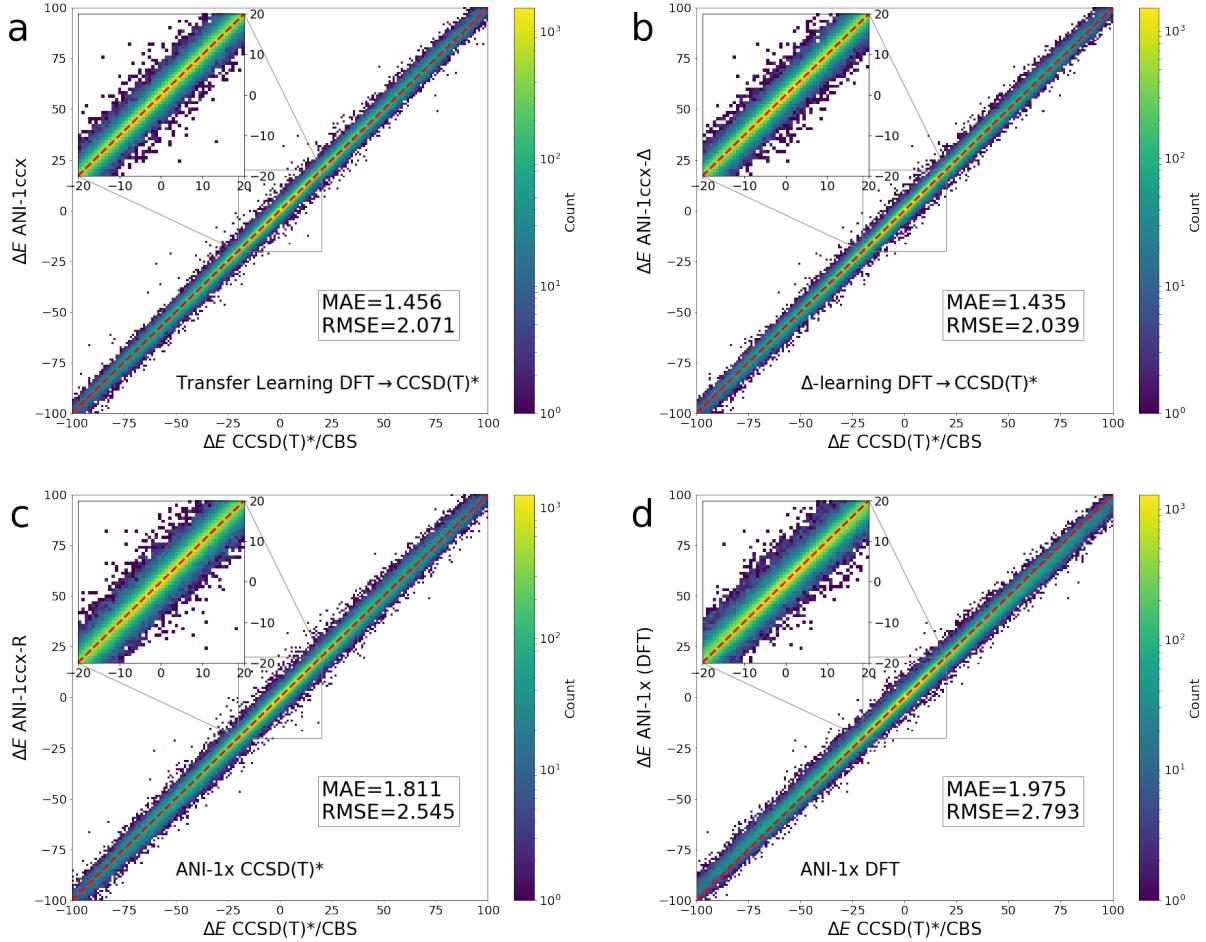


Supplementary Figure 1: Comparison of ANI model performance with and without transfer learning and dihedral active learning reparameterization on 45 torsion profiles containing the atomic elements C, H, N, and O. The ANI potentials with ‘-d’ appended are reparameterized for better torsions and with ‘-nd’ represent no reparameterization data was used during training. The grey dots represent the MAD of a given torsion scan vs. gold standard CCSD(T)/CBS. The box extends from the upper to lower quartile of the MAD distribution, the black horizontal line in the box is the median MAD. All ANI methods carried out using restrained optimization with ANI forces.

