

TorsionNet: A Deep Neural Network to Rapidly Predict Small Molecule Torsion Energy Profiles with the Accuracy of Quantum Mechanics

*Brajesh K. Rai^{*1}, Vishnu Sresht¹, Qingyi Yang², Ray Unwalla², Meihua Tu², Alan M. Mathiowetz², and Gregory A. Bakken³*

¹Simulation and Modeling Sciences and ²Medicine Design, Pfizer Worldwide Research Development and Medical, 610 Main Street, Cambridge, Massachusetts 02139, United States

³Digital, Pfizer, Eastern Point Road, Groton, Connecticut 06340, United States

KEYWORDS: Neural network, deep learning, machine learning, ML, quantum mechanics, QM, torsion, strain energy, CSD, PDB, conformational strain, cloud computing

ABSTRACT

Fast and accurate assessment of small molecule dihedral energetics is crucial for molecular design and optimization in medicinal chemistry. Yet, accurate prediction of torsion energy profiles remains a challenging task as current molecular mechanics methods are limited by insufficient coverage of druglike chemical space and accurate quantum mechanical (QM) methods are too expensive. To address this limitation, we introduce TorsionNet, a deep neural network (DNN) model specifically developed to predict small molecule torsion energy profiles with QM-level accuracy. We applied active learning to identify nearly 50k fragments (with elements H, C, N, O, F, S, and Cl) that maximized the coverage of our corporate library and leveraged massively parallel cloud computing resources to perform DFT torsion scan of these

fragments, generating a training dataset of 1.2 million DFT energies. By training TorsionNet on this dataset, we obtain a model that can rapidly predict the torsion energy profile of typical druglike fragments with DFT-level accuracy. Importantly, our method also provides a direct estimate of the uncertainty in the predicted profiles without any additional calculations. In this report, we show that TorsionNet can reliably identify the preferred dihedral geometries observed in crystal structures. We also present practical applications of TorsionNet that demonstrate how consideration of DNN-based strain energy leads to substantial improvement in existing lead discovery and design workflows. A benchmark dataset (TorsionNet500) comprising 500 chemically diverse fragments with DFT torsion profiles (12k DFT-optimized geometries and energies) has been created and is made freely available.

INTRODUCTION

Small molecule conformations in solution, solid state, and protein environment are determined primarily by the rotational energetics of constituent single bonds. Therefore, accurate determination of small molecule torsion energy profiles is a crucial step towards characterizing their physical and chemical properties and rationalizing their biological activities. The ability to rapidly and accurately predict preferred dihedral angles and estimate the energy penalty due to the deviation from the low-energy torsion angle is essential to the successful application of molecular modeling in small molecule discovery and design, including tasks such as generation of low-energy conformations¹, virtual screening^{2, 3}, and interactive exploration of structure-based design ideas for potency optimization.

Both data-driven and physics-based techniques have been applied to assess torsional preferences and to identify low-energy rotamers. Knowledge derived from crystal structure databases, e.g. CSD⁴ and PDB⁵, serves as a useful guide for structure-based design⁶⁻⁸ and is also used by many conformation generation programs⁹⁻¹³. By leveraging the statistics of the observed torsion angle ranges in structural databases, these programs generate physically and biologically relevant structures. Nevertheless, such data-driven analyses are often hindered by the underrepresentation of novel chemistries explored by medicinal chemistry programs.

Physics-based alternatives, on the other hand, usually rely on approximate molecular mechanics (MM) force fields¹⁴⁻¹⁸, expressing the total energy of a conformer using a combination of bonded and non-bonded terms, where each individual component is described by a fixed functional form and a pre-specified set of parameters. Dihedral terms are modeled as a Fourier series and the associated parameters are derived by fitting to torsion scan profiles of model systems obtained from quantum mechanics (QM) calculations. Recent force field development efforts have focused on improving the coverage of torsion parameters to ever-expanding small molecule chemical space, leading to a nearly exponential increase in the number of dihedral types in some force fields. For example, the number of torsion parameters in successive releases of OPLS force field has gone from 1.6k in the version OPLS_2005¹⁹ to 45k in OPLS2.1²⁰ to 147k in the latest version, OPLS3e¹⁸. Despite these notable efforts, molecular mechanics continues to be limited in accuracy due to its restrictive functional form and incomplete parameterization and performs poorly when applied to unseen chemical motifs such as those found in proprietary pharmaceutical libraries.

Ab initio methods do not have such limitations and have been shown to reproduce preferred dihedral geometries in a variety of organic molecules and fragments²¹⁻²⁵. In a recent publication, we reported our extensive analysis of torsional strain in crystal structures with Density Functional Theory (DFT)²³. Key insights from that exercise were that in the crystalline state molecules generally adopt low-energy conformations, consistent with gas-phase DFT calculations from molecular substructures, and that motifs with deviations from ideal gas-phase geometry are normally associated with a high-energy penalty and occur relatively infrequently, which in a great majority of cases can be attributed to external effects such as steric hindrance or inter-molecular interactions. We also demonstrated that reasonable estimates of conformational strain can be obtained by summing up the strain energies in each individual rotatable bond in the molecule. The strain energy of each rotatable bond is calculated independently using an *ab initio* torsion scan performed on an essential substructure fragment around that rotatable bond. However, despite the relative efficiency of the fragment-based approach, the high computational cost associated with QM calculations severely constrains the throughput of the method, making it impractical for use in virtual screening or interactive structure-based design.

We address this practical limitation of first principle methods by using machine learning (ML), which allows us to circumvent the computationally expensive QM calculations and provide several orders of magnitude speed-up while retaining the accuracy of *ab initio* calculations. In recent years, ML models, trained on reference QM data, have been applied to a variety of chemistry tasks^{26, 27}. One of the earliest examples was the use of a Kernel Ridge Regression model²⁸ that was trained on reference data from 6k randomly selected compounds from GDB7²⁹—a database of organic compounds with up to 7 heavy atoms—to predict atomization energies. In a subsequent publication, ML was applied to a problem of greater pharmaceutical relevance: Random Forest models were trained to produce *ab initio* quality atomic charges using a data set of QM ESP-fit charges from 80k druglike molecules³⁰. More recently, ML has been applied to predict various atomic and molecular properties³¹⁻³⁵ and to develop general-purpose molecular potentials³⁶⁻⁴⁰. Despite the reported rapid advancement in methods for training and validating ML models, most of the published QM/ML models in chemistry thus far have been limited to modeling the *ab initio* data generated from GDB compounds that are comprised of a small number of heavy atoms (typically 7 – 13 heavy atoms) and generally lack the level of structural and chemical complexity found in druglike molecules. Furthermore, relatively little attention has been given to test these models using experimental data or to assess their utility in practical medicinal chemistry workflows.

In this work, we focus on improving the accuracy of dihedral preference characterization and torsional strain estimation in druglike molecules by using a QM/ML model. We introduce TorsionNet, a deep neural network (DNN) model, which can rapidly generate a DFT-quality torsion energy profile of an arbitrary druglike fragment. We describe our active learning methodology, which in conjunction with massively parallel cloud computing resources, enabled extensive exploration of the chemical space represented in our corporate library, and allowed us to generate DFT torsion scan profiles of nearly 50k small molecule fragments for training and validation of our ML model. We assess the performance of TorsionNet by comparing DNN predictions with reference DFT calculations and extensively evaluate the ability of our model to identify preferred torsion angles in crystal structures. Application of DNN-based strain energies for practical computational drug design tasks is demonstrated with three proof-of-concept applications: conformation analysis, virtual screening, and crystal structure refinement.

