Title: Outsmarting quantum chemistry through transfer learning

Authors:

Justin S. Smith[1,2,3,†], Benjamin T. Nebgen[2,4,†], Roman Zubatyuk[2,5,†], Nicholas Lubbers[2,3], Christian Devereux[1], Kipton Barros[2], Sergei Tretiak[2,4,], Olexandr Isayev[6,*], Adrian E. Roitberg[1,*]*

Affiliations:

[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: Computer simulations are foundational to theoretical chemistry. Quantum-mechanical (QM) methods provide the highest accuracy for simulating molecules but have difficulty scaling to large systems. Empirical interatomic potentials (classical force fields) are scalable, but lack transferability to new systems and are hard to systematically improve. Automated, data-driven machine learning is close to achieving the best of both approaches. Here we use transfer learning to retrain a general purpose neural network potential, ANI-1x, on a dataset of gold standard QM calculations (CCSD(T)/CBS level) that is relatively small but designed to optimally span chemical space. The resulting potential, *ANI-1ccx*, approaches CCSD(T)/CBS accuracy on benchmarks for reaction thermochemistry, isomerization, and drug-like molecular torsions. ANI-1ccx is broadly

applicable to materials science, biology and chemistry, and billions of times faster than the parent CCSD(T)/CBS calculations.

One Sentence Summary: Approaching coupled cluster accuracy with a general-purpose neural network potential through transfer learning.

KEYWORDS. Transfer learning, active learning, machine learning, molecular potentials, force field, neural network

Main Text:

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, empirical potentials (*e.g.* classical force fields) are commonly employed to enable large scale dynamical simulation such as protein folding[14], ligand-protein docking[15] or the dynamics of dislocations in materials[16]. These models are often fragile; an empirical model fit to one system may not accurately model other systems[17]. An outstanding challenge is to simultaneously capture a great diversity of chemical processes with a single empirical model.

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[31,32], atomic partial charges[32–35], molecular dipoles[26,36,37], materials discovery[38–40], and protein-ligand complex scoring[41]. 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 empirical 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 chemistry[42]. 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 molecules[43]. This vast amount of data would be impractical to generate at a level of theory more accurate than DFT[44]. 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 atoms[45]. 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 molecule[48]. 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 molecular conformations[45]), and then we retrain to a much smaller dataset (about 500k select conformations from ANI-1x) at the CCSD(T)/CBS level of accuracy. The resulting general-purpose potential, ANI-1ccx, 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 Environment[54] package (ASE; https://wiki.fysik.dtu.dk/ase/).

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*[55] which is implemented in the ORCA software package[56]. 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 1: Computational cost of CCSD(T)/CBS and CCSD(T)*/CBS methods, and accuracy compared to reference calculations at the CCSD(T)-F12 level of theory[57]. S66 and W4-11 are standard benchmarks for interaction and atomization energies of small molecules[58,59]. See Table S1 for a more detailed comparison. All calculations are performed on an Intel Xeon E5-2630 v3 @ 2.40GHz CPU.

	CPU-core hours		**Mean absolute deviation from CCSD(T)-F12 (kcal/mol)**	
	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**

showing our CCSD(T)*/CBS approximation provides accurate energies in a computationally efficient way. 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 dataset[45] is used to train an initial DFT (to the ωB97x/6-31G* model chemistry[60]) 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.[42] 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.[45] 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-

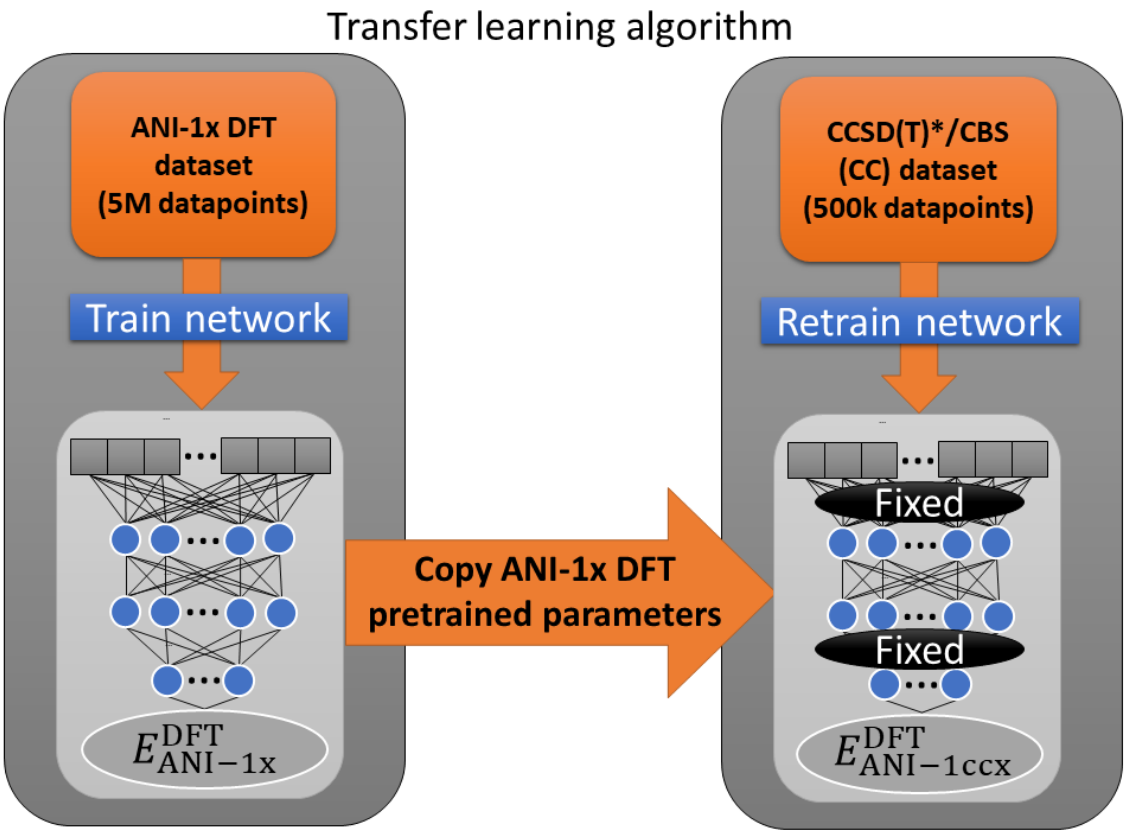


Figure 1: Diagram of the transfer learning technique evaluated in this work. Transfer learning starts from a pretrained ANI-1x DFT model, then retrains to higher accuracy CCSD(T)*/CBS data with some parameters fixed during training.

