scientific data

Check for updates

OPEN

DATA DESCRIPTOR

The QDπ dataset, training data for drug-like molecules and biopolymer fragments and their interactions

Jinzhe Zeng[1], Timothy J. Giese[1], Andreas W. Götz[2] & Darrin M. York[1]✉

The development of universal machine learning potentials (MLP) for small organic and drug-like molecules requires large, accurate datasets that span diverse chemical spaces. In this study, we introduce the QDπ dataset which incorporates data taken from several datasets. We use a query—by—committee active learning strategy to extract data from large datasets to maximize the diversity and avoid redundancy as relevant for neural network training to construct the QDπ dataset. The QDπ dataset requires only 1.6 million structures to express the chemical diversity of 13 elements from the various source datasets at the ωB97M-D3(BJ)/def2-TZVPPD level of theory. The QDπ dataset enables creation of flexible target loss functions for neural network training relevant to drug discovery, including information-dense data sets of relative conformational energies and barriers, intermolecular interactions, tautomers and relative protonation energies of drug-like compounds and biomolecular fragments. It is the hope that the high chemical information density and diversity contained in the QDπ dataset will provide a valuable resource for the development of new universal MLPs for drug discovery.

Background & Summary

The development and use of machine learning potentials (MLPs) for molecular simulations has seen a surge in interest[1–6], particularly in drug discovery applications that screen a large number of small organic molecules[7]. Many of the molecules encountered in the screening process may not have well—established, mature molecular mechanical force field parameters, and some may not have ever been synthesized before. This has led to interest in training universal MLP models that accurately reproduce ab initio energies and forces for a large diversity of molecules[8–14] and chemical systems[15,16] by considering the enormous number of possible atomic permutations, combinations, and conformational isomers[17,18]. The training of universal MLP models therefore requires extensive and accurate datasets that sample the diverse chemical space of organic and drug-like molecules.

For many years, benchmark datasets have been prepared that were intended to compare the quality of density functional and semiempirical quantum mechanical (QM) methods against highly accurate ab initio results;[19–23] however, the diversity of compounds contained in these datasets is far too narrow to train a universal MLP. This motivated the creation of very large datasets of stable conformations (geometry optimized structures), including the QMugs dataset[24], the QM40 dataset[25], various GDB datasets[17,26], and subsets[27,28]. An alternative strategy has been to construct datasets by collecting samples drawn from molecular dynamics (MD) simulation[29,30], such that the dataset contains thermally accessible conformations to improve MLP accuracy. Recent datasets such as ANI-1[31], OrbNet Denali[32], QM7-X[33], AIMNet-NSE[34], and SPICE[35] combine the two data generation methods by including a large number of geometry optimized chemical species and thermally-accessible structures.

It has become exceedingly expensive to calculate target molecular energies and atomic forces with ab initio QM methods as the size of the datasets have grown. Therefore, active learning strategies have been employed to remove redundant information within the datasets to limit their size without sacrificing chemical diversity[36,37]. The active learning strategy has been used to create datasets such as ANI-1x[38], ANI-2x[12], and the work by Yang *et al.*[39] These datasets have some limitations; for example, although the ANI-1ccx dataset[38] is calculated at the very accurate CCSD(T) level in the complete basis set limit, it lacks atomic forces, which has been found to be an important target property for model training[40]. Furthermore, the ANI-1x and ANI-2x datasets collect data with the ωB97X/6-31G* method[41] however, this method was later found[42] to produce atomic forces that differed

[1]Laboratory for Biomolecular Simulation Research, Institute for Quantitative Biomedicine, and Department of Chemistry and Chemical Biology, Rutgers University, Piscataway, NJ, 08854-8087, USA. [2]San Diego Supercomputer Center, University of California San Diego, La Jolla, CA, 92093, USA. ✉e-mail: Darrin.York@rutgers.edu

from ωB97M-D3(BJ)/def2-TZVPPD — one of the most accurate and robust density functional methods in the extensive GMTKN55 benchmark database comparisons[23] — by more than 2 kcal/mol/Å in mean squared error (MAE). In contrast, the SPICE dataset is calculated with the accurate ωB97M-D3(BJ)/def2-TZVPPD method but, as we shall demonstrate, the SPICE dataset does not cover the full chemical space expressed by the ANI datasets.

The present work introduces the the Quantum Deep Potential Interaction (QDπ) dataset for drug discovery force field development. The QDπ dataset contains 1.6 million molecular structures to express the chemical diversity of 13 elements, and the energies and forces are calculated with the accurate ωB97M-D3(BJ)/def2-TZVPPD method. Molecular conformations were taken from various source datasets including SPICE[35], ANI[12,38], GEOM[18], FreeSolv[43], RE[14], and COMP6[36]. We describe several strategies that were used to select structures in a manner that maximizes the chemical diversity while minimizing the number of expensive ab initio evaluations. These strategies include direct inclusion of a source dataset, relabeling small datasets, using active learning to prune large datasets[36], and using a combination of molecular dynamics and active learning to extend the breadth of very small datasets[36,37]. The goal of an active learning procedure is not to extract a unique subset of structures; instead, the goal is to extract the relevant information from a population that contains redundancies. Statistical analysis of the QDπ dataset shows that it offers more coverage than the individual SPICE and ANI datasets. Furthermore, the active learning procedure is shown to be an effective method to avoid including redundant training information from multiple datasets without sacrificing chemical diversity.