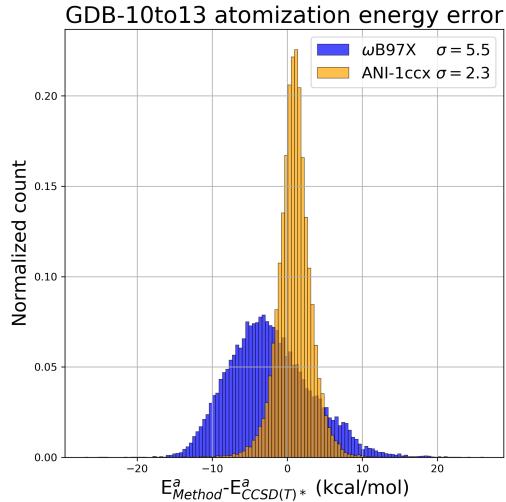
Δ -learning algorithm



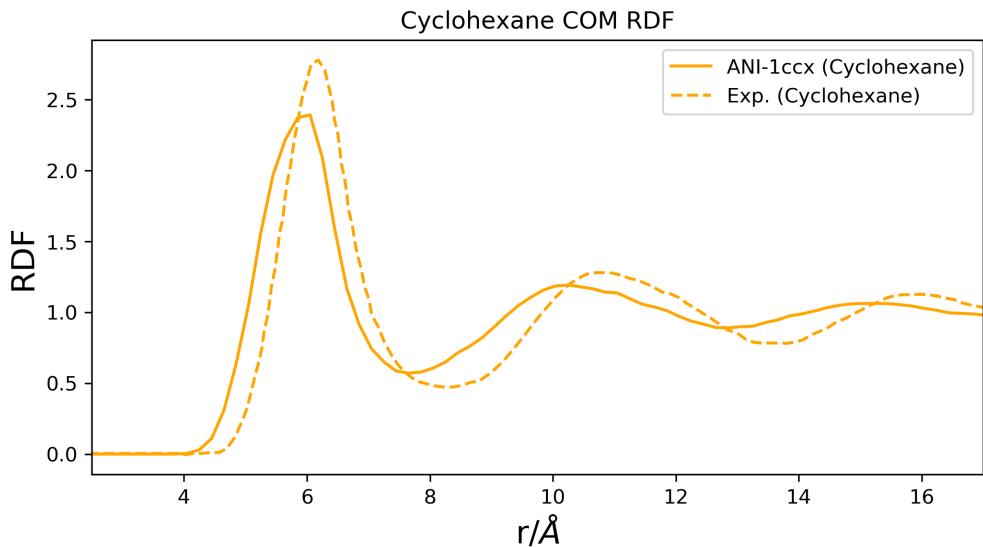
Supplementary Figure 2: Diagram of the delta learning techniques evaluated in this work. Δ -learning uses the pretrained ANI DFT model to predict an energy, then a second network is trained to correct the DFT prediction to better predict the CCSD(T)/CBS-extrapolated data. The resulting scheme requires two networks for prediction and is thus more expensive than the transfer learning.



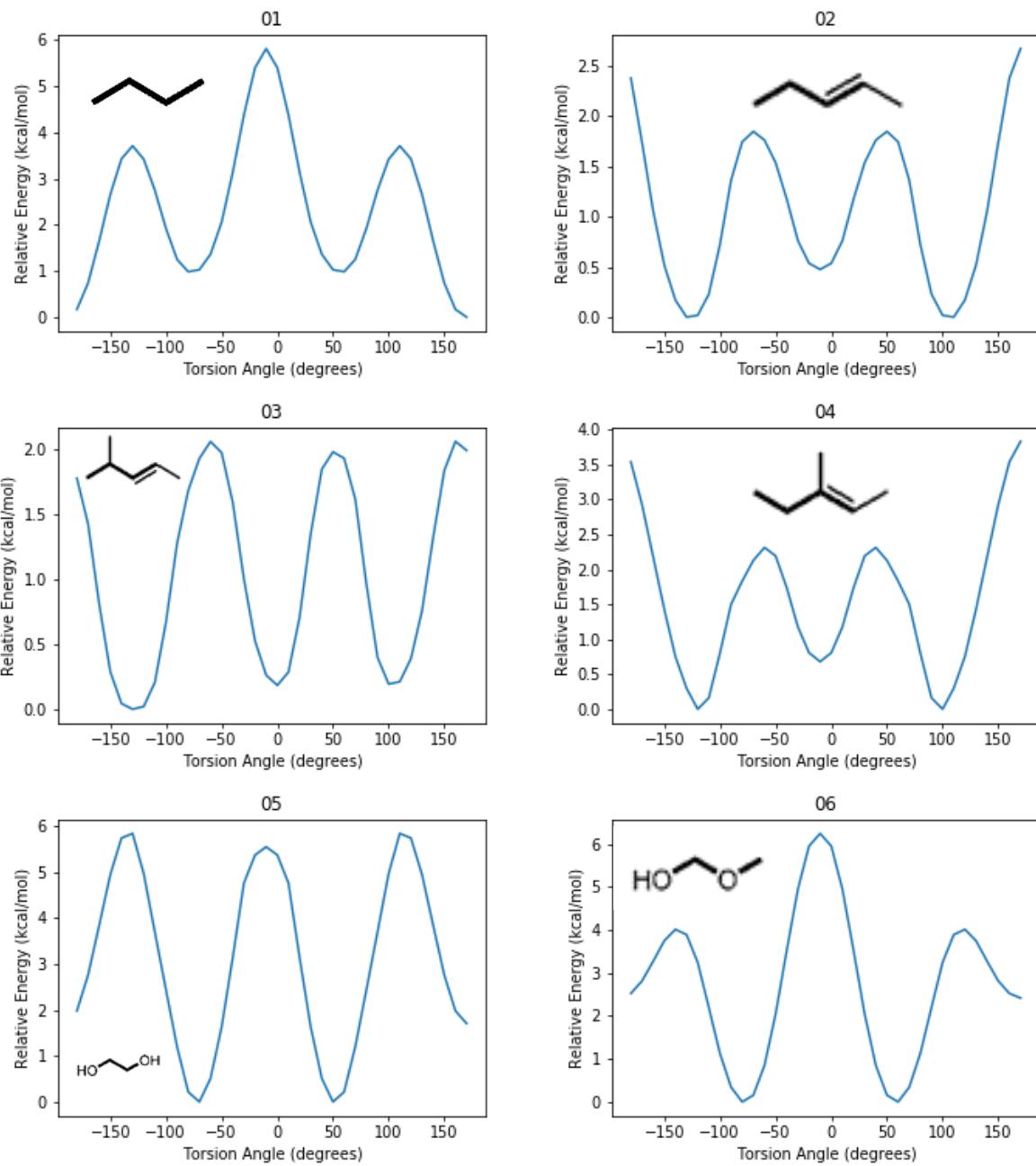
Supplementary Figure 3: Comparison of four different ANI potentials at predicting relative energies (ΔE) between all conformers for each molecule in the GDB-10to13 (CCSD(T)/CBS* computed) benchmark. These log-scale density correlation plots show a) ANI-1ccx (transfer learning-based CCSD(T)*/CBS trained model), b) ANI-1ccx- Δ (Δ -learning based CCSD(T)*/CBS trained model), c) ANI-1ccx-R (model trained only to the CCSD(T)*/CBS dataset), and d) ANI-1x (DFT trained model).