MATERIALS AND METHODS

Minimal Torsion Fragments

The minimal torsion fragment, as defined and described in our previous work²³, is the essential substructure that defines the chemical environment around a specified rotatable bond that significantly influences its torsional energy profile. This fragmentation is both necessary and sufficient to accurately describe torsional energy profiles. As described in subsequent sections, QM and DNN profiles are generated by scanning the specified dihedral in such minimal torsion fragments extracted from the parent molecule. We use minimal torsion fragments instead of the parent molecules themselves because scanning the dihedral in the parent molecule could be both computationally challenging and generate misleading torsion energy profiles due to long-range interactions.

Generation of Torsion Profiles

QM torsion energy profiles were generated by following the procedure described in our previous work²³, and is summarized here for completeness. The workflow is comprised of three stages. First, Omega¹⁰ was used to generate up to 20 low-energy conformations of the given fragment. However, in the absence of secondary rotatable bonds in many fragments, fewer than 20 conformers could be generated. In the second stage, with each Omega conformation as the starting point, gas-phase molecular mechanics scans were performed with MMFF94s⁴¹, using the OETorsionScan function from the OESzybki toolkit (release 2.1.0, OpenEye Scientific Software, Santa Fe, NM). The output from multiple scans was combined by taking the lowest energy structure at each intermediate step, yielding a single MM torsion energy profile. Finally, in the third stage, restrained gas-phase geometry optimization of each MM conformation was carried out using QM, first with HF/3-21G, followed by B3LYP/6-31G* (or B3LYP/6-31+G* for fragments that contain Sulfur or anions). The single-point energy of each optimized structure was calculated in the gas phase using B3LYP/6-31G** (or B3LYP/6-31+G** for fragments that

contain Sulfur or anions), producing a torsion profile that is comprised of 24 equally spaced points (separated by 15°) over the entire -180° to $+180^\circ$ range.

All QM calculations were performed using the Psi4 package⁴². The workflow described above was implemented using Python and was run on the AWS (Amazon Web Services) using the Orion framework (OpenEye Scientific Software, Santa Fe, NM).

Chemical Environment Representation using Symmetry Function

The local atomic environment around a given rotatable bond is represented by a fixed-length numerical vector, using a set of symmetry functions $\{G_r^\alpha, G_\theta^{\alpha\beta}, G_\phi^{\alpha\beta}\}$ that are invariant not only to translation and rotation of the whole molecule but also to the permutation of atoms that are of the same type. Symmetry functions G_r^α and $G_\theta^{\alpha\beta}$ were first proposed by Behler and Parrinello⁴³ to describe the radial and angular distributions, respectively, of the atoms in the neighborhood of a reference center. In this work, we extended the original Behler-Parrinello method by introducing torsion symmetry functions ($G_\phi^{\alpha\beta}$), which provides a more detailed description of the local dihedral environment that determines the torsion energy profile. Here, α and β represent one of the 9 atom types, one for each of the chemical elements H, C, N, O, F, S, and Cl, and 2 additional types for atoms with negative and positive formal charge.

A common component of all three symmetry functions is the cutoff function f_c , which defines the chemical environment around a given reference center

$$f_c(R_{ij}^\alpha) = \begin{cases} 0.5 \cdot \left[\cos\left(\frac{\pi R_{ij}^\alpha}{R_c}\right) + 1 \right] & \text{for } R_{ij}^\alpha \leq R_c \\ 0.0 & \text{for } R_{ij}^\alpha > R_c \end{cases} \quad (1)$$

where R_{ij}^α is the distance between the reference center i and the neighboring atom j of type α within the cutoff radius R_c .

Radial symmetry functions use a Gaussian-weighted sum of cutoff functions, f_c , to describe radial arrangement of neighboring atoms of type α around the reference center,

$$G_r^\alpha = Z_b Z_c \sum_{j=1}^N e^{-\eta(R_{ij}^\alpha - R_s)^2} \cdot f_c(R_{ij}^\alpha). \quad (2)$$

Here, Z_b and Z_c represent the atomic number of the two atoms that define the specified rotatable bond and the reference center i was chosen as the mid-point of that bond. Parameters η and R_s , which represent the width and shift of the Gaussians were set to 10^{-4} and 0, respectively. The radial distribution of each atom of type α around the reference center i was generated using seven different cutoff radii ($R_c = 1.5, 2.0, 2.5, 3.0, 4.0, 6.0$, and 10.0 Å), yielding 63 radial symmetry function elements (7 cutoff radii for each of the 9 atom types).

Angular symmetry functions depend on the angle θ_{ijk} , centered at i , and enclosed by vectors R_{ij}^α and R_{ik}^β that connect atoms j and k of type α and β , respectively, to the reference center. The sum over all cosines with respect to any neighbor j and k , each of which is weighted by a Gaussian of the three interatomic distances and the respective cutoff functions were calculated as

$$G_\theta^{\alpha\beta} = 2^{1-\zeta} Z_b Z_c \sum_j \sum_{k \neq j} \left[(1 + \lambda \cdot \cos \theta_{ijk})^\zeta \cdot e^{-\eta(R_{ij}^2 + R_{ik}^2 + R_{jk}^2)} \cdot f_c(R_{ij}^\alpha) \cdot f_c(R_{ik}^\beta) \cdot f_c(R_{jk}^{\alpha\beta}) \right]. \quad (3)$$

As in the case of the radial symmetry functions, Z_b and Z_c represent the atomic number of the two atoms that define the specified rotatable bond and the reference center i was chosen as the mid-point of that bond. Parameters ζ , λ , and η were set to 0.5, 0.5, and 10^{-4} , respectively. In order to keep the number of angular symmetry function elements small, only a single cutoff radius, $R_c = 4.0$ Å, was used. With nine atom types and one fixed set of parameters, our angular symmetry functions were described by 45 elements (number of unique atom-type pairs).

Torsion symmetry functions were defined by dihedral angles φ_{ijkl} that includes atoms j and k of the specified rotatable bond. Pairs i and l are any atoms of type α and β , respectively, but located on opposite side of the rotatable bond (jk).

$$G_\varphi^{\alpha\beta} = 2^{1-\zeta} Z_b Z_c \sum_j \sum_{k \neq j} \left[(1 + \lambda \cdot \cos \varphi_{ijkl})^\zeta \cdot e^{-\eta(R_{ij}^2 + R_{kl}^2 + R_{il}^2)} \cdot f_c(R_{ij}^\alpha) \cdot f_c(R_{kl}^\beta) \cdot f_c(R_{il}^{\alpha\beta}) \right]. \quad (4)$$

All parameters of the torsion symmetry function were identical to those we used for the angular component. However, to construct a more detailed representation of the dihedral environment, torsion symmetry functions were calculated using a range of cutoff radii ($R_c = 2.5, 3.5, 5.0$, and 10.0 \AA), generating 180 values for this component.

In addition to symmetry functions, we used five additional features that describe the geometric and chemical characteristics of the specified torsion. These include the dihedral angle, the length of the central rotatable bond, the distance between the two terminal atoms, and the product of the atomic numbers of the two central and the two terminal atoms. These additional features were combined with the radial, angular, and torsion symmetry function elements to produce a 293-dimensional atomic environment vector (AEV) that encodes the chemical environment around a given rotatable bond.

The key features of these symmetry functions include their smoothness to minor perturbation in geometries and their invariance to translation, rotation, and permutation, making these functions important and well-received descriptors for machine learning methods⁴⁴.