Methods

The QDπ dataset was constructed during the development of the QDπ-2 MLP model[42]. As previously mentioned, the QDπ dataset combines and extends several existing databases using a consistent ab initio reference theory. Specifically, the reference theory is the ωB97M-D3(BJ)/def2-TZVPPD Hamiltonian[44–47], as implemented in the PSI4 v1.7 software[48,49]. Many of the existing MLP datasets contain millions of molecular structures; however, they also contain a considerable amount of redundant information. We begin the discussion by introducing strategies for incorporating or expanding upon data from existing databases to limit the overall number of data points while retaining the chemical and structural information needed to accurately train a robust model. We then detail the contents of the existing databases serving as the foundation for the QDπ dataset.

Data generation. *Data selection methods.* We employed 4 strategies for incorporating molecular structures into the QDπ dataset from existing databases.

- Direct inclusion. If a source database is a collection of energies and forces evaluated at the ωB97M-D3(BJ)/def2-TZVPPD level of theory, then the entire database is directly incorporated within the QDπ dataset.
- Relabeling. If the source database does not provide ωB97M-D3(BJ)/def2-TZVPPD reference data, but the number of structures within the database is reasonably small, then we recalculate the energies and forces of each structure at ωB97M-D3(BJ)/def2-TZVPPD and include the results into the QDπ dataset. The geometries are not reoptimized at the reference level of theory because the purpose of the QDπ dataset is to train MLPs for use in MD simulation. As explained below, we perform some MD simulation specifically to extend the configurational space of the reference data.
- Active learning strategy to prune large datasets. If the source database contains a large number of structures, then it becomes impractical to recalculate the entire database at the ωB97M-D3(BJ)/def2-TZVPPD level of theory. We instead use a query—by—committee active learning strategy to reduce the number of ab initio calculations[36] by identifying and eliminating structures that do not introduce a significant amount of new information to train against. Each active learning cycle involves the training of 4 independent MLP models against the developing QDπ dataset with different random seeds. The energy and force standard deviations between the 4 models is calculated for each structure in the source database. If the energy and force standard deviations are below 0.015 eV/atom and 0.20 eV/Å, respectively, then the structure is not added to the QDπ dataset. A random subset of up to 20,000 structures from the remaining candidates is selected for labeling with ωB97M-D3(BJ)/def2-TZVPPD and included within the QDπ dataset. The thresholds on the standard deviation were chosen because, in the first cycle, the initial set of trained models used to bootstrap the active learning procedure produces standard deviations less than these values. These thresholds cannot be made arbitrarily small because the model's accuracy is inherently limited by its capacity to fit the data[50]. The active learning procedure terminates when each structure within the source database has either been included within or excluded from the QDπ dataset. We have implemented this strategy in the DP-GEN software[50].
- Active learning strategy to extend small datasets. If a source database consists of only a few optimized structures, we employ an active learning strategy with molecular dynamics (MD) simulation to search for additional thermally accessible conformations to include within the QDπ dataset[36,37]. At each active learning cycle, MD sampling is performed with each molecule in the source database using 1 of the 4 MLP models. The simulation length and sampling frequency varies between databases. A configuration is rejected if the energy and force standard deviations between the models are below 0.015 eV/atom and 0.20 eV/Å, respectively. A random subset of up to 20,000 structures from the candidate samples is selected for labeling and included in the QDπ dataset. The active learning procedure is terminated when the 4 models agree to within the specified tolerance for all explored samples. The details of molecular dynamics simulations vary depending on the subset and will be provided later.

MLP models used in the active learning methods. The active learning procedures were performed to either prune or expand existing datasets. For this purpose, the active learning procedures were used to train a semiempirical

Name	Elements	Conformations		Selection method
		before selection	after selection	
SPICE	H Li C N O F Na P S Cl K Br I	997570	997570	Direct inclusion
ANI	H C N O F S Cl	6046683	324294	Active learning strategy to prune large datasets
GEOM	H C N O F P S Cl Br I	31224028	23579	Active learning strategy to prune large datasets
FreeSolv-MD	H C N O F P S Cl Br	302400	76696	Active learning strategy to extend small datasets
RE	H C N O	3667	3667	Relabeling
RE-MD	H C N O F P S Cl Br	2217000	12019	Active learning strategy to extend small datasets
COMP6	H C N O	99317	99317	Relabeling
Total	H Li C N O F Na P S Cl K Br I	40890665	1537142	(3.8 %)

Table 1. The overall content of the QDπ neutral dataset.

Name	Elements	Conformations before/after selection		Selection method
SPICE	H Li C N O F Na P S Cl K Br I	107482	107482	Direct inclusion
RE	H C N O	616	616	Relabeling
RE-MD	H C N O F P S Cl Br	176000	4128	Active learning strategy to extend small datasets
Total	H Li C N O F Na P S Cl K Br I	284098	112226	(39.5 %)

Table 2. The overall content of the QDπ charged dataset.

quantum mechanical (SQM)/Δ MLP model[14]. A SQM/Δ MLP model supplements a standard semiempirical QM (or QM/MM) calculation with a MLP that is trained to reproduce the difference between SQM and ab initio energies and forces. The Δ MLP strategy is amenable to condensed phase QM/MM applications; the long-range electrostatics are evaluated with the inexpensive SQM potential and the Δ MLP is a short-range nonelectrostatic correction. One must also consider long-range dispersion interactions within QM/MM applications; however, in the present work we prepared a SQM/Δ MLP model for calculating energies and forces of small molecules. The demonstration does not parametrize QM/MM interactions, which is beyond the scope of this work. By building a Δ MLP correction for a SQM model, the MLP is not solely responsible for distinguishing between charged and neutral species; the physics of the underlying SQM potential models the distinction[14]. The SQM/Δ MLP models use the DFTB3/3ob Hamiltonian[51,52], and a two-body embedding DeepPot-SE model[53] with type embedding. The DeepPot-SE model is implemented in the DeePMD-kit software[54–56]; the mathematical expressions and an extended description of the neural network can be found in Ref. [56]. The atomic descriptors were calculated with a 6 Å cutoff radius and a 1 Å switching layer that ensures the energy and force corrections smoothly approach zero at the cutoff. The number of neurons in the embedding network, the fitting network, and the type embedding network are (25, 50, 100), (256, 256, 256, 1), and (8), respectively. The embedding submatrix contains 12 channels, and all networks use single-point precision.