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 committee[61]). 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 learning[45] 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 1) used to create ANI-1ccx. First, an ANI model is trained to the DFT dataset with the new active learning torsion data added, yielding a potential equivalent to the ANI-1x potential[45]. We then retrain the ANI-1x potential to the CCSD(T)*/CBS data with a portion of the optimizable network parameters held constant. 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. An alternative to transfer learning is Δ-learning[46]. 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 Figure S2 and Table S3.

Results

We compare the errors of ANI-1ccx (trained with transfer learning), ANI-1x (trained on DFT data only), and direct DFT calculations (ωB97X/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. 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 benchmark[45] 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. HC7/11[62] is a benchmark designed to gauge the accuracy of hydrocarbon reaction and isomerization energies. The ISOL6

Table 2: Accuracy in predicting conformer energy differences on the GDB-10to13 benchmark relative to CCSD(T)*/CBS. The mean absolute deviations (MAD) and root mean squared deviations (RMSD) are given in kcal/mol. Methods compared are the ANI-1ccx transfer learning potential, the ANI-1ccx-R trained only on coupled cluster data, the ANI-1x trained only on DFT data, and the DFT reference (ωB97X-D3).

	ANI-1ccx	ANI-1ccx-R	ANI-1x	ωB97X-D3
MAD	1.46	1.81	1.98	1.42
RMSD	2.07	2.55	2.80	2.05

benchmark[63] (a subset of the ISOL24/11 benchmark) measures isomerization energies for organic molecules. Finally, we test on the Genentech torsion benchmark[64], which contains 62 diverse organic molecule torsion profiles (45 containing only CHNO).

Table 2 provides mean absolute deviations (MAD) and root mean squared deviations (RMSD) for the ANI potentials and ωB97X-D3/6-31g*, on the GDB-10to13 benchmark from COMP6[45] benchmark suite. Reference values are recomputed at the CCSD(T)*/CBS level of theory. Table 2 only considers conformations within 100kcal/mol 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.[45] 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 (with D3 dispersion corrections applied) model has a 36% increase in RMSD over ANI-1ccx. ANI-1ccx performs as well as the original reference with D3 dispersion corrections added (ωb97X-D3/6-31G*) in the 100kcal/mol 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%.[65] Table S4 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 vs.

5.0 kcal/mol for DFT, which means the ANI-1ccx model generalizes better to high energy conformations than ωb97X-D3/6-31G*. Figure S3 shows correlation plots for the ANI models vs. CCSD(T)*/CBS.

Figure 2 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/mol 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/mol, while the DFT distribution is much wider, with a standard deviation of 7.5 kcal/mol. The MAD/RMSD for DFT vs. reference CCSD(T)*/CBS is 6.0/7.5 kcal/mol, while for ANI-1ccx it is 1.9/2.5 kcal/mol. Figure S4 shows an attempt to correct 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 5.7 kcal/mol with MAD/RMSD of 4.9/6.0 kcal/mol.

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

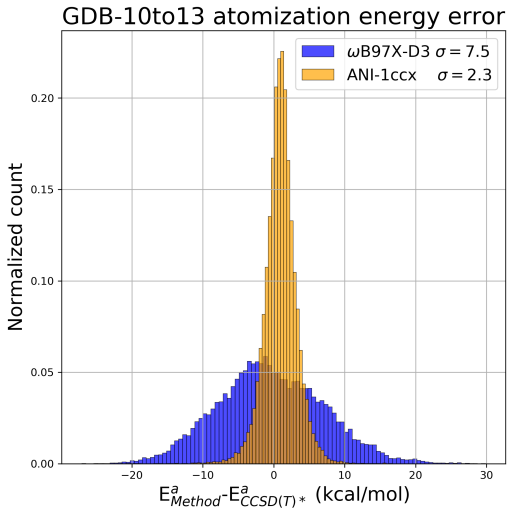


Figure 2: Accuracy in predicting atomization energies E^a on the GDB-10to13 benchmark relative to CCSD(T)*/CBS.

Table 3: Accuracy in predicting atomic forces on the GDB-10to13 benchmark relative to high-quality MP2/TZ reference calculations. Shown are the MAD/RMSD (in kcal/mol/Å) for individual force components. MP2/DZ uses a smaller basis set and is computationally more efficient than MP2/TZ.

	ANI-1ccx	**ANI-1x**	ωB97X-D3	**MP2/DZ**
MP2/TZ	3.4/5.3	4.7/7.1	3.7/5.9	4.6/5.9

to extreme computational expense with existing packages. However, MP2/cc-pVTZ (dubbed here as MP2/TZ) provides a high-quality alternative. Table 3 compares MP2/TZ force calculations on the GDB-10to13 benchmark to MP2/cc-pVDZ (MP2/DZ), ωb97x-D3/6-31G* (DFT/DZ), 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.

The HC7/11 and ISOL6 benchmarks address the calculation of reaction and isomerization energies and are depicted in Figure 3. For each reaction, reference energies and calculated energies are provided in Tables S7 and S8. Figure 3 shows the differences between the computed and the reference energies, for the reaction and isomerization energies individually for ωB97X-D3/6-31g*, 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. For each system in the benchmark, ANI-1ccx and ANI-1x (integrated with the Atomic Simulation Environment[54] package) was employed to optimize the reactants and products in each case. The LBFGS optimization algorithm was used for all geometry optimizations. Performing geometry optimizations with ANI-1ccx in this benchmark shows the applicability of ANI-1ccx forces in a realistic computational chemistry project setting. ωB97X-D3/6-31g* was also used to optimize the structures for comparison with the benchmarks. Our CCSD(T)*/CBS methods used the structures provided by the benchmarks. For the HC7/11 benchmark, the medium-sized basis DFT reference ωB97X-D3/6-31g* 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 16.0/21.7, 17.8/23.1, 1.9/2.0 and 1.5/1.8 kcal/mol, respectively.

Figure 3b displays a similar comparison for the five medium sized organic C, H, N, O containing molecules of the ISOL6[63] isomerization energy benchmark. A similar trend is seen here as with the HC7/11 benchmark, where ANI-1x deviations tend to correlate with the large deviations of its reference DFT. The transfer learning-based model follows the accuracy of the CCSD(T)*/CBS reference well. For the ISOL6 reactions shown in Figure 3b, overall MAD/RMSD for DFT, ANI-1x, ANI-1ccx, and CCSD(T)*/CBS are 4.0/4.7, 5.1/5.6, 1.5/1.6 and 1.0/1.1 kcal/mol, respectively.