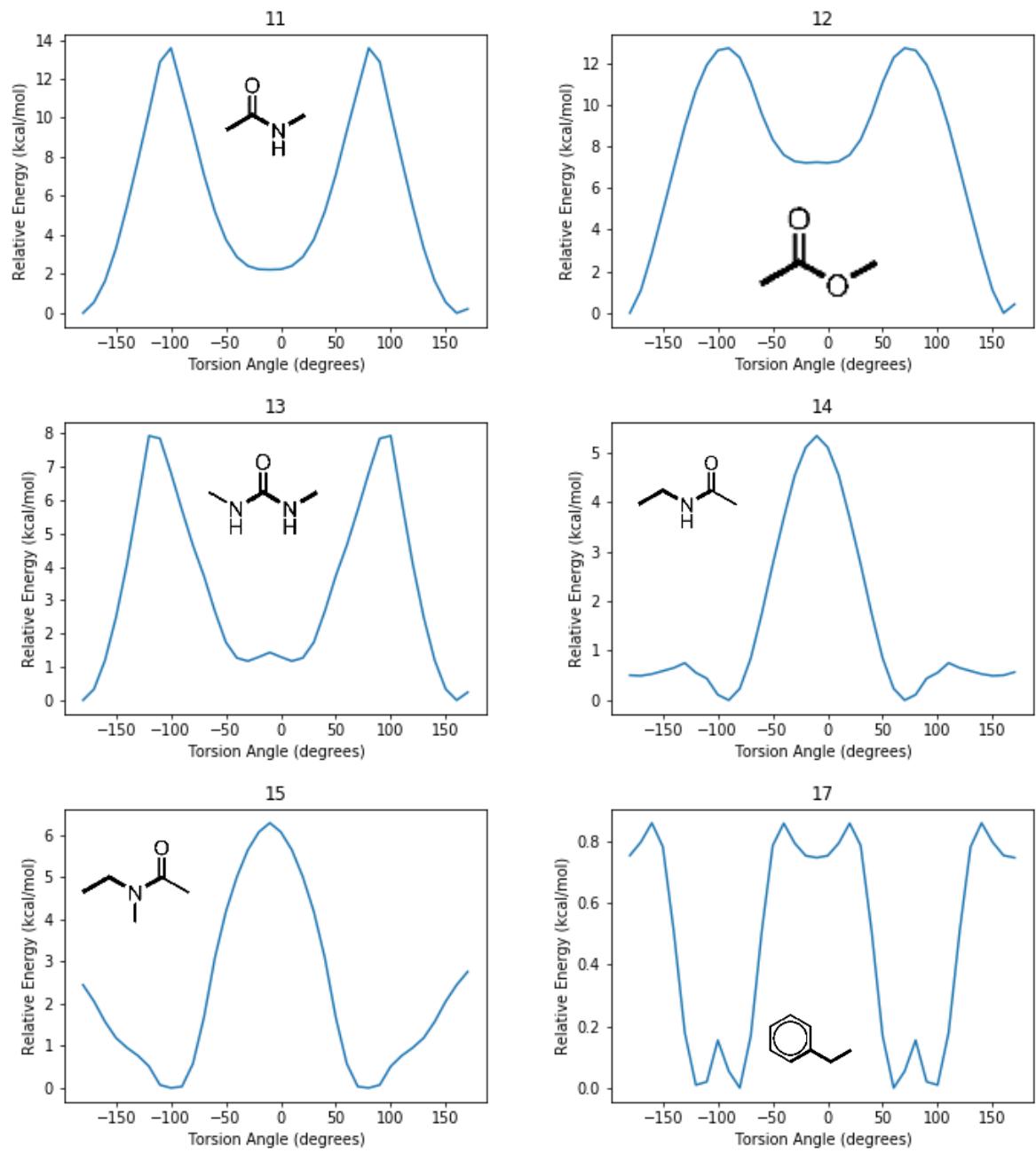


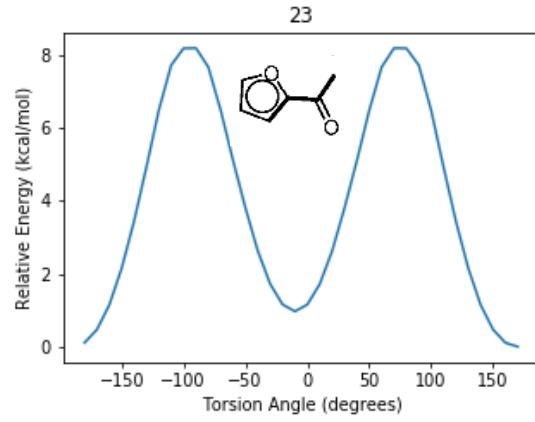
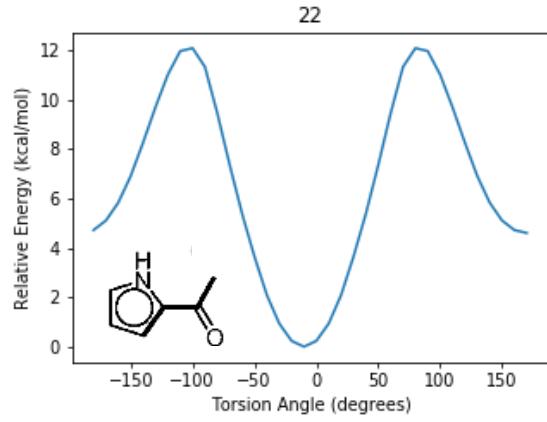
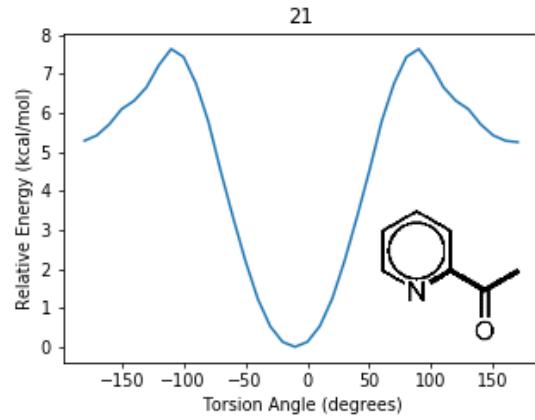
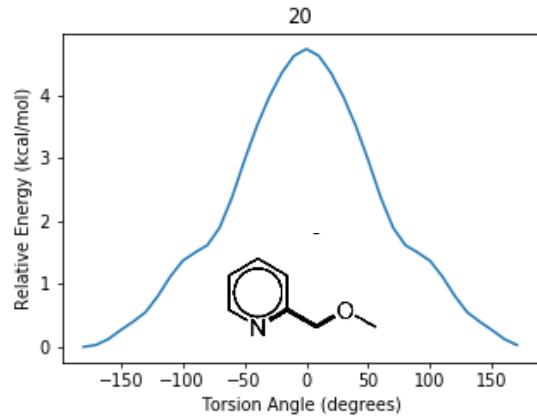
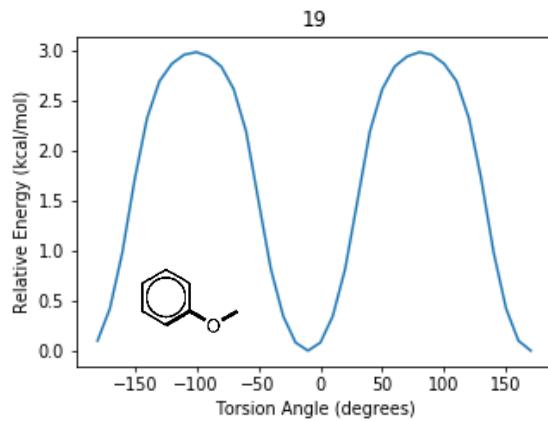
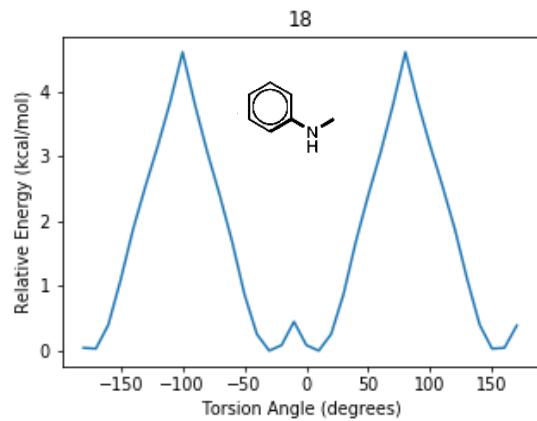
Supplementary Figure 4: Per atom corrected atomization energy difference (ΔE^a) distribution for $\omega\text{B97X}/6-31\text{G}^*$ and ANI-1ccx vs. CCSD(T)*/CBS reference data. An average correction of 18.5 kcal*mol⁻¹ was applied to the DFT data so it better fits the CCSD(T)*/CBS atomization energies. This correction came from a non-trivial linear fitting to the atomization energy difference between DFT and CCSD(T)*/CBS, based on the number of each atomic element. No such correction was applied to the ANI-1ccx data.