Content of the QDπ dataset. As previously stated, the structures within the QDπ dataset were taken from or expanded upon existing "source" datasets. Brief summaries of the source datasets are provided below. The QDπ dataset is partitioned into subsets containing neutral and charged molecules. Charged molecules are included because most of the existing MLP models poorly describe multiple charge/protonation states. Tables 1 and 2 summarize the overall content of neutral and charged QDπ subsets, respectively.

SPICE. The small-molecule/protein interaction chemical energies (SPICE)[35] dataset (v1.1.3) contains 1.1 million structures of dipeptides, solvated amino acids, DES370k monomers and dimers[57], molecules taken from PubChem, and ion pairs. The chemical space includes 15 elements (H, Li, C, N, O, F, Na, P, S, Cl, K, Br, and I), and the reference energies and forces were collected with the ωB97M-D3(BJ)/def2-TZVPPD Hamiltonian. This is the same *ab initio* reference theory used in the QDπ dataset. One could therefore include the SPICE data within the QDπ dataset without modification; however, we exclude some high energy outlier structures using the strategy discussed in Ref. [14]. Some of the high energy structures include situations where a covalent bond has been inadvertently broken, and the resulting diradical system may not be adequately modeled with a single determinant restricted wavefunction. The outliers are detected by grouping all conformations of a molecule, calculating the potential energy mean and standard deviation, and excluding a conformation if it differs by more than 8 standard deviations from the mean. Upon excluding the high energy structures, there are 997,570 remaining neutral structures and 107,482 charged structures which were directly added to the QDπ dataset.

SPICE v1.1.3 was the latest version when this work was conducted. SPICE has released a new major version (v2)[58] which could be considered adding into the later version of the QDπ dataset.

ANI. The accurate neural network engine for molecular energies (ANI) database is composed of ANI-1x[38] and ANI-2x[12] datasets. The chemical space includes 7 elements (H, C, N, O, F, S, and Cl), and the energies were originally prepared with the ωB97X/6-31G* Hamiltonian[41]. We excluded high energy configurations using the same strategy described for the SPICE dataset. After removing the outliers, the ANI dataset had 6,046,683 remaining structures. It was deemed too costly to relabel all the configurations with ωB97M-D3(BJ)/def2-TZVPPD; therefore, we used the active learning strategy to prune the available data. The active learning strategy relabeled 324,294 conformations (5.4%) for inclusion into the QDπ dataset.

GEOM. The Geometric Ensemble Of Molecules (GEOM) database[18] contains 37 million conformers of more than 450,000 organic molecules. The chemical space includes 10 elements (H, C, N, O, F, P, S, Cl, Br, and I). The construction of the QDπ dataset focused on the AICures subset of the GEOM database, which is a collection of 31.2 million configurations from 304,466 small-to-medium sized drug molecules. We used the active learning strategy to prune the configurations resulting in the inclusion of 23,579 (0.076%) structures into the QDπ dataset.

FreeSolv. We took 504 small molecules from the FreeSolv[43] database whose chemical space includes 9 elements (H, C, N, O, F, P, S, Cl, and Br). The FreeSolv database includes experimental solvation free energies rather than reference energies and forces of specific structures. To prepare data for the QDπ dataset, we solvated each solute with water in periodic boundary conditions, equilibrated the unit cell density, and used the "active learning from molecular dynamics" strategy to collect relevant conformations. Only the energy and forces of the isolated solute molecule are included in the QDπ dataset, which is taken from the solvated trajectory upon removing all solvent molecules and periodic boundary conditions. Initial solute geometries were prepared with the OpenBabel software[59] from the SMILES representation of molecules provided by FreeSolv. The periodic systems were prepared by solvating the molecules with 1140 4-point OPC waters[60]. The simulations were performed in the isothermal-isobaric ensemble at 298 K and 1 atm for 10 ps using a 1 fs time step, and solute configuration was saved every 50 fs. The solute was modeled with the DFTB3[51,52] QM/MM + QDπ-2 Δ MLP model[42]. The sampling was performed with the sander MD program[61], which we interfaced[42,62] to the DFTB+[63], xtb[64] and DeePMD-kit[56] software packages. The query—by—committee procedure added 76,696 conformations to the QDπ dataset after 3 cycles of active learning.

RE. The relative energy (RE) dataset[14] is a collection of small databases that collect relative energies at the ωB97X/6-31G* level of theory[41]. These include: HB375 × 10[21], AEGIS[65,66], Tautobase[67,68], TAUT15[23], amino acid model compounds[69], nucleic acid model compounds[69], PA26[23], and RegioSQM20[70]. In total, these datasets include 3,667 neutral and 616 charged molecules, which we labeled with ωB97M-D3(BJ)/def2-TZVPPD and included within the QDπ dataset. Furthermore, we extended the QDπ dataset by applying the active learning strategy to brief 1 ps gas phase molecular dynamics simulations of 2,217 unique neutral and 176 unique charged molecules. The active learning procedure produced 12,019 neutral conformations and 4128 charged conformations that were added to the QDπ dataset.

While the QDπ dataset only provides the potential energy of a conformation, the relative energies can be easily calculated by subtracting the potential energies of different conformations.

COMP6. The comprehensive machine-learning potential (COMP6) dataset[36] is a collection of benchmark databases labeled with ωB97X/6-31G*. The chemical space includes 4 elements (H, C, N, and O) in 99,317 conformations. The databases include structures taken from S66 × 8[19,20], ANI-MD, GDB[26,71], Tripeptides, and DrugBank[72]. All conformations were relabeled with ωB97M-D3(BJ)/def2-TZVPPD and added to the QDπ dataset.