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.[64] This benchmark provides a measure of accuracy for a model at reproducing potential energy profiles from a diverse set of molecular torsions. Figure 4 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.[64] 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. The red QM methods (first three from the left) are computationally intense QM methods which used MP2 restrained optimized structures, while the green QM methods were all optimized using their own forces. The ANI and MM models also carried out restrained optimizations using their own forces. The ANI-1x potential, trained to

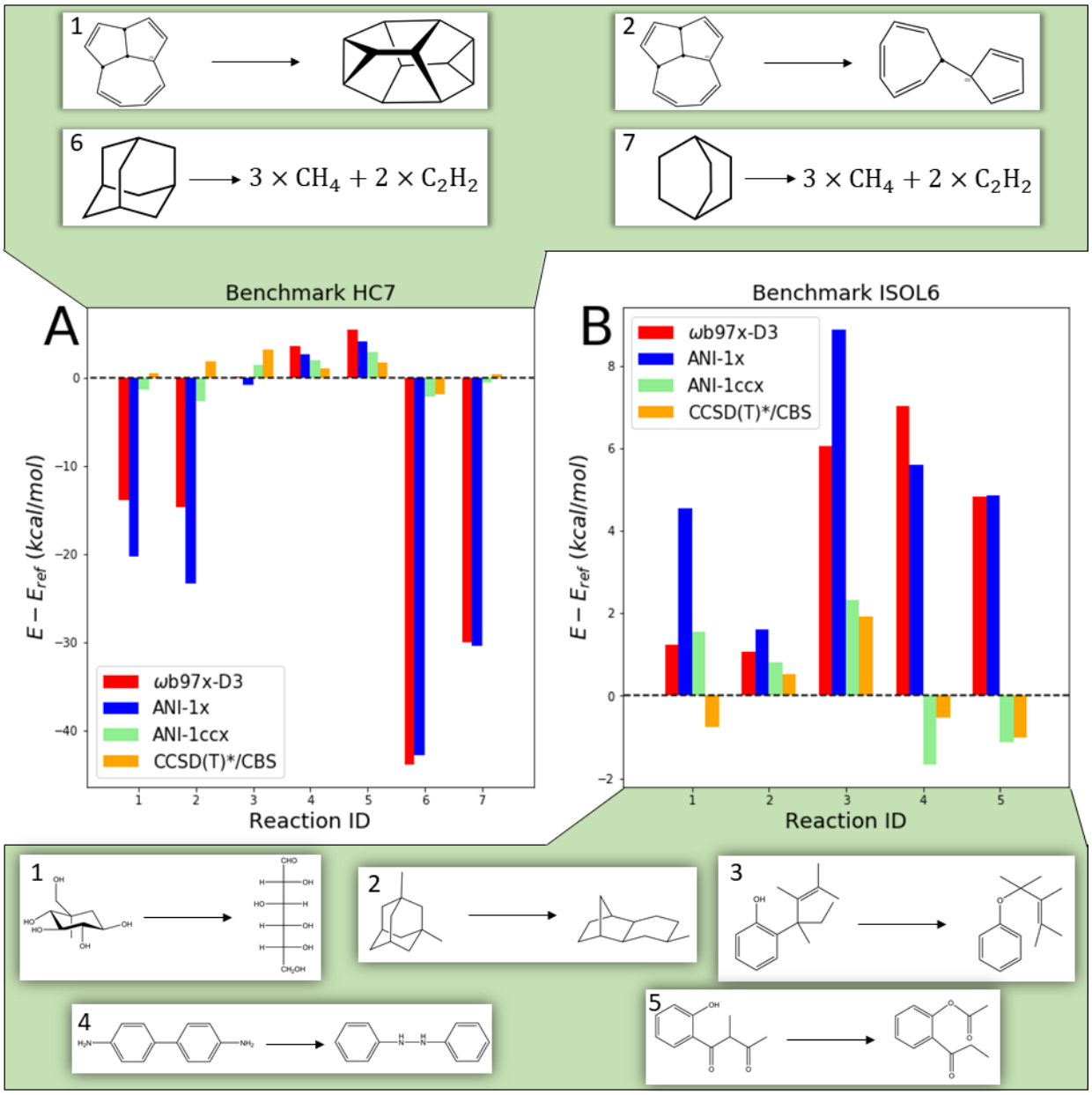


Figure 3: Accuracy in predicting reaction and isomerization energy differences on the **(a)** HC7/11 and **(b)** ISOL6 benchmarks, relative to CCSD(T)/CBS. Methods compared are the ANI-1ccx transfer learning potential, ANI-1x trained only on DFT data, the DFT reference (ωB97X-D3), and our coupled cluster 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.