Training Machine Learning Models

Machine learning regression models were trained to map a given torsion motif into a numerical value, corresponding to the relative energy of the specified torsion angle on the DFT profile (Fig. 1a). Model training and validation were done on a dataset $D = \{(X_1, y_1), \dots, (X_{24}, y_{24}), \dots, (X_{n \times 24}, y_{n \times 24})\}$, comprising 1.2 million data points that were obtained from DFT scans of n (50k) small molecule fragments, where X_i is a 293-dimensional AEV and y_i is the corresponding relative energy obtained from the reference DFT profile. It is important to emphasize that while y_i was the single point energy based on a DFT-optimized structure, the corresponding X_i vector was calculated using the intermediate conformation generated by MM scan of the given fragment, the output from the stage 2 of our Generation of Torsion Profiles workflow described above.

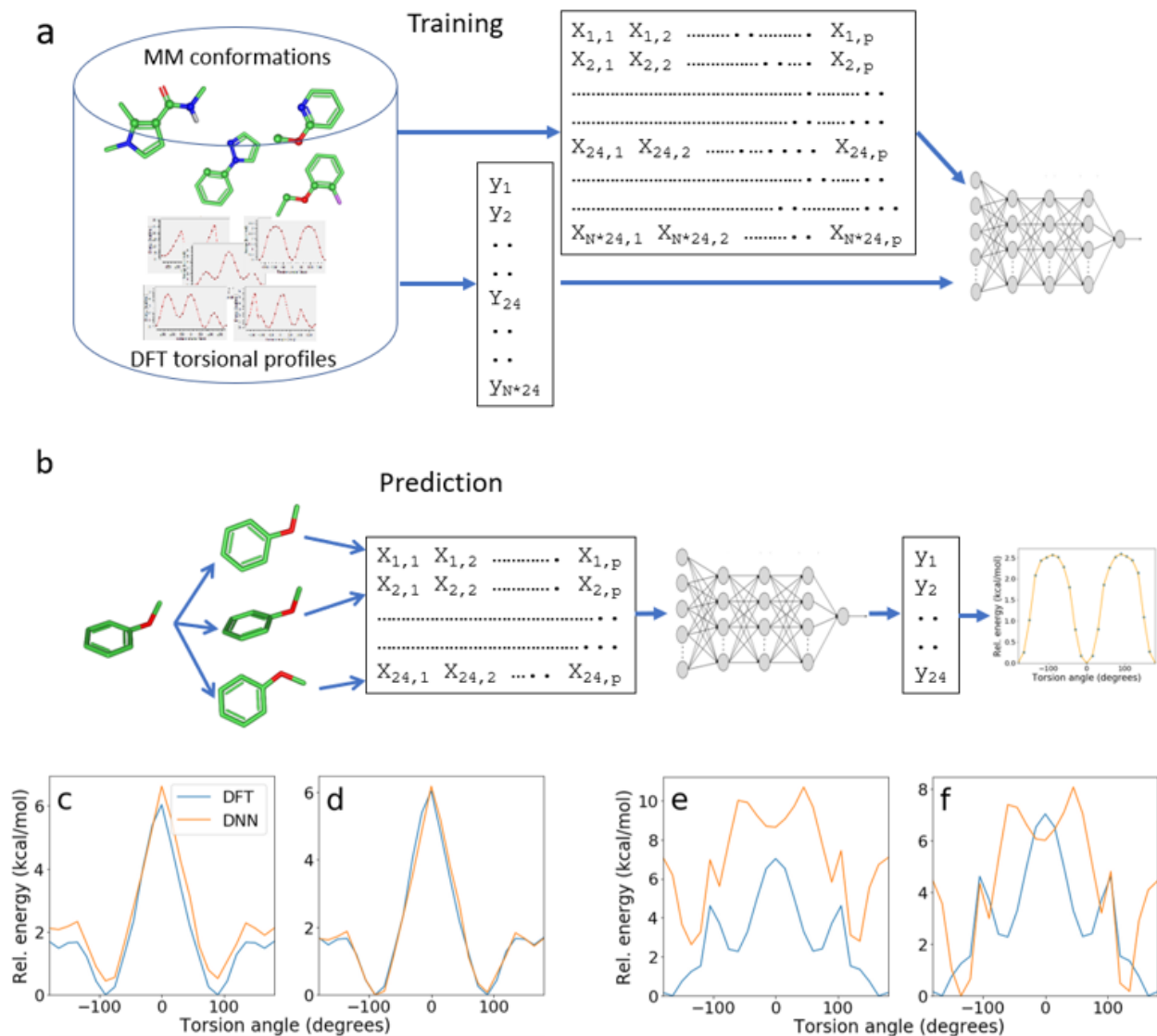


Figure 1. Illustration of training and prediction procedures. (a) Using a 3D MM conformation of the training set fragments, a machine learning model is trained to fit the corresponding DFT relative energy. Symmetry functions, which represent the local chemical environment around the specified rotatable bond, are provided as input features to the model. (b) A minimal torsion fragment around the specified torsion is generated, followed by an MM torsion scan to generate intermediate torsion conformers. Using symmetry functions as input, the ML model predicts the relative energy at each intermediate torsion angle, generating a torsion energy profile. (c-f) DFT and DNN torsion profiles of two motifs, selected to illustrate the effect of offset correction. (c and e) Raw DNN and the corresponding DFT profiles. (d and f) Offset-corrected DNN and the reference DFT profiles. While the predicted DNN profile in panel c closely resembles the reference DFT profile, it is shifted up by an offset value of 0.5 kcal/mol. Offset correction shifts the DNN profile down by 0.5 kcal/mol along y-axis, making the two profiles shown in panel d (DNN and DFT) nearly identical. The DNN profile in panel e, however,

displays a larger offset (~ 2.5 kcal/mol). A major discrepancy between DNN and DFT profiles persists even after applying offset correction (panel f).

Random Forest

The TreeBagger function from the MATLAB (Mathworks, Natick, MA) Statistics and Machine Learning toolbox was used to train Random Forest (RF) regression models on the available data (MM structure and DFT relative energies) during active learning, with AEVs as the input features and the relative energies obtained from the corresponding DFT profiles as the target. Each RF model was trained with default parameters, using an ensemble of 100 decision trees. Hyperparameter tuning was not performed as RF performance is known to be robust to a wide range of model parameters.

Deep Neural Network

As in the case of the Random Forest, DNN regression models were also trained using AEVs as the input features and the corresponding relative energies as the target. However, unlike RF, each input feature to the DNN model was first standardized by subtracting the mean and scaling to unit variance. The model architecture was defined using the Keras deep learning framework (version 2.2.0, <http://keras.io>) with the TensorFlow (version 1.7.0, www.tensorflow.org) backend. The network is comprised of 5 fully connected (FC) hidden layers, with 2930 nodes (corresponding to $10p$, where p is the number of input features) in the first hidden layer. Each subsequent hidden layer contained a decreasing number of nodes, $5p$, $3p$, $2p$, $1p$, respectively. The output of each FC layer was transformed by the ReLU activation function, followed by Batch Normalization and a drop out layer with a fixed dropout value of 0.2. The network was trained using 70% of the torsional profile data, with 10% of the DFT data held out for validation; the remaining 20% of the dataset was used as the test set. Weights in each layer were initialized using the default Glorot scheme. The network was optimized for logcosh loss using Adam with a fixed learning rate of 10^{-4} and default settings for other optimizer parameters ($\beta_1=0.99$, $\beta_2=0.999$). Models were trained with a batch size of 256

for 5000 epochs. Since no early stopping criteria were applied, each model was trained for a full 5000 epochs, while saving the best model (with the smallest validation loss).