Data Records

The QDπ dataset is provided in the DeePMD-kit HDF5 data file format[56], freely accessible from zenodo[73] under the CC BY 4.0 license. The structures (conformations) are organized into "groups" which share the same chemical formula. Data keys in each group are listed in Table 5. The elements array is a petite list of unique element symbols, and each entry in the atomic types array is an integer index of the elements array. The energies, coordinates, forces are stored in units of eV, Å, and eV/Å, respectively.

Technical Validation

Comparison between ωB97M-D3(BJ)/def2-TZVPPD and ωB97X/6-31G*. The developers of the SPICE dataset chose ωB97M-D3(BJ)/def2-TZVPPD because it was regarded as the most accurate density functional methods supported by PSI4 while being affordable enough to be applied within their budget[23]. Subsequent comparisons against the GMTKN55 dataset concluded that this level of theory was one of the best hybrid functionals available, especially for noncovalent interactions[45]. Furthermore, we previously compared ωB97X/6-31G* (the theory used in constructing the ANI datasets) to ωB97M-D3(BJ)/def2-TZVPPD using a series of datasets[42]. We extracted 5% of the data from the SPICE, ANI, GEOM, FreeSolv-MD, and RE-MD datasets and calculated the mean absolute difference in atomic forces between the 2 levels of theory. They were found to differ by 2.40, 3.15, 2.61, 3.17, and 2.50 kcal/mol/Å for the SPICE, ANI, GEOM, FreeSolv-MD, and RE-MD datasets, respectively. These differences substantially exceed the uncertainty in trained MLPs. Consequently, the ωB97M-D3(BJ)/def2-TZVPPD level of theory remains critical for training accurate MLPs.

Element	SPICE	ANI	GEOM	FreeSolv-MD	RE	RE-MD	COMP6	Total
H	14001078	2378509	492397	670342	35322	87919	1277560	18943127
Li	15	0	0	0	0	0	0	15
C	11408098	1347600	410273	542753	17789	81053	838496	14646062
N	2104132	587996	76872	43499	4196	25947	184592	3027234
O	2059694	565892	86665	107505	4648	13331	159822	2997557
F	356626	24344	6996	14986	0	446	0	403398
Na	88	0	0	0	0	0	0	88
P	35694	0	2824	3863	0	40	0	42421
S	484071	131086	20147	8408	0	1197	0	644909
Cl	233601	23253	4147	74704	0	662	0	336367
K	88	0	0	0	0	0	0	88
Br	83161	0	922	4524	0	180	0	88787
I	19609	0	106	0	0	0	0	19715

Table 3. Elements of the QDπ neural dataset.

Element	SPICE	RE	RE-MD	Total
H	1195340	35322	38453	1269115
Li	70	0	0	70
C	717700	17789	32208	767697
N	168369	4196	10829	183394
O	200698	4648	4257	209603
F	11277	0	60	11337
Na	5457	0	0	5457
P	6156	0	40	6196
S	23474	0	80	23554
Cl	13835	0	40	13875
K	6208	0	0	6208
Br	7126	0	0	7126
I	5587	0	0	5587

Table 4. Elements of the QDπ charged dataset.

Name	Key	Unit	Shape
Elements	`type_map.raw`	…	(N_{elements},)
Atomic types	`type.raw`	…	(N_{atoms},)
Coordinates	`set.000/coord.npy`	Å	($N_{\text{conformations}}$, $N_{\text{atoms}} \times 3$)
Energies	`set.000/energy.npy`	eV	($N_{\text{conformations}}$)
Forces	`set.000/force.npy`	eV/Å	($N_{\text{conformations}}$, $N_{\text{atoms}} \times 3$)
Non-PBC marker	`nopbc`	…	()
Net charge	`set.000/net_charge.npy`	1	($N_{\text{conformations}}$)

Table 5. Data keys in each group.

Diversity. Table 3 and 4 shows the number of times the different elements appear in the neutral and charged QDπ datasets, respectively. The most common elements are H, C, N, and O. The rarest elements are alkali and halide ions, which are taken mainly from the SPICE dataset. Figure 1 histograms the molecular size (the number of atoms) of conformations in the QDπ neutral and charged dataset. Most of conformations have fewer than 50 atoms; however, there are a few conformations with more than 100 atoms. There are two peaks in the histogram: one at 10 atoms and the other at 50 atoms, and there is a sudden drop in the population of configurations with more than 50 atoms. The sudden drop in the population at 50 atoms occurs because the SPICE dataset contains PubChem molecules with fewer than 50 atoms[35].

Effect of active learning on chemical diversity. To show the effectiveness of the query—by—committee data selection methodology, we compare the chemical diversity of the SPICE dataset neutral molecules to the original ANI dataset and to the subset of ANI conformations selected from the active learning procedure. The chemical diversity is characterized by calculating the descriptor output vectors of the QDπ-2 model and applying the t-Distributed Stochastic Neighbor Embedding (t-SNE) method[74] to map high-dimensional output vectors to

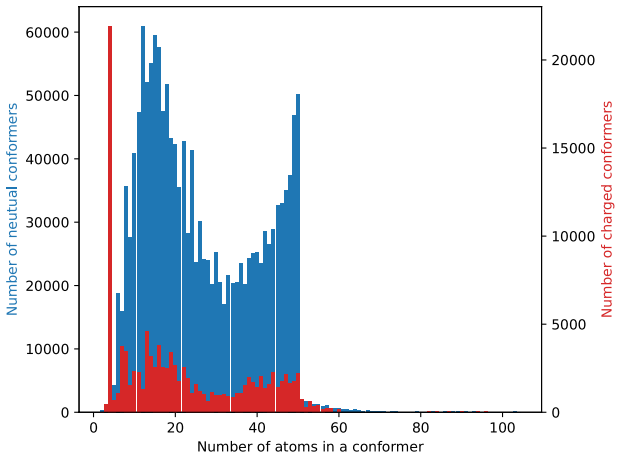

Fig. 1 A histogram of the number of atoms in neutral (left y-axis, in blue) and charged (right y-axis, in red) conformations.