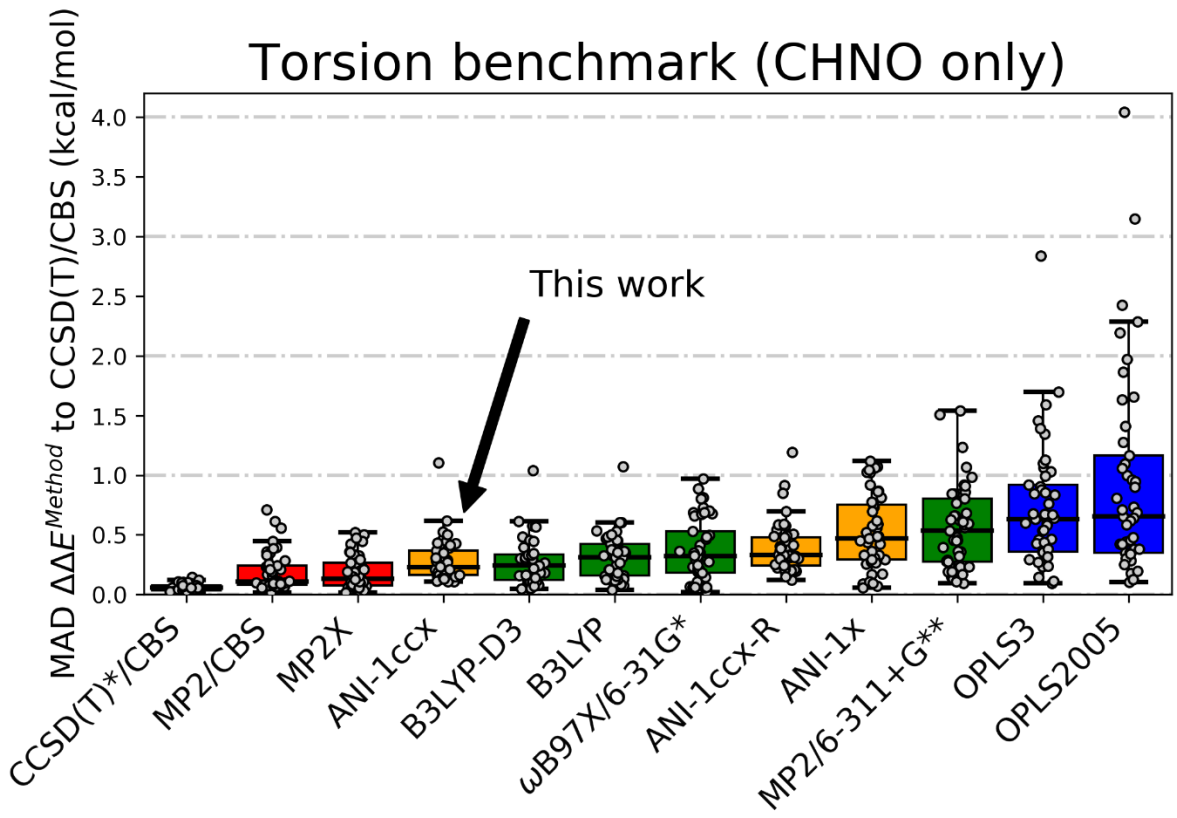


Figure 4: Accuracy in predicting torsional energies relevant to drug discovery. Methods compared are QM (red and green), molecular mechanics (blue) and ANI (orange) performance on 45 torsion profiles containing C, H, N and O atomic elements. The grey dots represent the MAD of a given torsion scan vs. gold standard CCSD(T)/CBS. The box extends from the upper to lower quartile, the black horizontal line in the box is the median, and the "whiskers" guide the eye to identify outliers as in Sellers et al[64].

the ANI-1x DFT dataset plus active learning-based dihedral corrections, obtains a median MAD of 0.47 kcal/mol 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, 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-D3/6-31g*) 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.* 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.[64] 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 and a small prefactor.

Conclusions and outlook

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. We conclude that subtle but important physics captured by gold-standard QM is modeled by ANI-1ccx. 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.

References and Notes:

(1) Roberts, R. M. Serendipity: Accidental Discoveries in Science. *Serendipity Accid. Discov. Sci. by Royst. M. Roberts* **1989**, *67* (12), 288.

(2) Berson, J. A. Discoveries Missed, Discoveries Made: Creativity, Influence, and Fame in Chemistry. *Tetrahedron* **1992**, *48* (1), 3–17.

(3) Pople, J. A. Quantum Chemical Models (Nobel Lecture). *Angew. Chemie Int. Ed.* **1999**, *38* (13–14), 1894–1902.

(4) Kohn, W. Nobel Lecture: Electronic Structure of Matter—wave Functions and Density Functionals. *Rev. Mod. Phys.* **1999**, *71* (5), 1253–1266.

(5) Purvis, G. D.; Bartlett, R. J. A Full Coupled-Cluster Singles and Doubles Model: The Inclusion of Disconnected Triples. *J. Chem. Phys.* **1982**, *76* (4), 1910–1918.

(6) Bartlett, R. J.; Musiał, M. Coupled-Cluster Theory in Quantum Chemistry. *Rev. Mod. Phys.* **2007**, *79* (1), 291–352.

(7) Crawford, T. D.; Schaefer, H. F. An Introduction to Coupled Cluster Theory for Computational Chemists; 2007; pp 33–136.

(8) Hobza, P.; Šponer, J. Toward True DNA Base-Stacking Energies: MP2, CCSD(T), and Complete Basis Set Calculations. *J. Am. Chem. Soc.* **2002**, *124* (39), 11802–11808.

(9) Feller, D.; Peterson, K. A.; Crawford, T. D. Sources of Error in Electronic Structure Calculations on Small Chemical Systems. *J. Chem. Phys.* **2006**, *124* (5), 054107.

(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.* **2011**, *7* (11), 3466–3470.

(11) Grimme, S. Density Functional Theory with London Dispersion Corrections. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **2011**, *1* (2), 211–228.

(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.* **2011**, *7* (1), 88–96.

(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.* **2017**, *115* (19), 2315–2372.

(14) Dill, K. A.; MacCallum, J. L. The Protein-Folding Problem, 50 Years On. *Science*. American Association for the Advancement of Science November 23, 2012, pp 1042–1046.

(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.* **2012**, *41* (1), 429–452.

(16) Meyers, M. A.; Mishra, A.; Benson, D. J. Mechanical Properties of Nanocrystalline Materials. *Progress in Materials Science*. Pergamon May 1, 2006, pp 427–556.

(17) Rauscher, S.; Gapsys, V.; Gajda, M. J.; Zweckstetter, M.; De Groot, B. L.; Grubmüller, H. Structural Ensembles of Intrinsically Disordered Proteins Depend Strongly on Force Field: A Comparison to Experiment. *J. Chem. Theory Comput.* **2015**, *11* (11), 5513–5524.

(18) Jordan, M. I.; Mitchell, T. M. Machine Learning: Trends, Perspectives, and Prospects. *Science*

(80-.). **2015**, *349* (6245), 255–260.

(19) Gil, Y.; Greaves, M.; Hendler, J.; Hirsh, H. Amplify Scientific Discovery with Artificial Intelligence. *Science (80-.).* **2014**, *346* (6206), 171–172.

(20) LeCun, Y. A.; Bengio, Y.; Hinton, G. E. Deep Learning. *Nature* **2015**, *521* (7553), 436–444.

(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-.).* **2018**, *360* (6385), 186–190.

(22) Klucznik, T.; Mikulak-Klucznik, B.; McCormack, M. P.; Lima, H.; Szymkuć, S.; Bhowmick, M.; Molga, K.; Zhou, Y.; Rickershauser, L.; Gajewska, E. P.; et al. Efficient Syntheses of Diverse, Medicinally Relevant Targets Planned by Computer and Executed in the Laboratory. *Chem* **2018**, *4* (3), 522–532.

(23) Schütt, K. T.; Arbabzadah, F.; Chmiela, S.; Müller, K. R.; Tkatchenko, A. Quantum-Chemical Insights from Deep Tensor Neural Networks. *Nat. Commun.* **2017**, *8*, 13890.

(24) Chmiela, S.; Tkatchenko, A.; Sauceda, H. E.; Poltavsky, I.; Schütt, K. T.; Müller, K.-R. Machine Learning of Accurate Energy-Conserving Molecular Force Fields. *Sci. Adv.* **2017**, *3* (5), e1603015.

(25) 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).

(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.* **2017**, *9* (8), 2261–2269.

(27) Behler, J. First Principles Neural Network Potentials for Reactive Simulations of Large Molecular and Condensed Systems. *Angew. Chemie Int. Ed.* **2017**, *56* (42), 12828–12840.

(28) Li, Z.; Kermode, J. R.; De Vita, A. Molecular Dynamics with On-the-Fly Machine Learning of Quantum-Mechanical Forces. *Phys. Rev. Lett.* **2015**, *114* (9), 096405.

(29) Glielmo, A.; Sollich, P.; De Vita, A. Accurate Interatomic Force Fields via Machine Learning with Covariant Kernels. *Phys. Rev. B* **2017**, *95* (21), 214302.

(30) Kruglov, I.; Sergeev, O.; Yanilkin, A.; Oganov, A. R. Energy-Free Machine Learning Force Field for Aluminum. *Sci. Rep.* **2017**, *7* (1), 8512.