Hyperparameter tuning was performed by exploring the ranges of values for the learning rate, batch size, and drop out that are commonly used when training deep neural networks. We also tuned the width and the depth of the network, testing architectures that contained between 2 and 6 [FC-ReLU-BN-dropout] blocks. The learning rate (with Adam optimizer) was varied from 1×10^{-3} to 1×10^{-6} with 10-fold increments. The batch size was sampled from one of the following values: 64, 128, 256, 512, 1024, 2048, 4096, and 8192. Drop out values of 0.05, 0.1, 0.2, 0.3, and 0.5 were explored, while keeping the same dropout rate for each layer.

Generating ML Torsion Energy Profiles

The procedure for generating ML torsion energy profiles is illustrated in Fig. 1b. For each test set fragment, the ML profile consists of 24 discrete points, representing the DNN relative energy of rotation intermediates around the specified single bond with uniform 15° separation in the torsion angle. As noted earlier, the input descriptors to the ML model were derived from MM-based conformations, as described above (see Generation of Torsion Profiles section). Next, AEVs, representing the chemical environment around the specified rotatable bond in each intermediate structure, were calculated, yielding a 24-by-293 matrix in which each row corresponds to an intermediate conformation. Finally, this matrix was used as the input to the ML model, which in turn produced a 24-dimensional output where each element corresponds to the predicted relative energy of each of the 24 intermediate dihedral angles between -180° and $+180^\circ$.

Since the relative energy of each intermediate structure is calculated independently, the minimum energy point on the predicted energy-vs-angle plot is nearly always offset away from the x-axis, necessitating an “offset correction” that shifts the entire profile along the y-axis by setting the global profile minimum to zero. This offset correction is illustrated by two examples shown in Fig. 1c-1f, where the global energy minimum of the two raw DNN profiles are shifted by 0.5 kcal/mol (Fig. 1c) and 2.5 kcal/mol (Fig. 1e); the offset-correction shifts these profiles along the y-axis, as shown in Fig. 1d and 1f, respectively.

Predicting Torsional Strain

The torsional strain of a given 3D conformation was determined as follows. First, the given molecule was decomposed into a set of overlapping substructures, generating one minimal torsion fragment corresponding to each constituent rotatable bond. Next, the DNN energy profile of each torsion motif was constructed as described above. The DNN profile was subsequently transformed into a continuous representation by cubic interpolation using `interp1d` function from `scipy`⁴⁵. The strain energy of each rotatable bond was obtained by calling the interpolation function with the specified dihedral angle in the given 3D conformation as input. Finally, the torsional strain of the given 3D conformation was calculated as the sum of strain energies from all N_τ constituent torsions, $E_{strain} = \sum_{i=1}^{N_\tau} E_{\tau, strain}$.

Estimating Prediction Error

We have found that profiles with offset value greater than 1 kcal/mol generally show greater discrepancies relative to the reference DFT data, even after offset correction has been applied (see Supplementary Fig. 2). Therefore, we used the magnitude of offset correction as a measure of prediction error in the ML profiles, treating profiles with offset value greater than 1 kcal/mol as low confidence predictions. This criterion was used during active learning to efficiently sample the chemical space, selecting fragments with poorly represented chemistries for DFT calculation at each subsequent iteration, as described next.

Efficient Chemical Space Coverage with Active Learning

Active learning was applied to select a subset of chemically diverse structures for DFT calculations from a rank-ordered collection of nearly 3 million unique minimal torsion fragments that were extracted from our corporate library (fragments were rank-ordered by their total number of occurrence in our corporate library; highest to lowest). As shown in Fig. 2, DFT torsion scans were performed in batches, starting with a set of 20k fragments with the highest frequency of occurrence, representing fragments with the highest relevance to our programs.

Data generated in this first iteration was used to train a Random Forest model. The choice of Random Forest as the ML algorithm in the active learning stage was motivated primarily by the ease of training as RF is known to produce a high performing model without requiring significant hyperparameter tuning. The RF model was subsequently applied to predict torsion energy profiles of the next 50k fragments in the list; and a subset of fragments with high prediction error, based on the profile offset threshold of 1.0 kcal/mol, in the corresponding RF torsion profiles, were selected for DFT scan in the next iteration. The output from the second iteration was appended to the data from the first iteration, closing the loop. This procedure was repeated eight times, generating nearly 50k DFT torsion energy profiles.