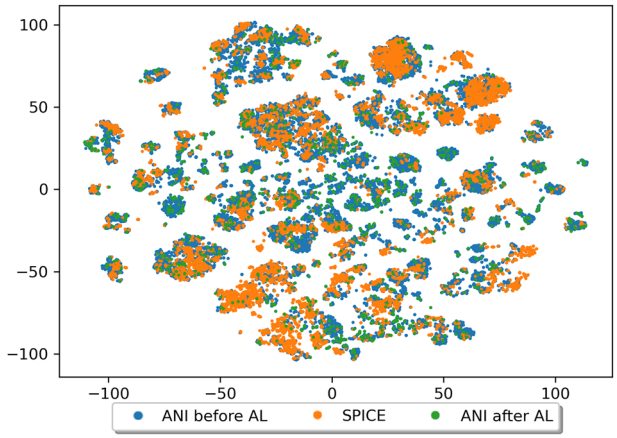

Fig. 2 2D parametric t-SNE embeddings. These embeddings are made from the descriptor of the QDπ-2 model for carbon atoms in the SPICE data set (orange), the ANI data set before active learning (AL) (blue), and the ANI data set after AL (green). The x and y coordinates are not easily amenable to physical interpretation, and arise from the t-SNE method that maps the high-dimensional chemical space to a 2D space such that similar data points appear close to one another in the map.

a 2-dimensional representation. Figure. 2 visualizes the 2-dimensional representation of the chemical diversity of carbon atoms. Although the axis are not easily amenable to physical interpretation, the important characteristic of the visual representation is the presence and separation of distinct clusters. The embedded representation of closely related chemical environments are often projected to similar values; therefore, the presence of multiple clusters is an indication of chemical diversity. There are areas in Fig. 2 where the SPICE (the orange colors) and ANI (the blue colors) distributions overlap with each other, and there are areas where they do not overlap. Areas lacking overlap suggest that the ANI dataset introduces new training information, whereas overlapping areas are indicative that the samples contain redundancies. Given the direct inclusion of the SPICE into QDπ dataset, an active learning procedure was used to extract a subset of samples from ANI. The distribution of samples extracted from ANI (the green colors in Fig. 2) broadly overlaps with the original ANI distribution — there are green dots in most of the blue clusters. The active learning procedure extracts fewer samples from the ANI distribution in areas where it significantly overlaps with the SPICE dataset (the orange clusters), as intended. The chemical diversity of the QDπ dataset includes both SPICE and the extracted ANI samples; therefore, the diversity expressed by the QDπ dataset is broader than the individual SPICE and ANI datasets.

Usage Notes

The dataset can be read and used by the DeePMD-kit package[56]. It can also be loaded and manipulated within python scripts with the aid of the dpdata software.[75] An example Python script which loads and prints the dataset is provided below.

```
import dpdata

# load all subsets

data = dpdata.MultiSystems()

data.from_deepmd_hdf5("data/neutral/spice.hdf5")

data.from_deepmd_hdf5("data/neutral/ani.hdf5")

data.from_deepmd_hdf5("data/neutral/geom.hdf5")

data.from_deepmd_hdf5("data/neutral/freesolvmd.hdf5")

data.from_deepmd_hdf5("data/neutral/re.hdf5")

data.from_deepmd_hdf5("data/neutral/remd.hdf5")

data.from_deepmd_hdf5("data/neutral/comp6.hdf5")

# dump combined data

data.to_deepmd_hdf5("qdpi-1.0.hdf5")

# print the summary of data

print(data)

# get subsystems

subsystems = list(data.systems.values())

# get the data from one of the subsystem

print(subsystems[0].data.keys())

print(subsystems[0].data)
```

Code availability

DP-GEN v0.12.0 (https://github.com/deepmodeling/dpgen) was used to perform active learning. Example DP-GEN training and active learning input files for pruning the ANI datasets can be downloaded at https://gitlab.com/RutgersLBSR/QDpiDataset. Python scripts for data labeling is located within the psi4qdp repository available at https://github.com/njzjz/psi4qdp.

Received: 7 January 2025; Accepted: 8 April 2025;
Published online: 25 April 2025