(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.* **2012**, *108* (5), 58301.

(32) Faber, F. A.; Hutchison, L.; Huang, B.; Gilmer, J.; Schoenholz, S. S.; Dahl, G. E.; Vinyals, O.; Kearnes, S.; Riley, P. F.; von Lilienfeld, O. A. Prediction Errors of Molecular Machine Learning Models Lower than Hybrid DFT Error. *J. Chem. Theory Comput.* **2017**, acs.jctc.7b00577.

(33) Bleiziffer, P.; Schaller, K.; Riniker, S. Machine Learning of Partial Charges Derived from High-Quality Quantum-Mechanical Calculations. *J. Chem. Inf. Model.* **2018**, *58* (3), 579–590.

(34) 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*. October 16, 2017, pp 4714–4758.

(35) Nebgen, B.; Lubbers, N.; Smith, J. S.; Sifain, A.; Lokhov, A.; Isayev, O.; Roitberg, A.; Barros, K.; Tretiak, S. Transferable Molecular Charge Assignment Using Deep Neural Networks. *arXiv* **2018**, Preprint at https://arxiv.org/abs/1803.04395.

(36) Sifain, Andrew E.; Lubbers, Nicholas; Nebgen, Benjamin T.; Smith, Justin S.; Lokhov, Andrey Y.; Isayev, Olexandr; Roitberg, Adrian E.; Barros, Kipton; Tretiak, S. Discovering a Transferable Charge Assignment Model Using Machine Learning. *ChemRxiv* **2018**, Preprint at https://doi.org/10.26434/chemrxiv.6638.

(37) Gastegger, M.; Behler, J.; Marquetand, P. Machine Learning Molecular Dynamics for the Simulation of Infrared Spectra. *Chem. Sci.* **2017**, *8* (10), 6924–6935.

(38) Ramprasad, R.; Batra, R.; Pilania, G.; Mannodi-Kanakkithodi, A.; Kim, C. Machine Learning in Materials Informatics: Recent Applications and Prospects. *npj Comput. Mater.* **2017**, *3* (1), 54.

(39) Raccuglia, P.; Elbert, K. C.; Adler, P. D. F.; Falk, C.; Wenny, M. B.; Mollo, A.; Zeller, M.; Friedler, S. A.; Schrier, J.; Norquist, A. J. Machine-Learning-Assisted Materials Discovery Using Failed Experiments. *Nature* **2016**, *533* (7601), 73–76.

(40) Isayev, O.; Oses, C.; Toher, C.; Gossett, E.; Curtarolo, S.; Tropsha, A. Universal Fragment Descriptors for Predicting Properties of Inorganic Crystals. *Nature Communications*. Nature Publishing Group June 5, 2017, p 15679.

(41) Ragoza, M.; Hochuli, J.; Idrobo, E.; Sunseri, J.; Koes, D. R. Protein-Ligand Scoring with Convolutional Neural Networks. *J. Chem. Inf. Model.* **2017**, *57* (4), 942–957.

(42) 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.

(43) 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* **2017**, *4*, 170193.

(44) 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.* **2018**, *14* (5), 2341–2352.

(45) 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**, *148* (24).

(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.* **2015**, *11* (5), 2087–2096.

(47) Bartók, A. P.; De, S.; Poelking, C.; Bernstein, N.; Kermode, J. R.; Csányi, G.; Ceriotti, M. Machine Learning Unifies the Modeling of Materials and Molecules. *Sci. Adv.* **2017**, *3* (12), e1701816.

(48) Chmiela, S.; Sauceda, H. E.; Müller, K.-R.; Tkatchenko, A. Towards Exact Molecular Dynamics Simulations with Machine-Learned Force Fields. *arXiv* **2018**, Preprint at https://arxiv.org/abs/1802.09238.

(49) Taylor, M. E.; Stone, P. Transfer Learning for Reinforcement Learning Domains : A Survey. *J. Mach. Learn. Res.* **2009**, *10* (Jul), 1633–1685.

(50) Pan, S. J.; Yang, Q. A Survey on Transfer Learning. *IEEE Transactions on Knowledge and Data Engineering*. October 2010, pp 1345–1359.

(51) Turki, T.; Wei, Z.; Wang, J. T. L. Transfer Learning Approaches to Improve Drug Sensitivity Prediction in Multiple Myeloma Patients. *IEEE Access* **2017**, *5*, 7381–7393.

(52) Rosenbaum, L.; Dörr, A.; Bauer, M. R.; Frankmboeckler; Zell, A. Inferring Multi-Target Qsar

Models with Taxonomy-Based Multi-Task Learning. *J. Cheminform.* **2013**, *5* (7), 33.

(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*; ACM Press: New York, New York, USA, 2007; pp 193–200.

(54) Hjorth Larsen, A.; JØrgen Mortensen, J.; Blomqvist, J.; Castelli, I. E.; Christensen, R.; Dułak, M.; Friis, J.; Groves, M. N.; Hammer, B.; Hargus, C.; et al. The Atomic Simulation Environment - A Python Library for Working with Atoms. *Journal of Physics Condensed Matter*. IOP Publishing July 12, 2017, p 273002.

(55) Riplinger, 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.* **2016**, *144* (2).

(56) Neese, F. The ORCA Program System. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **2012**, *2* (1), 73–78.

(57) Adler, T. B.; Knizia, G.; Werner, H. J. A Simple and Efficient CCSD(T)-F12 Approximation. *J. Chem. Phys.* **2007**, *127* (22), 221106.

(58) 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.

(59) 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.

(60) Chai, J. Da; Head-Gordon, M. Systematic Optimization of Long-Range Corrected Hybrid Density Functionals. *J. Chem. Phys.* **2008**, *128* (8), 084106.

(61) Seung, H. S.; Opper, M.; Sompolinsky, H. Query by Committee. In *Proceedings of the fifth annual workshop on Computational learning theory - COLT '92*; ACM Press: New York, New York, USA, 1992; pp 287–294.

(62) 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.

(63) 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.

(64) 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.

(65) 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**, *148* (24), 241733.

(66) Pordes, R.; Petravick, D.; Kramer, B.; Olson, D.; Livny, M.; Roy, A.; Avery, P.; Blackburn, K.; Wenaus, T.; Würthwein, F.; et al. The Open Science Grid. In *Journal of Physics: Conference Series*; IOP Publishing, 2007; Vol. 78, p 012057.

(67) Sfiligoi, I.; Bradley, D. C.; Holzman, B.; Mhashilkar, P.; Padhi, S.; Würthwein, F. The Pilot Way to Grid Resources Using GlideinWMS. In *2009 WRI World Congress on Computer Science and*

Information Engineering, CSIE 2009; IEEE, 2009; Vol. 2, pp 428–432.