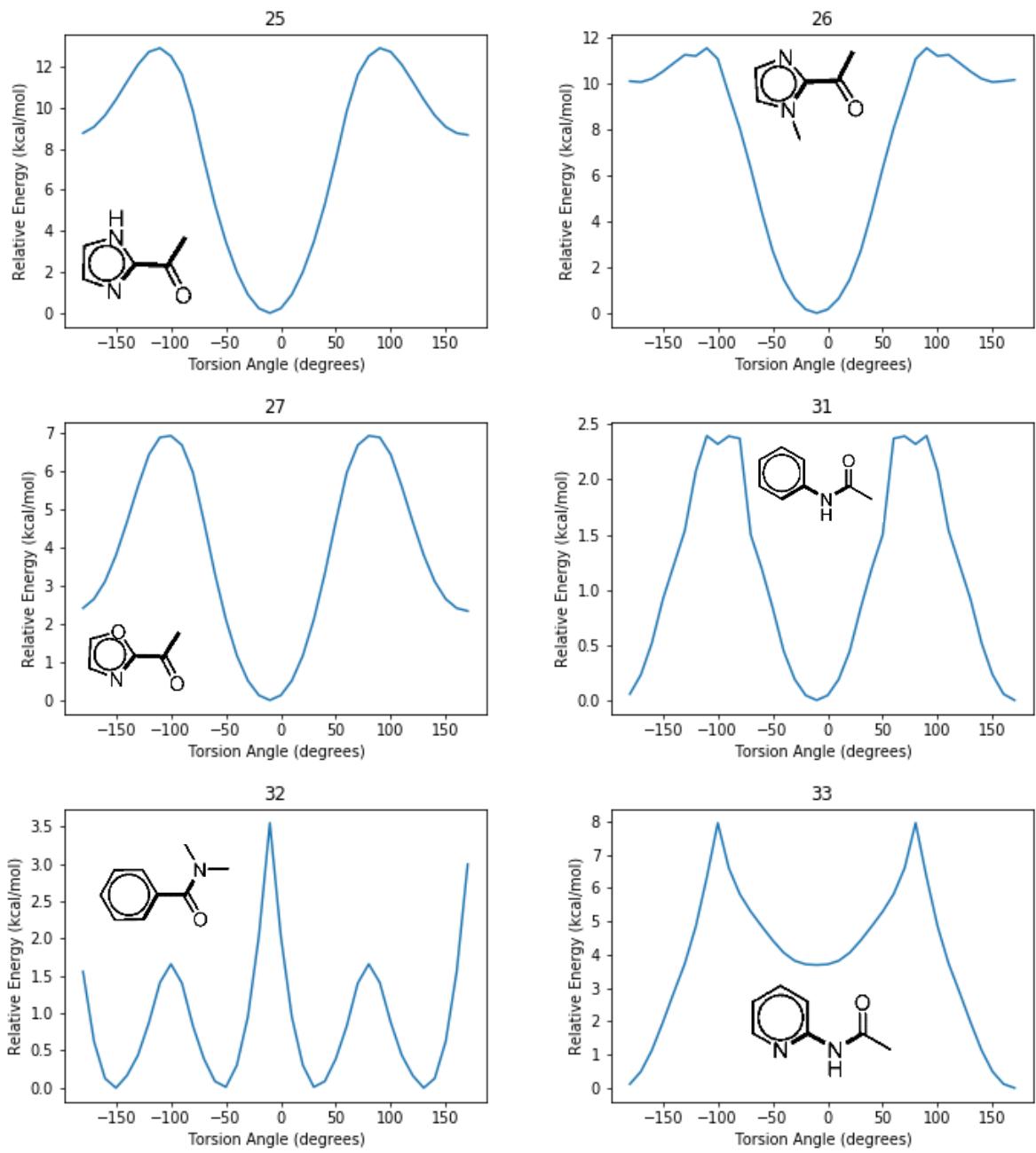


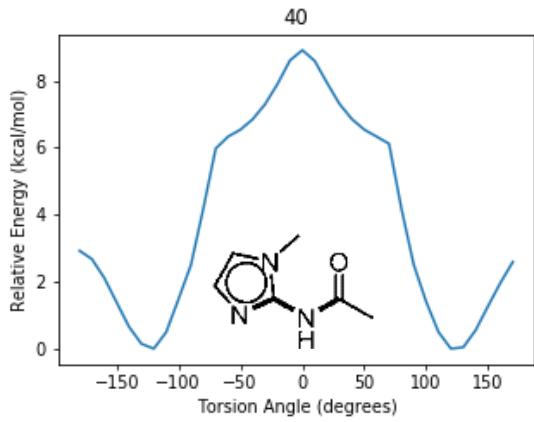
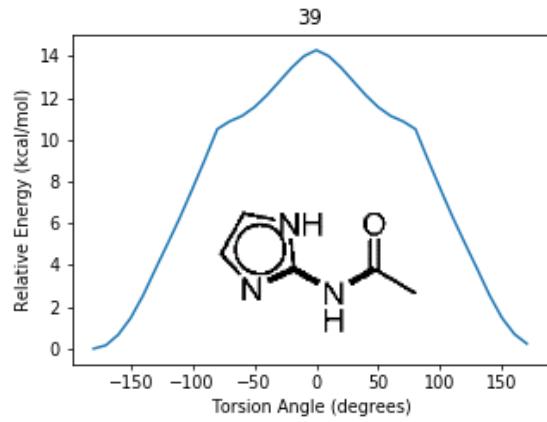
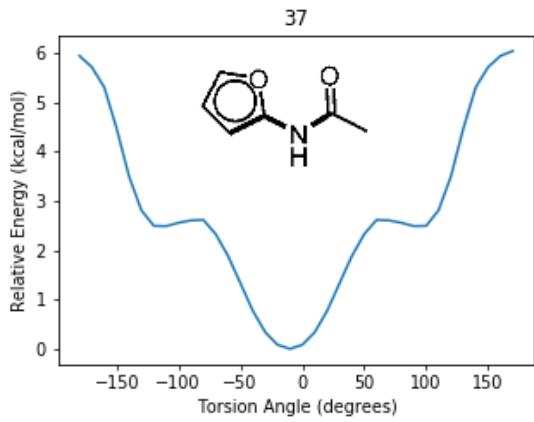
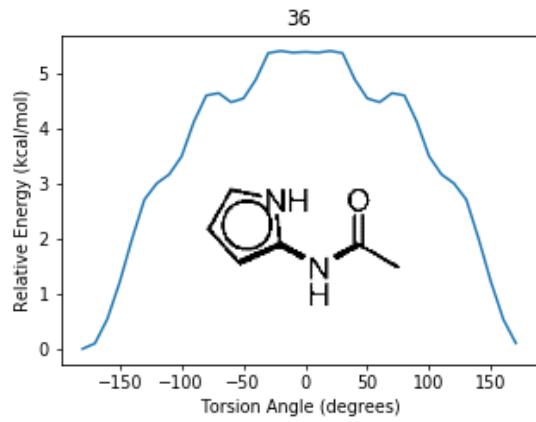
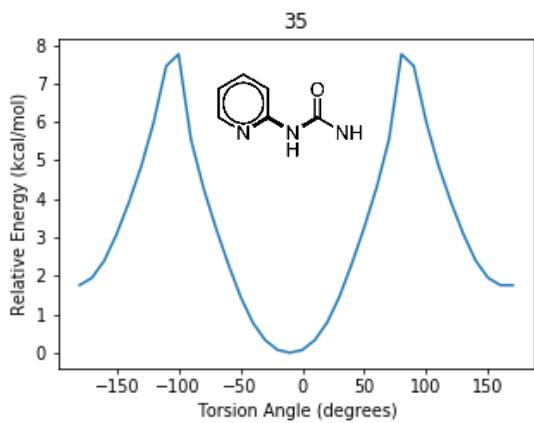
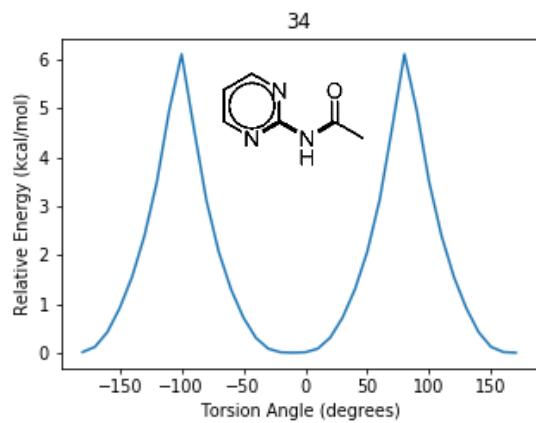
Supplementary Figure 5: Comparison of C-C center of mass radial distribution functions (RDF) from experiment²² and ANI-1ccx. Simulations was carried out with the NVT ensemble using the Langevin thermostat set to 300K.