References

1. Behler, J. Perspective: Machine learning potentials for atomistic simulations. *J. Chem. Phys.* **145**, 170901, https://doi.org/10.1063/1.4966192 (2016).
2. Butler, K. T., Davies, D. W., Cartwright, H., Isayev, O. & Walsh, A. Machine learning for molecular and materials science. *Nature* **559**, 547–555, https://doi.org/10.1038/s41586-018-0337-2 (2018).
3. Noé, F., Tkatchenko, A., Müller, K.-R. & Clementi, C. Machine Learning for Molecular Simulation. *Annu. Rev. Phys. Chem.* **71**, 361–390, https://doi.org/10.1146/annurev-physchem-042018-052331 (2020).
4. Pinheiro Jr, M., Ge, F., Ferré, N., Dral, P. O. & Barbatti, M. Choosing the right molecular machine learning potential. *Chem. Sci.* **12**, 14396–14413, https://doi.org/10.1039/d1sc03564a (2021).
5. Manzhos, S. & Carrington Jr, T. Neural Network Potential Energy Surfaces for Small Molecules and Reactions. *Chem. Rev.* **121**, 10187–10217 (2021).
6. Zeng, J., Cao, L. & Zhu, T. Neural network potentials. In Dral, P. O. (ed.) *Quantum Chemistry in the Age of Machine Learning*, chap. 12, 279–294 (Elsevier, 2022).
7. Polishchuk, P. G., Madzhidov, T. I. & Varnek, A. Estimation of the size of drug-like chemical space based on GDB-17 data. *J. Comput.-Aided Mol. Des.* **27**, 675–9, https://doi.org/10.1007/s10822-013-9672-4 (2013).
8. Schütt, K., Sauceda, H., Kindermans, P., Tkatchenko, A. & Müller, K. SchNet - A Deep Learning Architecture for Molecules and Materials. *J. Chem. Phys.***148**, 241722 (2018).
9. Simeon, G. & De Fabritiis, G. Tensornet: Cartesian tensor representations for efficient learning of molecular potentials. In Oh, A. *et al.* (eds.) *Advances in Neural Information Processing Systems*, vol. 36, 37334–37353 (Curran Associates, Inc., 2023).
10. Batzner, S. *et al.* E(3)-equivariant graph neural networks for data-efficient and accurate interatomic potentials. *Nat. Commun.* **13**, 2453, https://doi.org/10.1038/s41467-022-29939-5 (2022).
11. Batatia, I., Kovács, D. P., Simm, G. N. C., Ortner, C. & Csányi, G. Mace: higher order equivariant message passing neural networks for fast and accurate force fields. In Koyejo, S. *et al.* (eds.) *Advances in Neural Information Processing Systems*, vol. 35, 11423–11436 (Curran Associates Inc., 2022).
12. Devereux, C. *et al.* Extending the Applicability of the ANI Deep Learning Molecular Potential to Sulfur and Halogens. *J. Chem. Theory Comput.* **16**, 4192–4202, https://doi.org/10.1021/acs.jctc.0c00121 (2020).