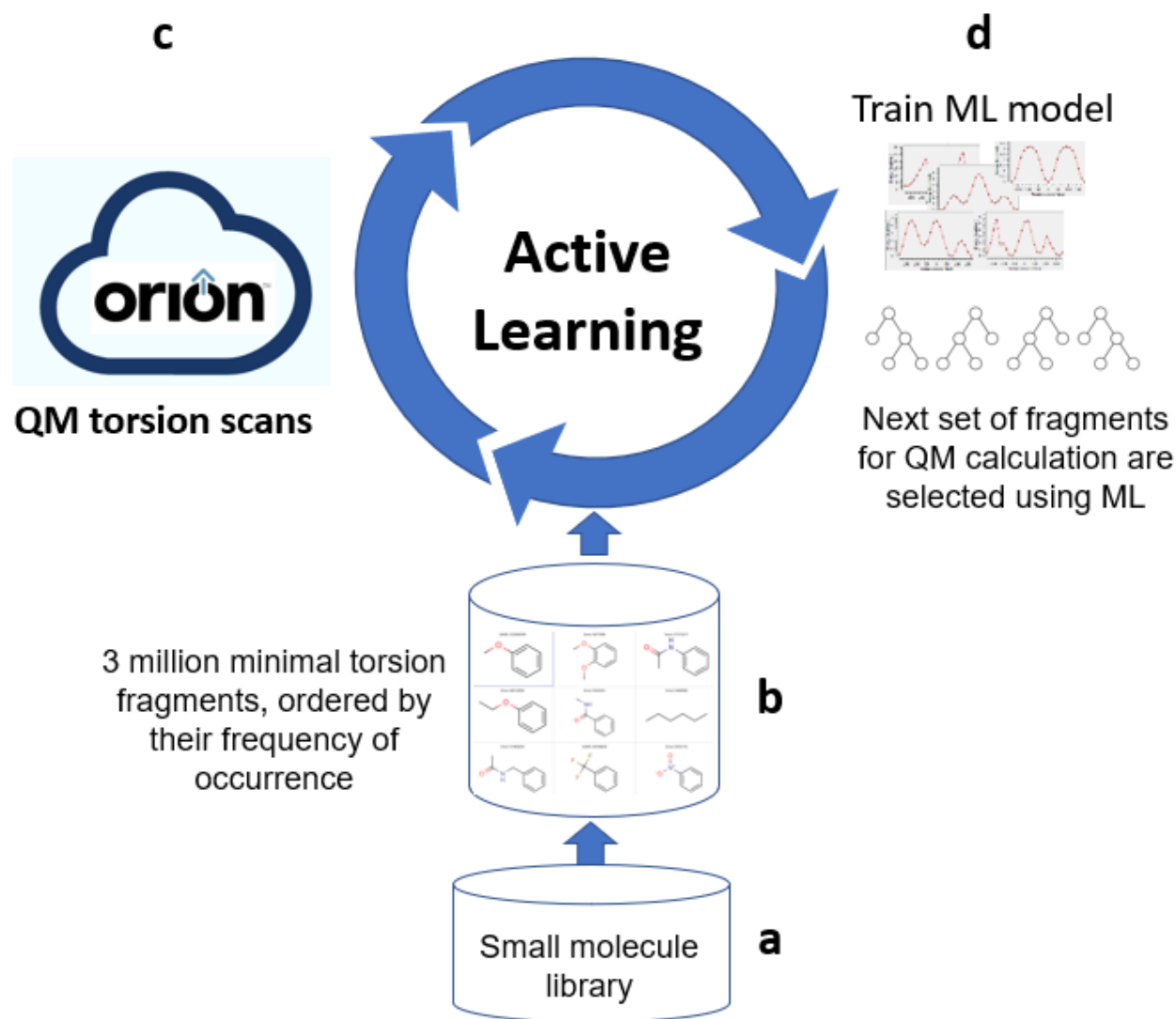


Figure 2. Active learning procedure for generating DFT torsion profiles from chemically diverse fragments. (a) Our in-house small molecule library was used to generate minimal torsion fragments. (b) Fragments were ordered based on their frequency of occurrence in the small molecule library, keeping the fragments with high medicinal chemistry relevance to near the top of the list. (c) In the first iteration of active learning, DFT torsion scans of 20k highest ranking (most relevant) fragments were performed on AWS using the Orion framework. (d) An ML model to predict torsion profiles was trained using available data (DFT relative energies and fragment structures); the model was then used to generate torsion profiles of the next 50k fragments in the collection (b). The fragments with high prediction error based on the profile offset threshold of 1 kcal/mol, representing chemistries not represented in the current training set, were selected for the next iteration of DFT calculations, model training, and selection of fragments with novel chemistries for the subsequent round.

TorsionNet500: A Benchmark Dataset of 500 Diverse Fragments with DFT Torsion Profiles

A diverse collection of 500 fragments and the associated DFT profiles were selected from our dataset of 50k fragments. This dataset comprises of a variety of acyclic, aryl, biaryl, and other motifs and should serve as a useful benchmark to test and validate the current and future QM/ML methods. The DFT energies and optimized geometries of intermediate conformations from this subset are available as part of the Supplementary Information.

RESULTS AND DISCUSSIONS

Employing active learning and large-scale cloud computational resources, we extensively explored the available chemical space of our corporate library, completing nearly 50k DFT torsion scan calculations within approximately three weeks, a task that would have required several years on typical high performance compute clusters that is accessible to most research organizations. Fragments selected for DFT calculations via our active learning approach exhibit broad structural and chemical diversity, ranging from simple acyclic systems to more complex substructures comprised of multiple heterocycles and substituents. The key physicochemical properties of these fragments are presented in Supplementary Fig. 1.

Below, we present the performance of TorsionNet and demonstrate its applications to common drug discovery tasks. First, we compared the accuracy of TorsionNet profiles with respect to the reference DFT results. This is followed by qualitative and quantitative comparison of predicted energy profiles with observed torsion angle distributions, rigorously testing the ability of our model to identify the preferred dihedral angle and geometries in crystal structures. Finally, we demonstrate practical applications of our method in conformational analysis, structure-based virtual screening, and refinement of ligand structures in the PDB.

Comparison with DFT

We compared the TorsionNet energies and profiles to the reference DFT calculations using out-sample (unseen) motifs in a 5-fold cross-validation test set. Fig. 3 (a-d) shows TorsionNet and the corresponding DFT profiles of four motifs that are common in medicinal chemistry. Overall,

TorsionNet profiles of these motifs are in excellent agreement with the corresponding DFT results. The ability of TorsionNet to capture fine details of DFT profiles is particularly notable. For example, even the low-energy barrier at 0^0 in the DFT profile shown in Fig. 3b is correctly reproduced by TorsionNet. More importantly, TorsionNet is able to predict both the location and relative energy value of local minima on DFT profiles with high accuracy, which is critical for the successful downstream utilization of these profiles in other molecular modeling tasks such as conformation generation and virtual screening.

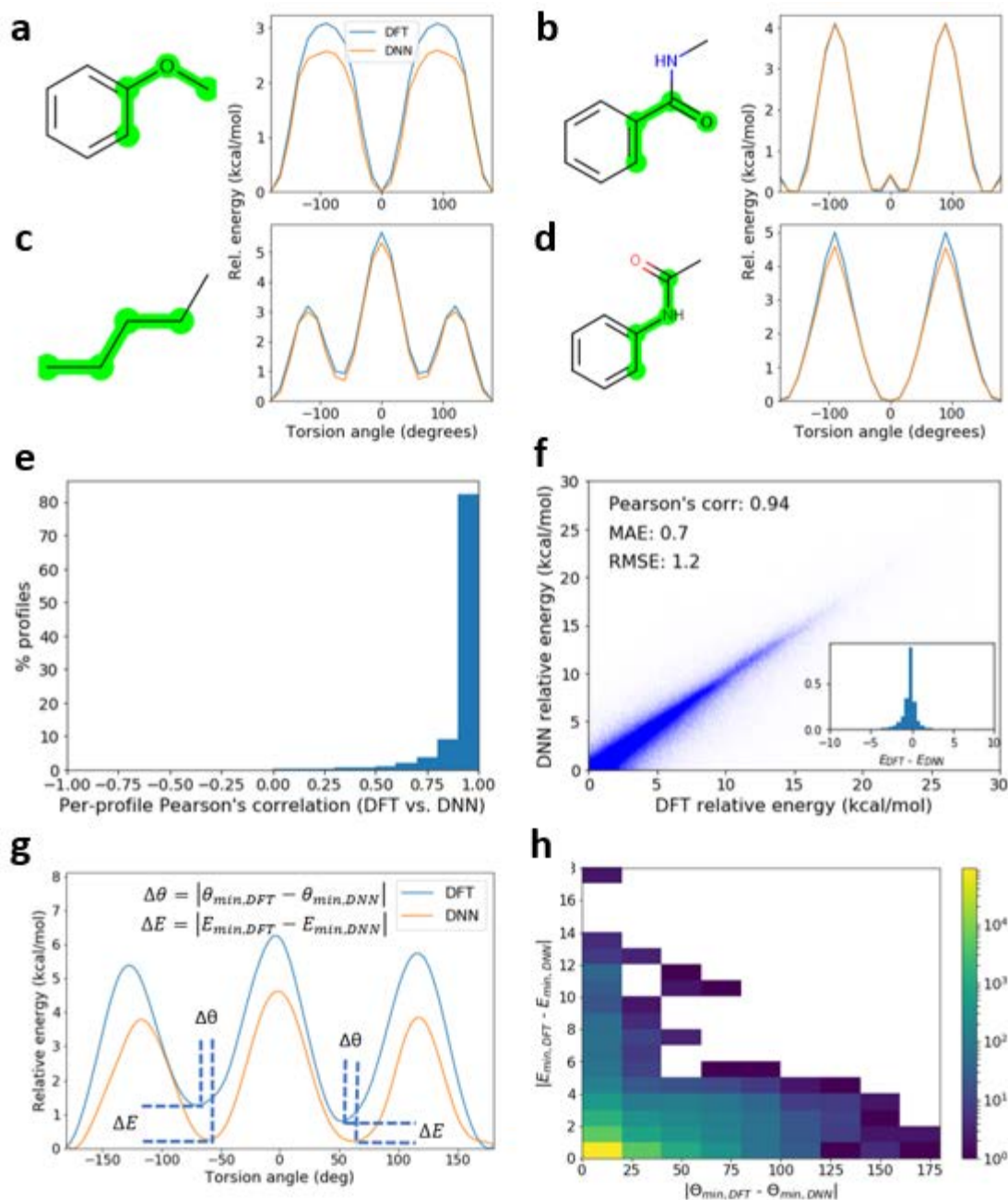


Figure 3. (a-d) Comparison of TorsionNet profiles to the corresponding DFT results for four frequently observed fragments. Each profile was generated by scanning the dihedral angle marked in green within the fragment depicted immediately to the left of the profile. (e) Histogram of Pearson's correlation between DNN and DFT torsion energy profiles in cross-validation test set. (f) Comparison of cross-validated DNN relative energies to the corresponding reference DFT energies from 1.2 million intermediate torsion conformations in 50k fragments. Relative energy of each intermediate conformation was calculated with respect to the lowest energy point on the corresponding torsion profile. Inset (panel f): normalized histogram of discrepancy in DNN relative energies with respect to DFT. (g) Schematics of calculating torsion angle and relative energy differences between corresponding local minima

from DFT and DNN. (h) 2D Histogram of DNN prediction error in torsion angle (x-axis) and relative energy (y-axis) of 106k low-energy minima in the test set profiles with respect to DFT. The histogram is rendered using a logarithmic color scale according to the number of data points in each bin, yellow (largest) to purple (smallest).