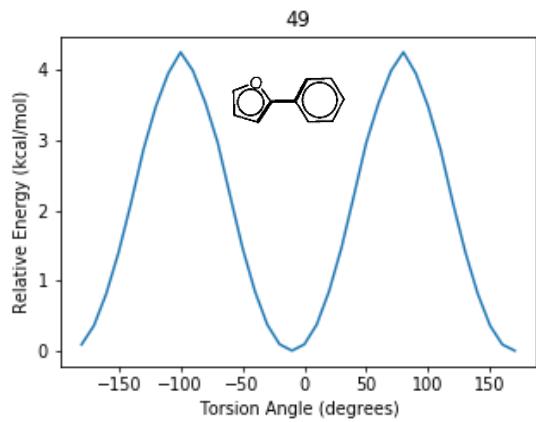
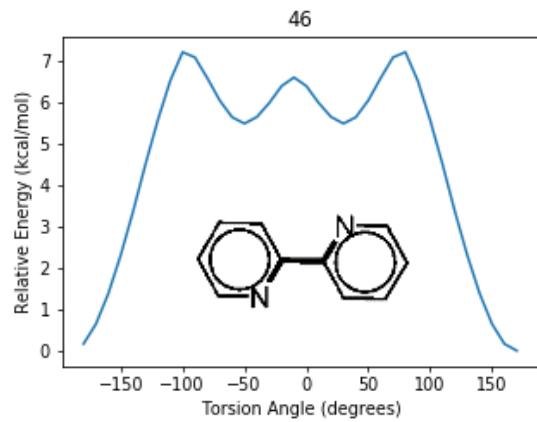
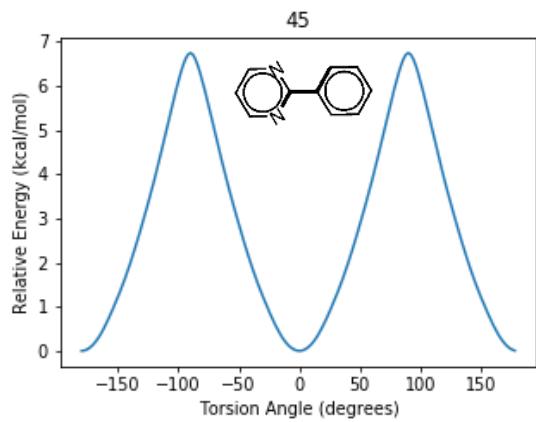
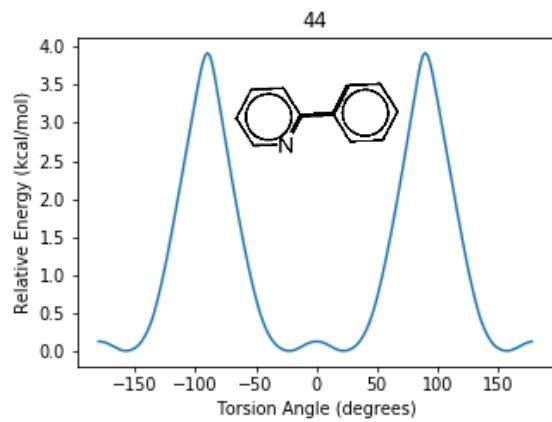
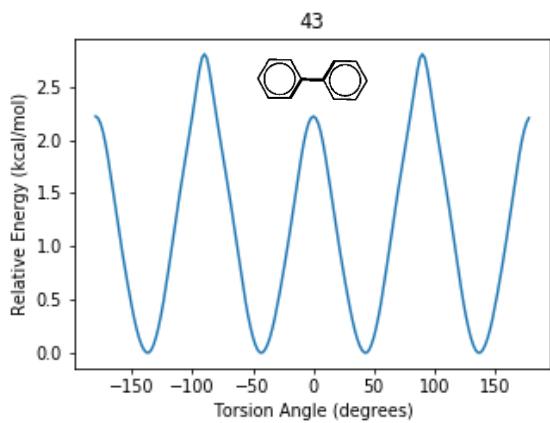
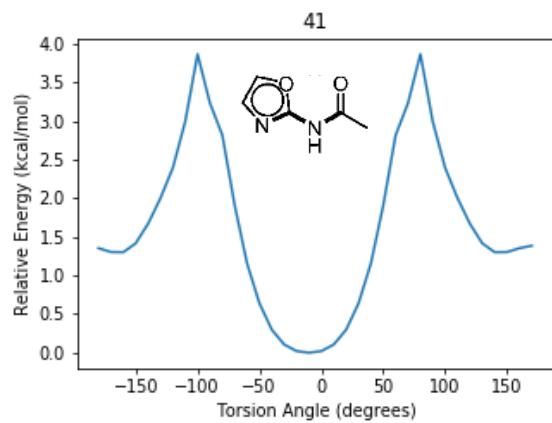