13. Zheng, P., Zubatyuk, R., Wu, W., Isayev, O. & Dral, P. O. Artificial intelligence-enhanced quantum chemical method with broad applicability. *Nat. Commun.* **12**, 7022, https://doi.org/10.1038/s41467-021-27340-2 (2021).
14. Zeng, J., Tao, Y., Giese, T. J. & York, D. M. QDπ: A Quantum Deep Potential Interaction Model for Drug Discovery. *J. Chem. Theory Comput.* **19**, 1261–1275 (2023).
15. Zhang, D. *et al.* DPA-2: a large atomic model as a multi-task learner. *npj Comput. Mater* **10**, 293, https://doi.org/10.1038/s41524-024-01493-2 (2024).
16. Batatia, I. *et al.* A foundation model for atomistic materials chemistry. *arXiv* 2401.00096 https://doi.org/10.48550/arXiv.2401.00096 (2024).
17. Ruddigkeit, L., van Deursen, R., Blum, L. C. & Reymond, J.-L. Enumeration of 166 Billion Organic Small Molecules in the Chemical Universe Database GDB-17. *J. Chem. Inf. Model.* **52**, 2864–2875 (2012).
18. Axelrod, S. & Gómez-Bombarelli, R. GEOM, energy-annotated molecular conformations for property prediction and molecular generation. *Sci. Data* **9**, 185, https://doi.org/10.1038/s41597-022-01288-4 (2022).
19. Goerigk, L., Kruse, H. & Grimme, S. Benchmarking Density Functional Methods against the S66 and S66 × 8 Datasets for Non-Covalent Interactions. *Chem. Phys. Chem.* **12**, 3421–3433, https://doi.org/10.1002/cphc.201100826 (2011).
20. Brauer, B., Kesharwani, M. K., Kozuch, S. & Martin, J. M. L. The S66 × 8 benchmark for noncovalent interactions revisited: explicitly correlated ab initio methods and density functional theory. *Phys. Chem. Chem. Phys.* **18**, 20905–20925, https://doi.org/10.1039/C6CP00688D (2016).
21. Řezáč, J. Non-Covalent Interactions Atlas Benchmark Data Sets: Hydrogen Bonding. *J. Chem. Theory Comput.* **16**, 2355–2368, https://doi.org/10.1021/acs.jctc.9b01265 (2020).
22. Řezáč, J. Non-Covalent Interactions Atlas Benchmark Data Sets 2: Hydrogen Bonding in an Extended Chemical Space. *J. Chem. Theory Comput.* **16**, 6305–6316 (2020).
23. Goerigk, L. *et al.* A look at the density functional theory zoo with the advanced GMTKN55 database for general main group thermochemistry, kinetics and noncovalent interactions. *Phys. Chem. Chem. Phys.* **19**, 32184–32215 (2017).
24. Isert, C., Atz, K., Jiménez-Luna, J. & Schneider, G. QMugs, quantum mechanical properties of drug-like molecules. *Sci. Data* **9**, 273, https://doi.org/10.1038/s41597-022-01390-7 (2022).
25. Madushanka, A., Moura Jr, R. T. & Kraka, E. QM40, Realistic Quantum Mechanical Dataset for Machine Learning in Molecular Science. *Sci. Data* **11**, 1376 (2024).
26. Blum, L. C. & Reymond, J.-L. 970 Million Druglike Small Molecules for Virtual Screening in the Chemical Universe Database GDB-13. *J. Am. Chem. Soc.* **131**, 8732–8733 (2009).
27. Rupp, M., Tkatchenko, A., Müller, K.-R. & von Lilienfeld, O. Fast and Accurate Modeling of Molecular Atomization Energies with Machine Learning. *Phys. Rev. Lett.* **108**, 058301 (2012).
28. Ramakrishnan, R., Dral, P. O., Rupp, M. & von Lilienfeld, O. A. Quantum chemistry structures and properties of 134 kilo molecules. *Sci. Data* **1**, 140022, https://doi.org/10.1038/sdata.2014.22 (2014).
29. Chmiela, S. *et al.* Machine learning of accurate energy-conserving molecular force fields. *Sci. Adv.* **3**, 1603015, https://doi.org/10.1126/sciadv.1603015 (2017).
30. Schütt, K. T., Arbabzadah, F., Chmiela, S., Müller, K. R. & Tkatchenko, A. Quantum-chemical insights from deep tensor neural networks. *Nat. Commun.* **8**, 13890, https://doi.org/10.1038/ncomms13890 (2017).
31. 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).
32. Christensen, A. S. *et al.* OrbNet Denali: A machine learning potential for biological and organic chemistry with semi-empirical cost and DFT accuracy. *J. Chem. Phys.* **155**, 204103, https://doi.org/10.1063/5.0061990 (2021).
33. Hoja, J. *et al.* QM7-X, a comprehensive dataset of quantum-mechanical properties spanning the chemical space of small organic molecules. *Sci. Data* **8**, 43 (2021).
34. Zubatyuk, R., Smith, J. S., Nebgen, B. T., Tretiak, S. & Isayev, O. Teaching a neural network to attach and detach electrons from molecules. *Nat. Commun.* **12**, 4870, https://doi.org/10.1038/s41467-021-24904-0 (2021).
35. Eastman, P. *et al.* SPICE, A Dataset of Drug-like Molecules and Peptides for Training Machine Learning Potentials. *Sci. Data* **10**, 11, https://doi.org/10.1038/s41597-022-01882-6 (2023).
36. 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**, 241733–241743 (2018).
37. Zhang, L., Lin, D.-Y., Wang, H., Car, R. & E, W. Active learning of uniformly accurate interatomic potentials for materials simulation. *Phys. Rev. Materials* **3**, 23804, https://doi.org/10.1103/PhysRevMaterials.3.023804 (2019).
38. Smith, J. S. *et al.* The ANI-1ccx and ANI-1x data sets, coupled-cluster and density functional theory properties for molecules. *Sci Data* **7**, 134 (2020).
39. Yang, M. *et al.* Ab initio accuracy neural network potential for drug-like molecules. *ChemRxiv* https://doi.org/10.26434/chemrxiv-2024-sq8nh (2024).
40. Christensen, A. S. & von Lilienfeld, O. A. On the role of gradients for machine learning of molecular energies and forces. *Mach, Learn.,: Sci, Technol,* **1**, 45018, https://doi.org/10.1088/2632-2153/abba6f (2020).
41. Chai, J.-D. & Head-Gordon, M. Systematic optimization of long-range corrected hybrid density functionals. *J. Chem. Phys.* **128**, 084106, https://doi.org/10.1063/1.2834918 (2008).
42. Giese, T. J. *et al.* Software Infrastructure for Next-Generation QM/MM-ΔMLP Force Fields. *J. Phys. Chem. B* **128**, 6257–6271, https://doi.org/10.1021/acs.jpcb.4c01466 (2024).
43. Mobley, D. L. & Guthrie, J. P. FreeSolv: a database of experimental and calculated hydration free energies, with input files. *J. Comput. Aid. Mol. Des.* **28**, 711–720, https://doi.org/10.1007/s10822-014-9747-x (2014).
44. Mardirossian, N. & Head-Gordon, M. ωB97M-V: A combinatorially optimized, range-separated hybrid, meta-GGA density functional with VV10 nonlocal correlation. *J. Chem. Phys.* **144**, 214110, https://doi.org/10.1063/1.4952647 (2016).
45. Najibi, A. & Goerigk, L. The Nonlocal Kernel in van der Waals Density Functionals as an Additive Correction: An Extensive Analysis with Special Emphasis on the B97M-V and ωB97M-V Approaches. *J. Chem. Theory Comput.* **14**, 5725–5738, https://doi.org/10.1021/acs.jctc.8b00842 (2018).
46. Rappoport, D. & Furche, F. Property-optimized gaussian basis sets for molecular response calculations. *J. Chem. Phys.* **133**, 134105, https://doi.org/10.1063/1.3484283 (2010).
47. Weigend, F. & Ahlrichs, R. Balanced basis sets of split valence, triple zeta valence and quadruple zeta valence quality for H to Rn: Design and assessment of accuracy. *Phys. Chem. Chem. Phys.* **7**, 3297–305, https://doi.org/10.1039/b508541a (2005).
48. Smith, D. G. A. *et al.* PSI4 1.4: Open-source software for high-throughput quantum chemistry. *J. Chem. Phys.* **152**, 184108, https://doi.org/10.1063/5.0006002 (2020).
49. Lehtola, S., Steigemann, C., Oliveira, M. J. & Marques, M. A. Recent developments in libxc — A comprehensive library of functionals for density functional theory. *Softwarex* **7**, 1–5, https://doi.org/10.1016/j.softx.2017.11.002 (2018).
50. Zhang, Y. *et al.* DP-GEN: A concurrent learning platform for the generation of reliable deep learning based potential energy models. *Comput. Phys. Commun.* **253**, 107206, https://doi.org/10.1016/j.cpc.2020.107206 (2020).
51. Gaus, M., Cui, Q. & Elstner, M. DFTB3: Extension of the seld-consistent-charge density-functional tight-binding method (SCC-DFTB). *J. Chem. Theory Comput.* **7**, 931–948 (2011).
52. Gaus, M., Goez, A. & Elstner, M. Parametrization and Benchmark of DFTB3 for Organic Molecules. *J. Chem. Theory Comput.* **9**, 338–354 (2013).