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. 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 (20180213ER). 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[66,67] which is supported by the National Science Foundation award 1148698, and the U.S. DOE Office of Science.

Supplementary Materials:
Methods
Table S1 – S7
Fig S1 – S4
References (1 – 21)

Supplementary information for: "Outsmarting quantum chemistry with transfer learning"

Justin S. Smith[1,2,3,†], Ben Nebgen[2,4,†], Roman Zubatyuk[2,5,†], Nicholas Lubbers[2,3], Christian Devereux[1], Kipton Barros[2], 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)

S1 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.[9] They were benchmarked against two high-quality and distinct datasets: S66[10] for intermolecular interaction energies relevant to biomolecules and FH51[11] 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 for S66 and from 0.38 to 0.16 kcal/mol for HF51, compared to NormalPNO.[9] 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"[12] extrapolation scheme originally proposed by Hobza and coworkers[13]. 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 Halkier[14] and Helgaker[15] (α_{34} = 5.46, β_{34} = 3.05 [16]). The difference between CCSD(T) and MP2 energies was estimated with the cc-pVTZ basis set as shown in Equation 1.

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

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 Eq. 1 was estimated with 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 benchmark[17] for weak intermolecular interactions and the extensive W4-11 benchmark[18] for thermochemistry. CCSD(T)*/CBS performs excellent on both benchmarks (See Table S1).

In Table S1, the extrapolation schemes are noted with the "small" basis set, which is used to calculate ΔCCSD(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 ΔTightPNO term is the difference between TightPNO and NormalPNO energy using a basis set with lower cardinal number. The data in 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 ΔTightPNO approach helps to achieve much lower error, compared to NormalPNO, without sacrificing much computational efficiency.

Table S1: 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/mol**	
	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 [17]
[c] Reference data from [18], averaged for all subsets of W4-11.
[d] Original reference data from [10]
[e] Revised S66 data from [19]

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. Table S2 provides details of each model architecture used in this work.

Table S2: Fully connected neural network architectures for each atom type.

Hydrogen Network Architecture				
	Layer1	**Layer2**	**Layer3**	**Layer4**
Nodes	160	128	96	1
Activation	CELU[6]	CELU	CELU	Linear
Regularization	L2 (1.0E-4)	L2 (1.0E-5)	L2 (1.0E-6)	None
Carbon Network Architecture				
	Layer1	**Layer2**	**Layer3**	**Layer4**
Nodes	144	112	96	1
Activation	CELU	CELU	CELU	Linear
Regularization	L2 (1.0E-4)	L2 (1.0E-5)	L2 (1.0E-6)	None
Oxygen and Nitrogen Network Architecture				
	Layer1	**Layer2**	**Layer3**	**Layer4**
Nodes	128	112	96	1
Activation	CELU	CELU	CELU	Linear
Regularization	L2 (1.0E-4)	L2 (1.0E-5)	L2 (1.0E-6)	None

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.[7] 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
- $\eta^{Radial} = [16]$
- $R_s^{Radial} = [0.900000, 1.168750, 1.437500, 1.706250, 1.975000, 2.243750, 2.51250, 2.781250, 3.050000, 3.318750, 3.587500, 3.856250, 4.125000, 4.39375, 4.662500, 4.931250]$ Å

Angular Parameters:

- Angular cutoff = 3.5 Å
- $\eta^{Angular}$ = [8]
- R_s^{Radial} = [0.900000, 1.550000, 2.200000, 2.850000] Å
- ζ = [32.00000]
- θ_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[5,7] 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
- ADAM[8] 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

Table S2 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.

Table S2: 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**

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[1–3] database containing less than 20 total atoms. We also include the SMILES strings for molecules in the Genentech torsion benchmark[4] 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 $\hat{\rho} = 0.2$ kcal/mol is used in this work)[5] 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 0.2Å 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. Figure S1 compares ANI-1ccx and ANI-1x with and without the active learning generated dihedral corrections.

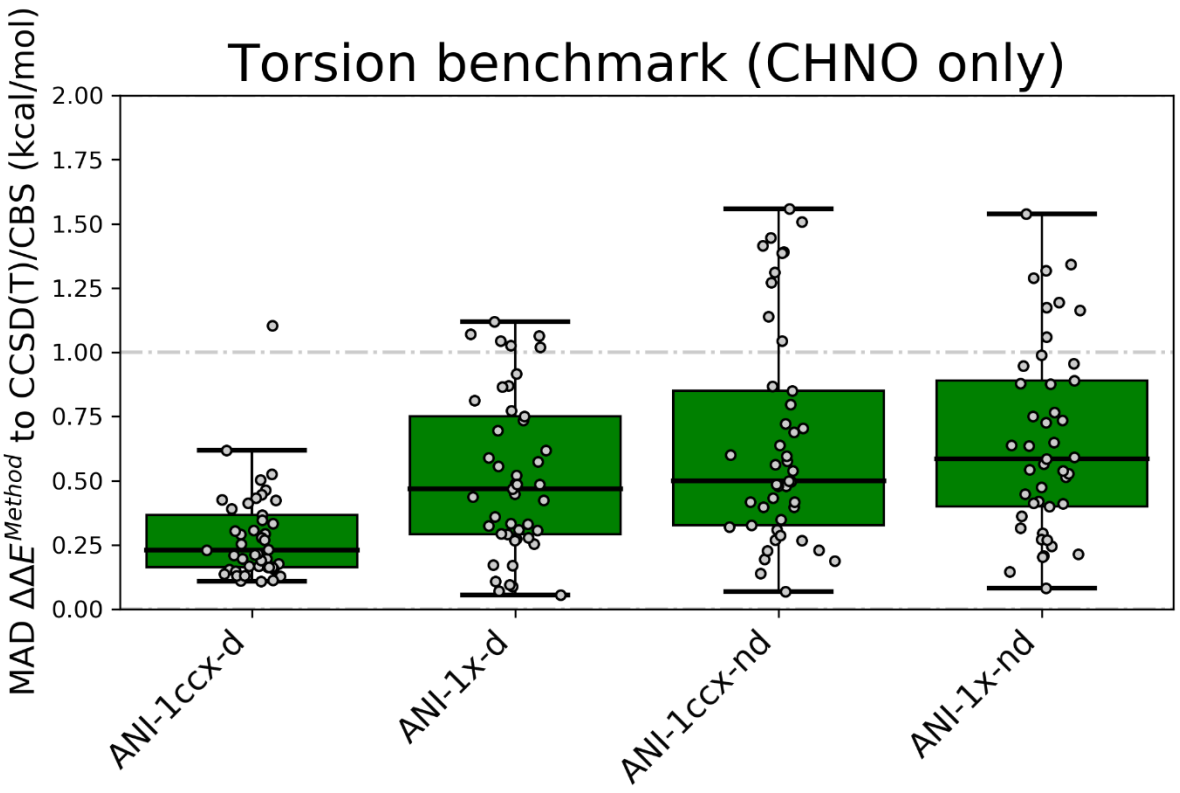


Figure S1: Comparison of ANI model performance with and without transfer learning and dihedral active learning reparameterization on 45 torsion profiles containing the atomic elements C, H, N, and O. The ANI potentials with '-d' appended are reparametrized for better torsions and with '-nd' represent 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.