We extensively compared TorsionNet and DFT profiles for the entire cross-validation test set. Over 95% of the out-sample test set profiles were predicted with high confidence, i.e., the offset in these DNN profiles, which as we discussed earlier provides a measure of prediction accuracy, was less than 1 kcal/mol. In the following discussions, we present the performance of our model based on this high-confidence prediction subset.

Fig. 3e shows the distribution of Pearson’s correlation for each profile pair (DNN vs. DFT) in the test set. We find a strong correlation between the two sets of profiles, with 82% of the profile pairs exhibiting Pearson’s correlation ≥ 0.9 . Furthermore, as shown in Fig. 3f, the DNN relative energies of 1.2 million intermediate points that make up these profiles are in excellent agreement with DFT, with Pearson’s correlation of 0.94, median absolute error of 0.3 kcal/mol, mean absolute error (MAE) of 0.7 kcal/mol and root mean squared error (RMSE) of 1.2 kcal/mol.

Low-energy minima in *ab initio* torsion profiles have been shown to be a good indicator of the preferred torsion angles of matching substructures in crystal structures²¹⁻²³, making QM, in particular DFT, an ideal, though not necessarily practical, method to assess dihedral preferences. We, therefore, evaluated how well low-energy minima observed in test set DFT profiles could be reproduced by our DNN model through the following procedure. First, we identified 106k low-lying minima (relative energy ≤ 4.2 kcal/mol; this cutoff was chosen as it corresponds to a 3-fold difference in binding affinity) from our test set DFT profiles. Energy minima ($E_{min,DFT}$) were determined by minimizing the interpolation function that approximates a given torsion profile, using multiple initial guesses, each of which corresponds to a local minimum on the discrete torsion profile. Next, as illustrated in Fig. 3g, for each DFT local minimum ($\theta_{min,DFT}$) we determined the nearest minimum on the corresponding TorsionNet profile by minimizing the interpolation function of that profile using $\theta_{min,DFT}$ as the initial guess. Finally, for each pair of local minima identified in the previous step,

the absolute difference between the corresponding torsion angles, $|\theta_{min,DFT} - \theta_{min,DNN}|$, and the relative energy pairs, $|E_{min,DFT} - E_{min,DNN}|$, were calculated. The result, as displayed as a 2D histogram plot in Fig. 3h, show that for an overwhelming majority of the datapoints in our test set, TorsionNet is able to reproduce DFT local minima with a high degree of accuracy: 84% of DNN local minima occur within 20° to the nearest DFT minimum with energy difference less than 1 kcal/mol.

We also evaluated the performance of our model on the TorsionNet500 benchmark. As summarized in Table 1, TorsionNet shows excellent performance in reproducing the reference DFT profiles and energies of the fragments in this set.

Table 1. Performance of TorsionNet and ANI-2x on the TorsionNet500 benchmark dataset. Torsion profiles and relative energies obtained from each method were compared to the reference DFT data.

Method	% of profiles with Pearson’s correlation ≥ 0.9	Pearson’s correlation, Rel. E	MAE (kcal/mol)	RMSE (kcal/mol)	% of local minima with $\Delta\theta \leq 20^\circ$ and $\Delta E \leq 1$ kcal/mol
TorsionNet	79	0.91	0.7	1.3	83
ANI-2x	54	0.75	1.4	2.1	66

Profile Offset as a Measure of Prediction Accuracy

Profile offsets, which were used as part of active learning to efficiently explore the chemical space, provide a robust and reliable measure of TorsionNet prediction accuracy. As illustrated in Fig. 1c-1f, DNN profiles with large offsets show greater discrepancy to the corresponding reference DFT profile, even after offset correction has been applied. For the entire test set of 50k fragments, we examined the relationship between offset value and error in the offset-corrected DNN profiles relative to the reference DFT results. As shown in Supplementary Fig. 2, with

increasing offset, DNN profiles exhibit a larger discrepancy with respect to DFT. For example, the MAE of DNN relative energies of a subset of profiles with low offset (absolute offset value ≤ 0.25 kcal/mol) is about 3x smaller compared to another subset with higher offset (absolute offset value ≥ 1.0 kcal/mol): 0.5 kcal/mol vs. 1.7 kcal/mol. This ability of our method to identify less accurate predictions at no additional computational cost has significant practical utility since in prospective applications, it is important that the models not only make correct predictions but also that they can determine when the predictions might not be accurate.

Comparison with ANI-2x

We assessed the performance of TorsionNet relative to ANI-2x⁴⁶, a recent neural network-based energy model that has been shown to significantly outperform common MM and semi-empirical methods³⁶. We compared how closely torsion energy profiles generated from this method match those calculated from DFT using the same fragments in our test set. Supplementary Figure 3 shows a detailed comparison of ANI-2x relative to DFT. Distribution of per-profile Pearson's correlation between ANI-2x and DFT (Suppl. Fig. 3a) is shifted to significantly lower values compared to the corresponding distribution shown for TorsionNet in Fig 3e. Whereas 82% of TorsionNet profiles have ≥ 0.9 Pearson's correlation to DFT, only 59% of ANI-2x profiles exceed this threshold. Comparison of the relative energies of intermediate torsion angle conformations shows a similar trend; in comparison to ANI-2x (Suppl. Fig. 3b), TorsionNet shows significantly higher Pearson's correlation (0.94 vs. 0.83) and smaller discrepancy (MAE of 0.7 kcal/mol for DNN vs. 1.2 kcal/mol for ANI-2x) with respect to DFT (Fig. 3f). Our method also outperforms ANI-2x on the TorsionNet500 benchmark as shown in Table 1.

Agreement between DNN Profiles and Preferred Torsion Angles in Crystal Structures

As further validation of TorsionNet, we compared DNN profiles to the dihedral angle distribution of matching substructures in crystal structures. We used an existing dataset comprised of torsion motifs and the corresponding histogram of observed dihedral angle distribution from our earlier work²³, where DFT profiles were compared to preferred dihedral

angles observed in crystal structures. Here, we repeated that analysis, but used TorsionNet-generated profiles instead of DFT.

For a representative set of four frequently occurring motifs, TorsionNet profiles and the corresponding histogram of torsion angles derived from CSD and PDB are shown in Fig. 4. A high degree of complementarity between calculated profiles and observed distributions can be seen; prominent peaks on the histogram plots, which indicate energetically favorable torsion angles, coincide with low-energy minima on the DNN profiles, further demonstrating the predictive power of TorsionNet.