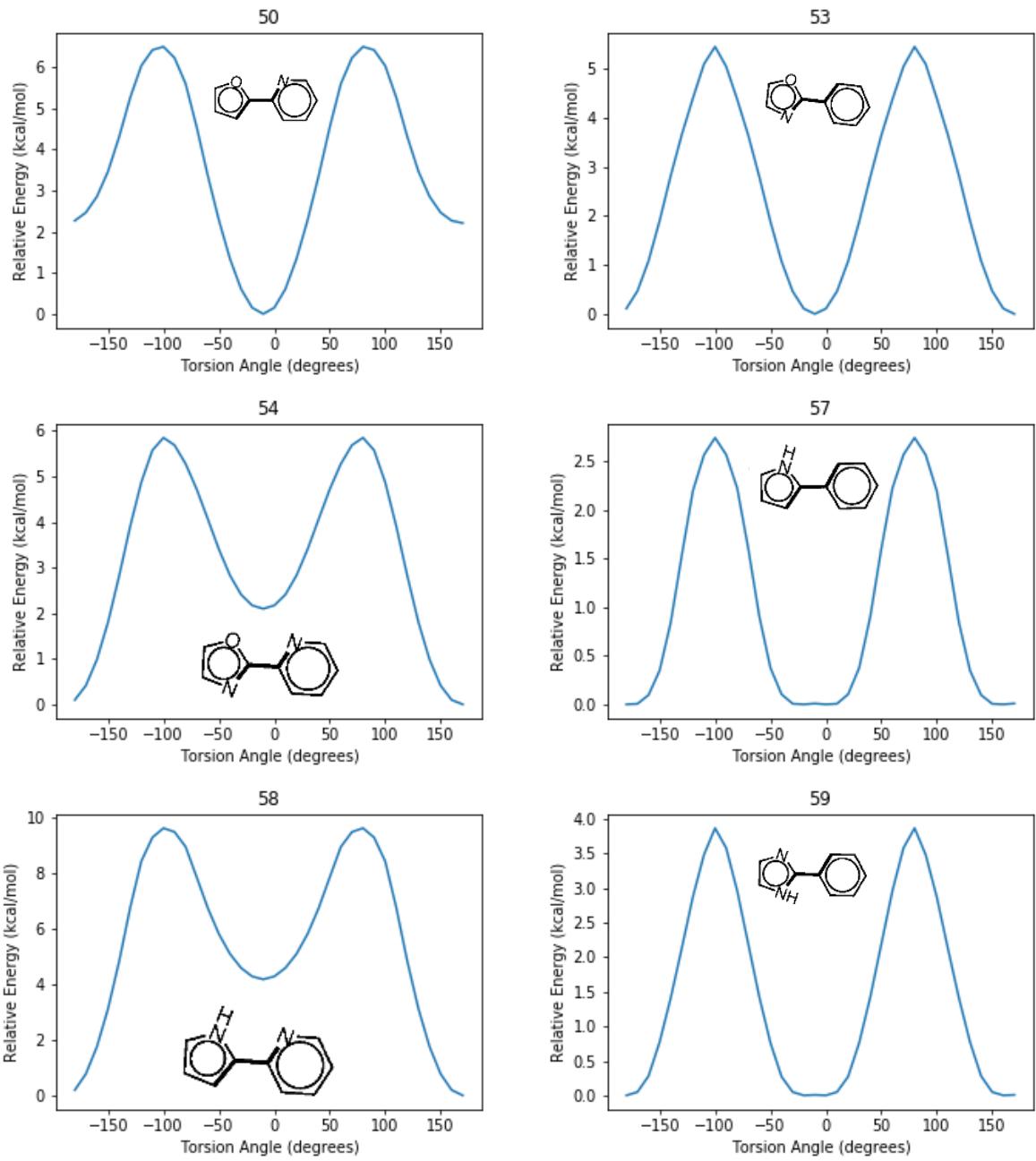


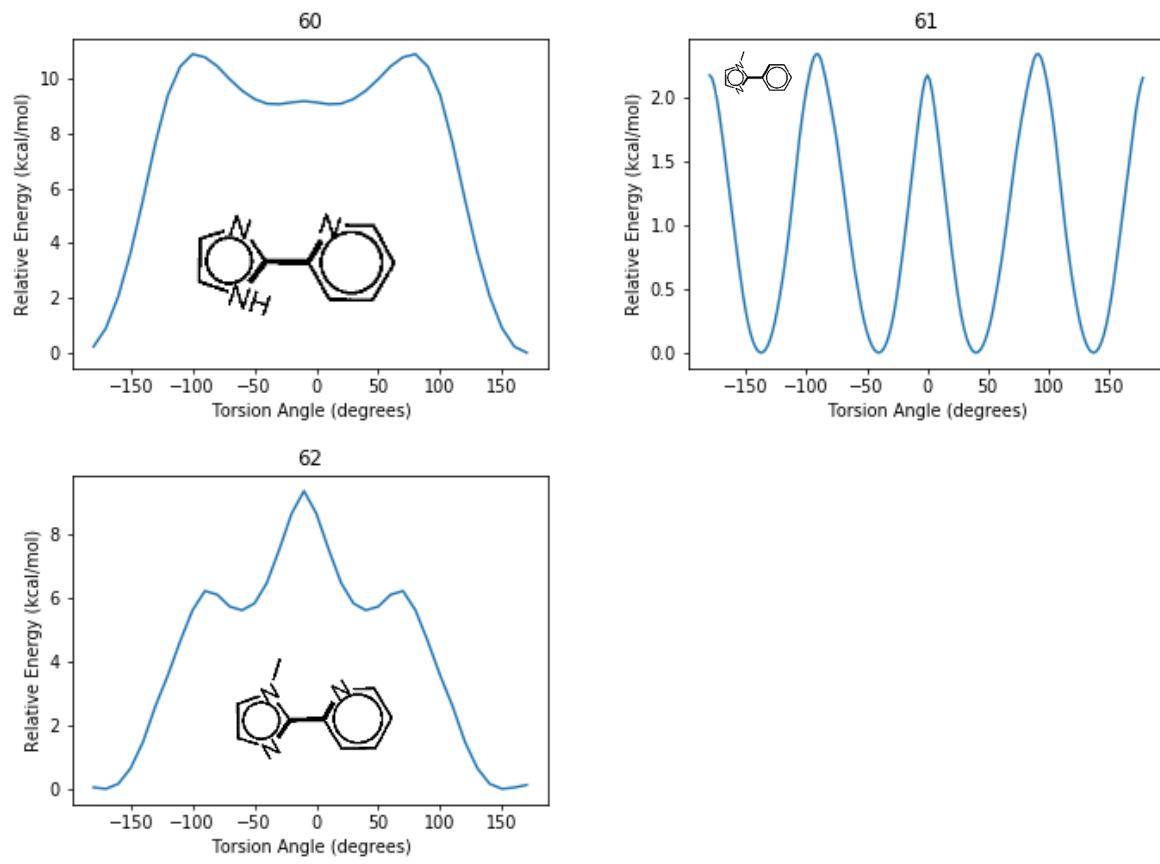












Supplementary Figure 6: Dihedrals scans with ANI-ccx used to compute the errors listed in Figure 4. The molecules are numbered according to reference 19, listing only atoms with C,H,N and O.