Table S3. Computed conformer ΔE on the GDB-10to13 benchmark (all conformers within 100 kcal/mol 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.

	ANI-1ccx	ANI-1ccx-Δ
MAD	1.46±0.02	1.44±0.04
RMSD	2.07±0.04	2.04±0.15

Table S4. GDB-10to13 *benchmark results comparing various ANI potentials and DFT.* Errors for conformer energy differences (**ΔE**) and potential energies (**E**) for all ANI potentials. ωB97x-D3/6-31G* 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.

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
ωB97x-D3/6-31G*	3.26	--	4.99	--	--	--	--	--

Table S5: 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.

	CCSD(T)*/CBS	***ANI-1ccx***	***ωb97x/6-31g(d)***	***ANI-1x***
HC7	1.5/1.8	1.9/2.0	16.0/21.7	17.8/23.1
ISOL6	1.0/1.1	1.5/1.6	4.0/4.7	5.1/5.6

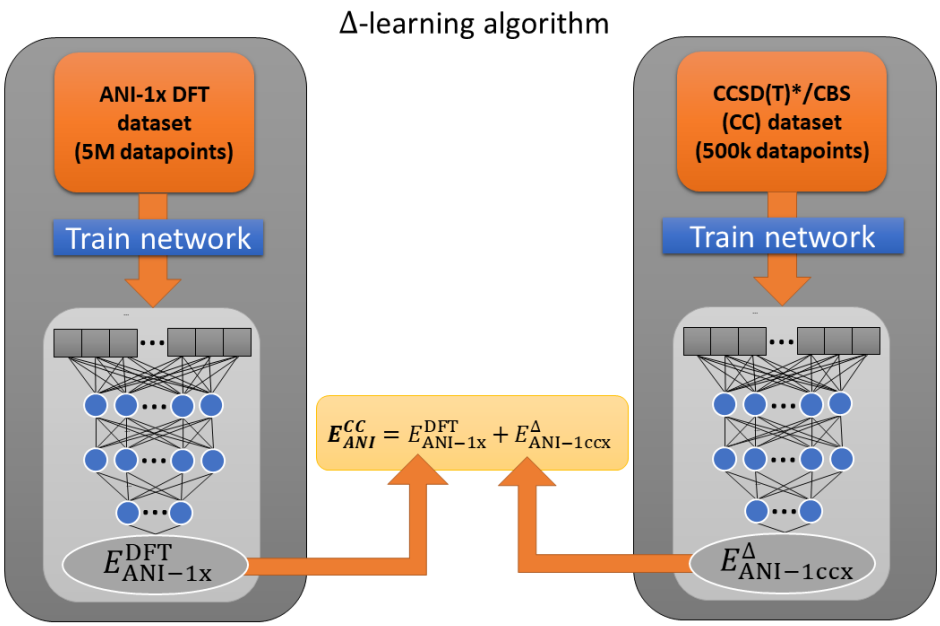


Figure S2: Diagram of the delta learning techniques evaluated in this work. Δ-learning uses the pretrained ANI DFT model to predict an energy, then a second network is trained to correct the DFT prediction to better predict the CCSD(T)/CBS-extrapolated data. The resulting scheme requires two networks for prediction and is thus more expensive than the transfer learning.

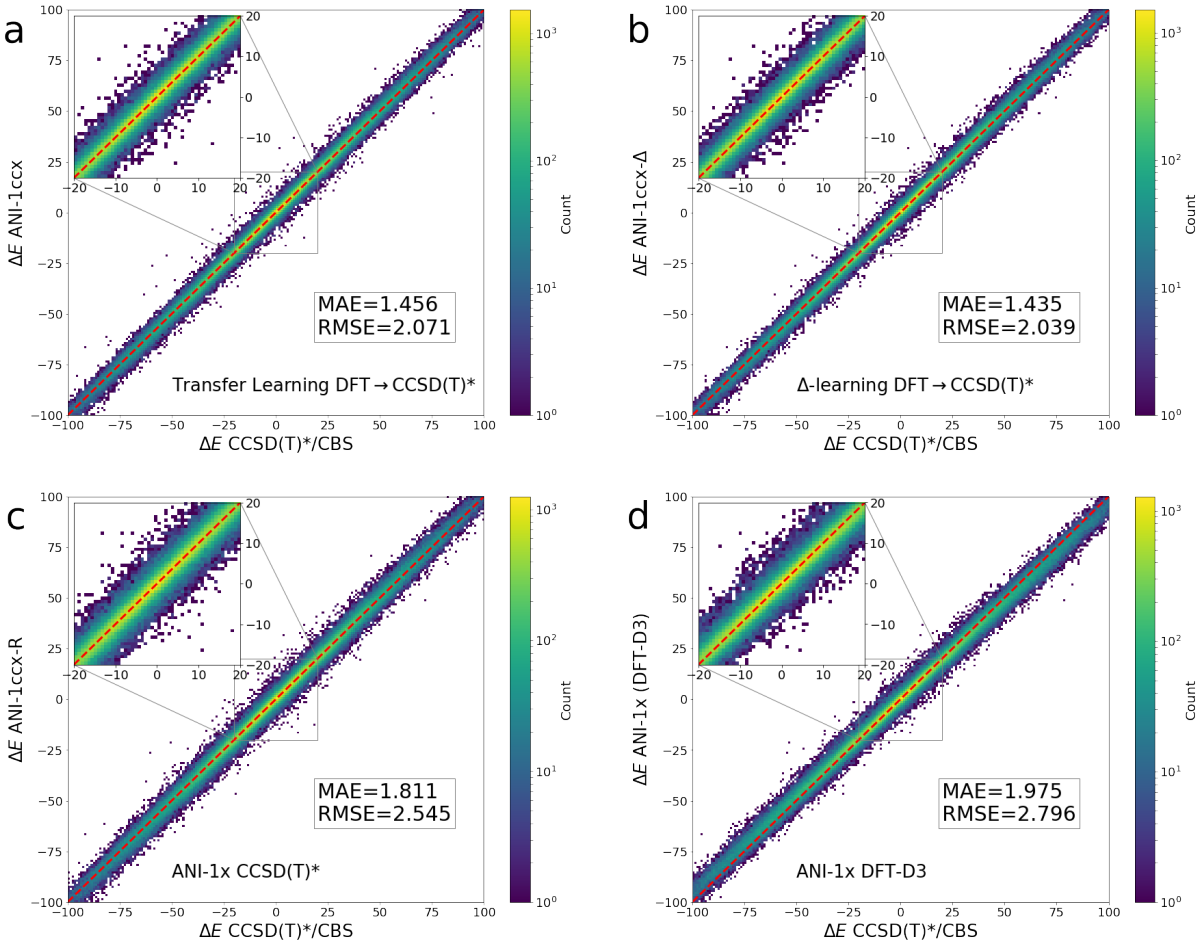


Figure S3: Comparison of four different ANI potentials at predicting relative energies (ΔE) between all conformers for each molecule in the GDB-10to13 (CCSD(T)/CBS* computed) benchmark. These log-scale density correlation plots show a) ANI-1ccx (transfer learning-based CCSD(T)*/CBS trained model), b) 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).