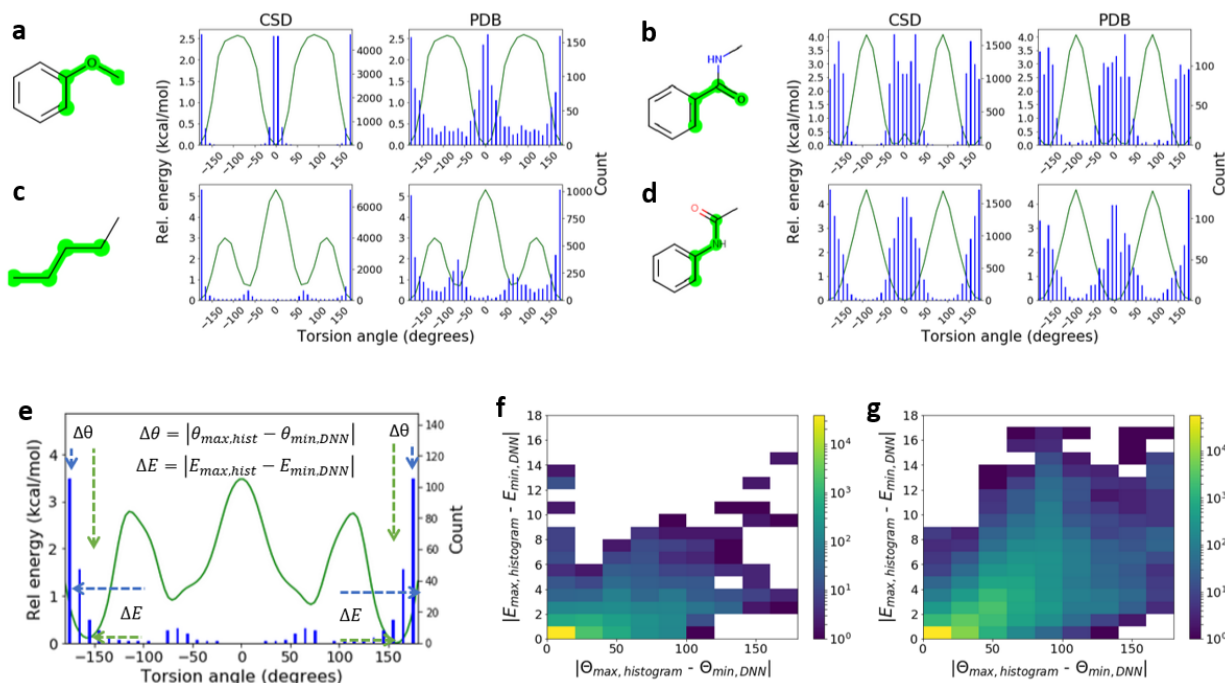


Figure 4. (a-d) Comparison of DNN torsion energy profile and torsion angle distribution of commonly observed fragments. Each torsion profile was generated by rotating the highlighted torsion (green) in the fragment shown on the left. Histograms of torsion angle distributions from the CSD and the PDB are shown in the middle and right panels, respectively. (e) Schematics illustrating how we calculated torsion angle deviations between histogram peak and the corresponding local energy minima. (f-g) 2D histograms of torsion angle deviations between the preferred crystal conformation and the nearest local minimum on the corresponding TorsionNet profile. (f) CSD. (g) PDB. Each histogram is rendered using a logarithmic color scale according to the number of data points in each bin, yellow (largest) to purple (smallest).

We performed a more comprehensive assessment of the ability of our model to identify preferred dihedral geometries in CSD and PDB crystal structures. As illustrated by the schematics presented in Fig. 4e, for each fragment in this set, we calculated the deviation of each prominent histogram peak ($\theta_{max,hist}$) to the nearest local minimum on the corresponding DNN profile ($\theta_{min,DNN}$) as well as relative energy difference at those two dihedral angles ($|E_{max,hist} - E_{min,DNN}|$). To identify local energy minima, we first transformed each discrete torsion profile into a continuous function by cubic interpolation using `interp1d` function from `scipy`⁴⁵. Subsequently, the local energy minimum, $\theta_{min,DNN}$, was determined by minimizing the interpolation function (using the `minimize` function from `scipy.optimize` module) with $\theta_{max,hist}$ as the initial guess. The distribution of $\Delta\theta$ and ΔE derived from the CSD (39k histogram peaks in 14k motifs) and the PDB (57k histogram peaks in 20k motifs) datasets is shown as a 2D histogram plot in Fig. 4f and Fig. 4g, respectively. Both CSD and PDB distributions are highly concentrated near the origin, bottom-left corner of the plot, indicating a high level of agreement between experimental data (preferred torsion angle in crystal structures) and DNN predictions (energetically favorable torsion angle) for most data points. Specifically, we find that 86% of CSD histogram peaks and 65% from the PDB set are located within 20° and 1.0 kcal/mol from the nearest local energy minimum. Encouragingly, these results are very similar to what we previously observed based on a similar analysis of the same dataset using DFT torsion profiles, giving us further confidence that computationally expensive QM calculations can be replaced by TorsionNet without any noticeable loss of accuracy in most cases.

Nonetheless, larger discrepancies between histogram peaks and predicted low-energy minima can be seen both for CSD (Fig. 4f) and to a larger extent for PDB sets (Fig. 4g). As demonstrated in our earlier DFT-based analysis, most of these outliers can be attributed to external factors such as intermolecular interactions, steric clash, and crystal packing effects.

Application 1: Crystal Structure Ranking in Conformational Ensemble

Generating physically and biologically relevant conformations is critical to various drug discovery applications, including virtual screening and structure-based design. A variety of

conformational sampling algorithms and programs^{10-13, 47, 48} have been developed, and many of these programs, such as Omega¹⁰, utilize the knowledge base of allowed torsion angles to limit conformational sampling to physically relevant regions of dihedral angle space. Conformational sampling is usually followed by energy-based minimization and ranking to prioritize low-energy conformations that are more likely to be observed in the solid-state or as a bioactive structure in complex with a protein. Thus, the ability of a conformer generator to generate and preferentially rank physically relevant structures from a vast number of possible low-energy conformations highly depends on the quality of energy function used. We, therefore, tested whether TorsionNet could rank experimental structures seeded in conformational pools with greater efficiency compared to an alternate force field-based approach.

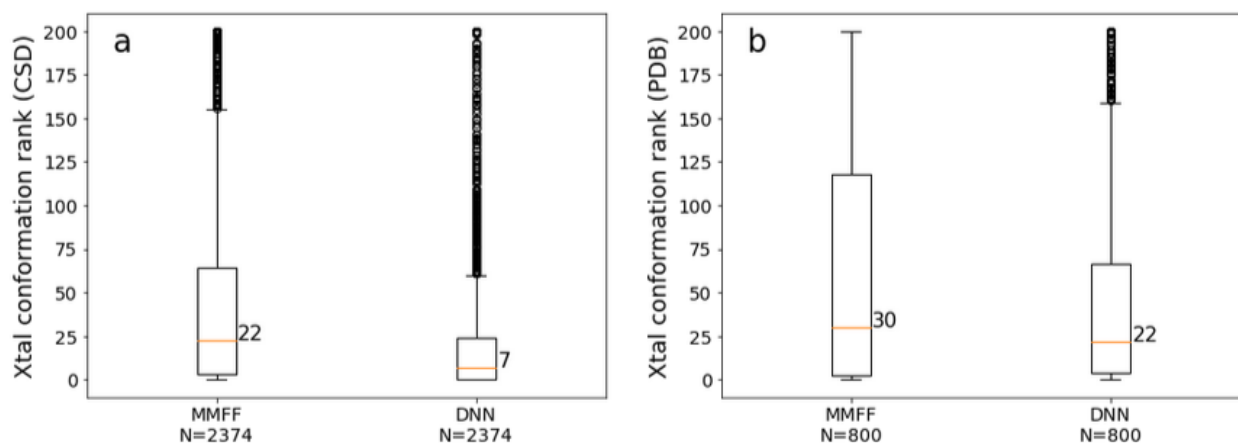


Figure 5. Ranking of crystal structures seeded in conformation ensembles from (a) CSD and (b) PDB datasets. For each test set structure, conformers and crystal structures were ranked using MMFF energy (excluding the stretch and angle bend components) and TorsionNet strain energy. Here, lower rank represents better performance.