Supplementary References:

- (1) Liakos, D. G.; Sparta, M.; Kesharwani, M. K.; Martin, J. M. L.; Neese, F. Exploring the Accuracy Limits of Local Pair Natural Orbital Coupled-Cluster Theory. *J. Chem. Theory Comput.* **2015**, *11* (4), 1525–1539. <https://doi.org/10.1021/ct501129s>.
- (2) Řezáč, J.; Riley, K. E.; Hobza, P. S66: A Well-Balanced Database of Benchmark Interaction Energies Relevant to Biomolecular Structures. *J. Chem. Theory Comput.* **2011**, *7* (8), 2427–2438. <https://doi.org/10.1021/ct2002946>.
- (3) Friedrich, J.; Hänchen, J. Incremental CCSD(T)(F12*)|MP2: A Black Box Method to Obtain Highly Accurate Reaction Energies. *J. Chem. Theory Comput.* **2013**, *9* (12), 5381–5394. <https://doi.org/10.1021/ct4008074>.
- (4) Liakos, D. G.; Neese, F. Improved Correlation Energy Extrapolation Schemes Based on Local Pair Natural Orbital Methods. *J. Phys. Chem. A* **2012**, *116* (19), 4801–4816. <https://doi.org/10.1021/jp302096v>.
- (5) Hobza, P.; Sponer, J. Toward True DNA Base-Stacking Energies: MP2, CCSD(T), and Complete Basis Set Calculations. *J. Am. Chem. Soc.* **2002**, *124* (39), 11802–11808.
- (6) Halkier, A.; Helgaker, T.; Jørgensen, P.; Klopper, W.; Olsen, J. Basis-Set Convergence of the Energy in Molecular Hartree–Fock Calculations. *Chem. Phys. Lett.* **1999**, *302* (X), 437–446. [https://doi.org/10.1016/S0009-2614\(99\)00179-7](https://doi.org/10.1016/S0009-2614(99)00179-7).
- (7) Helgaker, T.; Klopper, W.; Koch, H.; Noga, J. Basis-Set Convergence of Correlated Calculations on Water. *J. Chem. Phys.* **1997**, *106* (23), 9639–9646. <https://doi.org/10.1063/1.473863>.
- (8) Neese, F.; Valeev, E. F. Revisiting the Atomic Natural Orbital Approach for Basis Sets: Robust Systematic Basis Sets for Explicitly Correlated and Conventional Correlated Ab Initio Methods? *J. Chem. Theory Comput.* **2011**, *7* (1), 33–43. <https://doi.org/10.1021/ct100396y>.
- (9) Kesharwani, M. K.; Karton, A.; Sylvetsky, N.; Martin, J. M. L. The S66 Non-Covalent Interactions Benchmark Reconsidered Using Explicitly Correlated Methods Near the Basis Set Limit*. *Aust. J. Chem.* **2018**, 238–248. <https://doi.org/10.1071/CH17588>.
- (10) Karton, A.; Daon, S.; Martin, J. M. L. W4-11: A High-Confidence Benchmark Dataset for Computational Thermochemistry Derived from First-Principles W4 Data. *Chem. Phys. Lett.* **2011**, *510* (4–6), 165–178. <https://doi.org/10.1016/j.cplett.2011.05.007>.
- (11) 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. <https://doi.org/10.1039/C6SC05720A>.
- (12) Smith, J. S.; Nebgen, B.; Lubbers, N.; Isayev, O.; Roitberg, A. E. Less Is More: Sampling Chemical Space with Active Learning. *J. Chem. Phys.* **2018**. <https://doi.org/10.1063/1.5023802>.
- (13) Kingma, D.; Ba, J. Adam: A Method for Stochastic Optimization. *arXiv* **2014**, Preprint at <https://arxiv.org/abs/1412.6980>.
- (14) Jupp, S.; Malone, J.; Bolleman, J.; Brandizi, M.; Davies, M.; Garcia, L.; Gaulton, A.; Gehant, S.; Laibe, C.; Redaschi, N.; et al. The EBI RDF Platform: Linked Open Data for the Life Sciences. *Bioinformatics* **2014**, *30* (9), 1338–1339. <https://doi.org/10.1093/bioinformatics/btt765>.
- (15) Bento, A. P.; Gaulton, A.; Hersey, A.; Bellis, L. J.; Chambers, J.; Davies, M.; Krüger, F. A.; Light, Y.; Mak, L.; McGlinchey, S.; et al. The ChEMBL Bioactivity Database: An Update. *Nucleic Acids Res.* **2014**, *42* (D1), D1083–D1090. <https://doi.org/10.1093/nar/gkt1031>.

- (16) Davies, M.; Nowotka, M.; Papadatos, G.; Atkinson, F.; van Westen, G.; Dedman, N.; Ochoa, R.; Overington, J. MyChEMBL: A Virtual Platform for Distributing Cheminformatics Tools and Open Data. *Challenges* **2014**, *5* (2), 334–337. <https://doi.org/10.3390/challe5020334>.
- (17) Sellers, B. D.; James, N. C.; Gobbi, A. A Comparison of Quantum and Molecular Mechanical Methods to Estimate Strain Energy in Druglike Fragments. *J. Chem. Inf. Model.* **2017**, *57* (6), 1265–1275. <https://doi.org/10.1021/acs.jcim.6b00614>.
- (18) Řezáč, J.; Riley, K. E.; Hobza, P. Extensions of the S66 Data Set: More Accurate Interaction Energies and Angular-Displaced Nonequilibrium Geometries. *J. Chem. Theory Comput.* **2011**, *7* (11), 3466–3470. <https://doi.org/10.1021/ct200523a>.
- (19) Barron, J. T. Continuously Differentiable Exponential Linear Units. *arXiv* **2017**, Preprint at <https://arxiv.org/abs/1704.07483>.
- (20) Peverati, R.; Zhao, Y.; Truhlar, D. G. Generalized Gradient Approximation That Recovers the Second-Order Density-Gradient Expansion with Optimized across-the-Board Performance. *J. Phys. Chem. Lett.* **2011**, *2* (16), 1991–1997. <https://doi.org/10.1021/jz200616w>.
- (21) Luo, S.; Zhao, Y.; Truhlar, D. G. Validation of Electronic Structure Methods for Isomerization Reactions of Large Organic Molecules. *Phys. Chem. Chem. Phys.* **2011**, *13* (30), 13683. <https://doi.org/10.1039/c1cp20834a>.
- (22) Falkowska, M.; Bowron, D. T.; Manyar, H. G.; Hardacre, C.; Youngs, T. G. A. Neutron Scattering of Aromatic and Aliphatic Liquids. *ChemPhysChem* **2016**, *17* (13), 2043–2055. <https://doi.org/10.1002/cphc.201600149>.