53. Zhang, L. *et al*. End-to-end symmetry preserving inter-atomic potential energy model for finite and extended systems. Bengio, S. *et al*. (eds.) Advances in Neural Information Processing Systems 31, 4436–4446 (Curran Associates, Inc., 2018).
54. Wang, H., Zhang, L., Han, J. & E, W. DeePMD-kit: A deep learning package for many-body potential energy representation and molecular dynamics. *Comput. Phys. Commun.* **228**, 178–184, https://doi.org/10.1016/j.cpc.2018.03.016 (2018).
55. Liang, W., Zeng, J., York, D. M., Zhang, L. & Wang, H. Learning deepmd-kit: A guide to building deep potential models. Wang, Y. & Zhou, R. (eds.) *A Practical Guide to Recent Advances in Multiscale Modeling and Simulation of Biomolecules*, chap. 6, 1–20 https://doi.org/10.1063/9780735425279_006 (AIP Publishing, 2023).
56. Zeng, J. *et al*. DeePMD-kit v2: A software package for deep potential models. *J. Chem. Phys.* **159**, 054801, https://doi.org/10.1063/5.0155600 (2023).
57. Donchev, A. G. *et al*. Quantum chemical benchmark databases of gold-standard dimer interaction energies. *Sci. Data* **8**, 55 (2021).
58. Eastman, P., Pritchard, B. P., Chodera, J. D. & Markland, T. E. Nutmeg and spice: Models and data for biomolecular machine learning 2406.13112 (2024).
59. O'Boyle, N. M. *et al*. Open Babel: An open chemical toolbox. *J Cheminform* **3**, 33 (2011).
60. Izadi, S., Anandakrishnan, R. & Onufriev, A. V. Building Water Models: A Different Approach. *J. Phys. Chem. Lett.* **5**, 3863–3871, https://doi.org/10.1021/jz501780a (2014).
61. Case, D. A. *et al*. AmberTools. *J. Chem. Inf. Model.* **63**, 6183–6191 (2023).
62. Tao, Y. *et al*. Amber free energy tools: Interoperable software for free energy simulations using generalized quantum mechanical/molecular mechanical and machine learning potentials. *J. Chem. Phys.* **160**, 224104 (2024).
63. Hourahine, B. *et al*. DFTB+, a software package for efficient approximate density functional theory based atomistic simulations. *J. Chem. Phys.* **152**, 124101 (2020).
64. Bannwarth, C. *et al*. Extended tight-binding quantum chemistry methods. *WIREs Comput. Mol. Sci.* **11**, e01493 (2020).
65. Eberlein, L. *et al*. Tautomeric Equilibria of Nucleobases in the Hachimoji Expanded Genetic Alphabet. *J. Chem. Theory Comput.* **16**, 2766–2777 (2020).
66. Biondi, E. & Benner, S. A. Artificially Expanded Genetic Information Systems for New Aptamer Technologies. *Biomedicines* **6**, 53 (2018).
67. Wahl, O. & Sander, T. Tautobase: An Open Tautomer Database. *J. Chem. Inf. Model.* **60**, 1085–1089, https://doi.org/10.1021/acs.jcim.0c00035 (2020).
68. Wieder, M., Fass, J. & Chodera, J. D. Fitting quantum machine learning potentials to experimental free energy data: predicting tautomer ratios in solution. *Chem. Sci.* **12**, 11364–11381 (2021).
69. Moser, A., Range, K. & York, D. M. Accurate Proton Affinity and Gas-Phase Basicity Values for Molecules Important in Biocatalysis. *J. Phys. Chem. B* **114**, 13911–13921, https://doi.org/10.1021/jp107450n (2010).
70. Ree, N., Göller, A. H. & Jensen, J. H. RegioSQM20: improved prediction of the regioselectivity of electrophilic aromatic substitutions. *J. Cheminform.* **13**, 10, https://doi.org/10.1186/s13321-021-00490-7 (2021).
71. Fink, T., Bruggesser, H. & Reymond, J.-L. Virtual Exploration of the Small-Molecule Chemical Universe below 160 Daltons. *Angew. Chem. Int. Ed.* **44**, 1504–1508, https://doi.org/10.1002/anie.200462457 (2005).
72. Law, V. *et al*. DrugBank 4.0: shedding new light on drug metabolism. *Nucleic Acids Res.* **42**, 1091–1097, https://doi.org/10.1093/nar/gkt1068 (2014).
73. Zeng, J., Giese, T., Goetz, A. & York, D. The QDπ dataset, training data for drug-like molecules and biopolymer fragments and their interactions https://doi.org/10.5281/zenodo.14970869 (2025).
74. van der Maaten, L. & Hinton, G. Visualizing data using t-sne. *Journal of Machine Learning Research* **9**, 2579–2605 (2008).
75. dpdata: a python package for manipulating atomistic data. https://github.com/deepmodeling/dpdata. Accessed: 2024-12-17.

Acknowledgements

The authors are grateful for financial support provided by the National Institutes of Health (No. GM107485 to D.M.Y.) and the National Science Foundation (CSSI Frameworks Grant No. 2209718 to D.M.Y. and 2209717 to A.G.). Computational resources were provided by the Office of Advanced Research Computing (OARC) at Rutgers, The State University of New Jersey; the Advanced Cyberinfrastructure Coordination Ecosystem: Services & Support (ACCESS) program, which is supported by National Science Foundation grants #2138259, #2138286, #2138307, #2137603, and #2138296 (supercomputer Expanse at SDSC through allocation CHE190067); and supercomputer Aerosol at SDSC.

Author contributions

J.Z. conceived the experiments, J.Z. conducted the experiments, J.Z. analyzed the results. All authors reviewed the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to D.M.Y.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by-nc-nd/4.0/.

© The Author(s) 2025