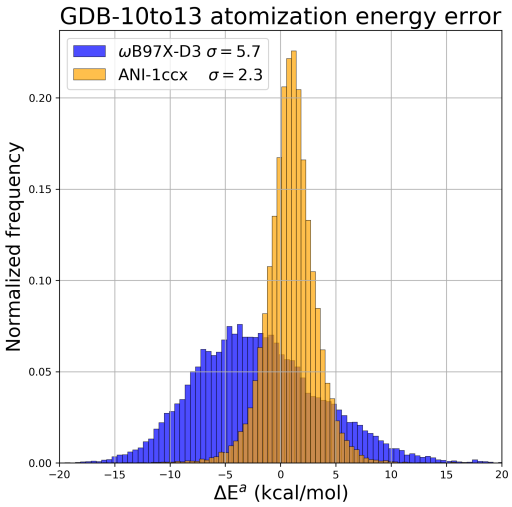
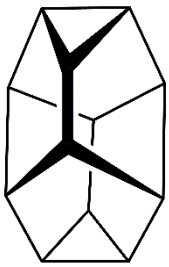
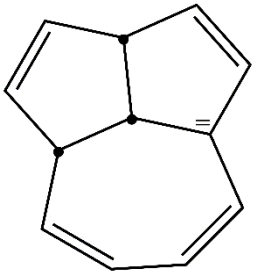


Figure S4: Per atom corrected atomization energy difference (ΔE^a) distribution for ωB97X-D3/6-31G* and ANI-1ccx vs. CCSD(T)*/CBS reference data. An average correction of 33 kcal/mol was applied to the DFT data so it better fits the CCSD(T)*/CBS atomization energies. This correction came from a non-trivial 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.

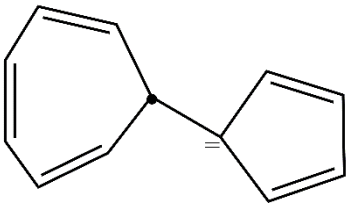
Table S6: Reference and calculated energies (kcal/mol) for the HC7/11 benchmark. Reference calculations are obtained from Peverati, Zhao, and Truhlar [a]. The rows of the table correspond to the index given in the HC7/11 portion of Figure 3 in the main article.



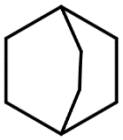
E1 (1)



E2 (22)



E3 (31)



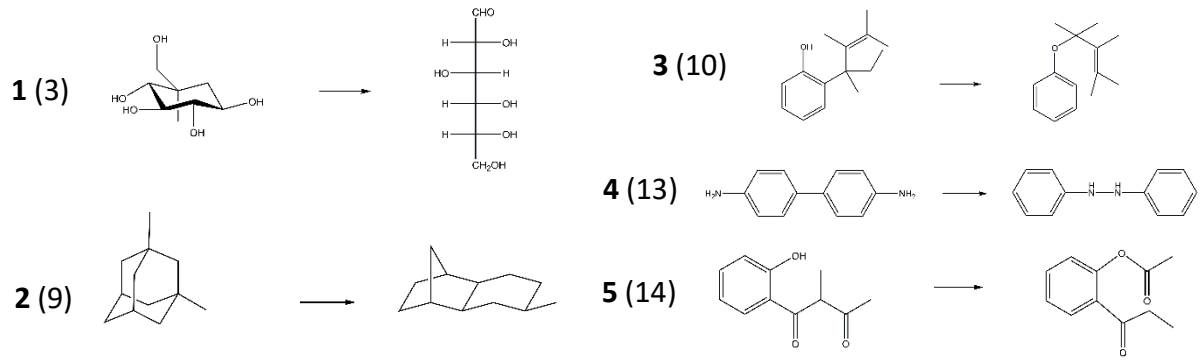
E4

(Bicyclo[2.2.2]octane)

Reaction details	Reference[a]	ANI-1ccx	CCSD(T)*/CBS	ANI-1x	ωb97x/6-31g(d)
E2 → E1	14.34	15.62	13.78	34.60	28.18
E3 → E1	25.02	27.74	23.12	48.33	39.66
Octane-a → Octane-b	1.90	0.37	-1.25	2.65	1.70
$4CH_4 + C_6H_{14} \rightarrow 5C_2H_6$	9.81	7.86	8.73	7.18	6.27
$6CH_4 + C_8H_{18} \rightarrow 7C_2H_6$	14.84	11.93	13.06	10.76	9.31
Adamantane → $3CH_4 + 2C_2H_2$	193.99	196.16	195.92	236.80	237.96
E4 → $3CH_4 + 2C_2H_2$	127.22	127.76	126.87	157.66	157.28

[a] Reference data from[20]

Table S7: Reference and calculated energies (kcal/mol) 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	ωb97x/6-31g(d)
1	9.77	8.22	10.54	5.23	8.53
2	21.76	20.95	21.23	20.15	20.69
3	6.82	4.49	4.90	-2.08	0.75
4	33.52	35.20	34.06	27.93	26.50
5	5.30	6.43	6.30	0.45	0.48

[a] Reference data from[21]
[b] Reference data from[20]

(1) 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.

(2) 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.

(3) 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.

(4) 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.

(5) 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**, *148* (24), 241733.

(6) Barron, J. T. Continuously Differentiable Exponential Linear Units. *arXiv* **2017**, Preprint at https://arxiv.org/abs/1704.07483.

(7) 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.

(8) Kingma, D.; Ba, J. Adam: A Method for Stochastic Optimization. *arXiv* **2014**, Preprint at https://arxiv.org/abs/1412.6980.

(9) 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.

(10) 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.

(11) 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.

(12) 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.

(13) 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.

(14) 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.

(15) Helgaker, T.; Klopper, W.; Koch, H.; Noga, J. Basis-Set Convergence of Correlated Calculations on Water. *J. Chem. Phys.* **1997**, *106* (23), 9639–9646.

(16) 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.

(17) 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.

(18) 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.

(19) 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.

(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.

(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.