We analyzed two sets of crystal structures from the CSD and the PDB: the CSD subset, which was adopted from our earlier work²³, and the PDB subset, where the structures were derived from the Platinum Diverse Dataset⁴⁹. Structures in these two subsets were selected from parent databases (CSD or PDB) based on a number of druglike or structure quality filters as described in respective publications. To better assess the role of torsional strain in this analysis, we applied

additional filters and removed molecules containing fewer than four or greater than six rotatable bonds. Molecules, where fewer than 30 conformers could be generated by Omega with default settings, were also discarded. The distributions of key physicochemical properties of the resulting dataset, comprising 2.4k structures from CSD and 800 from PDB, are shown in Supplementary Figures 4 and 5. For each crystal structure in the two sets, we generated up to 200 conformations using the Omega Python toolkit (release 2.9.1, OpenEye Scientific Software, Santa Fe, NM) with default settings. Next, we combined each conformation ensemble with its corresponding crystal structure. We then rank-ordered conformations in each ensemble using TorsionNet and MMFF¹⁴ energies. To avoid the outsized influence of stretch and angle bend components on the total force field energy, which can occur due to even the slightest deviation of bond lengths and bond angles from their pre-specified ideal value, we excluded contributions of these terms from the total MMFF energy.

We used the ranking of crystal structures in the corresponding conformation ensemble as the assessing criterion: a lower rank represents better performance. The boxplots in Fig. 5 show the distribution of crystal structure rank using TorsionNet and MMFF. TorsionNet significantly outperformed MMFF for the CSD set (median rank 7 vs 22) and also produced a lower median rank for the PDB set (median rank 22 vs 30). That torsional strain alone can produce such improvement in crystal structure rank is remarkable; it reinforces the importance of torsions in determining small molecule conformations, and further highlights the need for fast and accurate methods to assess dihedral energetics.

While TorsionNet produces a substantially better ranking of CSD crystal structures compared to MMFF, the improvement observed for the PDB set is relatively small, where a greater number of Omega conformations with lower energy relative to the bioactive structures are observed. This would imply that protein-bound ligand conformations exhibit higher strain compared to the small molecule structures in CSD. However, as our previous DFT-based analysis has shown, higher strain in PDB structures are usually due to artifacts of crystal structure refinement and can be attributed to the generally lower quality of PDB structures compared to those found in the CSD. Therefore, the relatively smaller amount of improvement of TorsionNet in the PDB set is more

likely to be a result of poorer quality of these structures rather than the limitation of the DNN model specific to bioactive conformations.

Application 2: Structure-based Virtual Screening Hit Rates

Structure-based docking scores have been mostly developed to capture the key components of protein-ligand interaction through empirical terms, generally seeking a balance between speed and accuracy. Ligand conformational strain, an important factor in total binding affinity, is either neglected by most docking programs or characterized by approximate molecular mechanics energy functions⁵⁰. We investigated the potential benefit of applying TorsionNet to structure-based virtual screening, examining whether applying strain energy correction as a post-processing step to score docking hits can improve hit rates.

We utilized a proprietary docking dataset comprised of four protein targets, each with 100 known actives that were designed and optimized for potency against their respective target by in-house medicinal chemistry programs. In addition, a common set of 100k known inactive compounds, selected based on internal HTS screening against each individual target, were included as a decoy set. Docking was performed against each target as described in Supplementary Note 2. Next, the top-ranked docking poses (one pose for each molecule) were scored with GlideScore⁵¹. Additionally, the DNN torsional strain energy of the same set of docked poses was calculated using the procedure described in Materials and Methods. Subsequently, the docked poses were rank-ordered based on GlideScore or a composite score, $GlideScore + w \cdot E_{strain}$, representing a weighted combination of GlideScore and the torsional strain energy. The weight, w , of the strain energy component in the composite score was manually adjusted and set to 0.25, as it produced a better hit rate across all four targets compared to other values we explored. Finally, the hit rate of the top 1% poses was calculated for all four targets based on the ranking by GlideScore and the composite, strain corrected, scores. As shown in Fig. 6, strain correction results in substantial improvement in the recovery rate of actives over the baseline (GlideScore), ranging from 10% improvement for target D to over 100% improvement for target A.

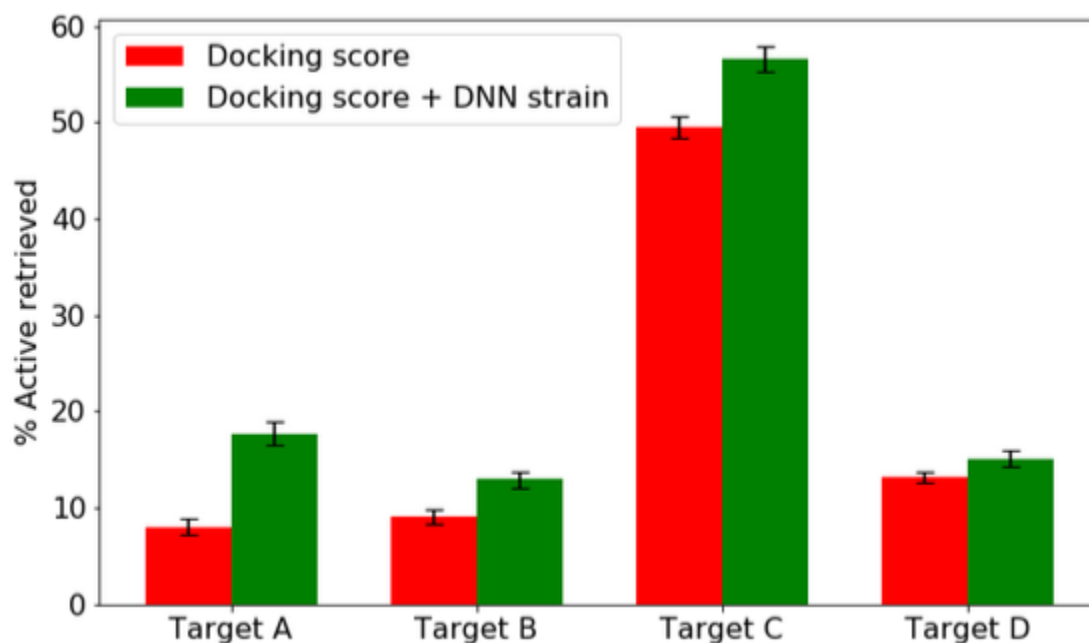


Figure 6. Virtual screening hit rate obtained using top 1% of the docked poses against four proprietary targets. For each target, a library of compounds comprised of 100 known actives against that target and a common set of 100k known inactive compounds identified from in-house HTS screening campaigns against these four targets was used for docking. Docking poses were ranked using Glide score (red) and re-ranked based on DNN-based strain correction to the Glide score (green). Error bars represent 95% confidence interval, calculated using 100 bootstrap samples.

The impact of strain energy correction will likely vary depending on the physicochemical characteristics of the ligand and the binding site. For example, strain correction can be expected to have a larger impact on molecules with more rotatable bonds. In contrast, the effect of strain correction will be less pronounced when the binding is dominated by strong polar or salt-bridge interactions, as in the case of target D. Nonetheless, the results presented here clearly highlight the importance of incorporating a strain energy term into existing scoring functions to reduce the high false-positive rate that most virtual screening tools suffer from.

It should also be noted that the dataset we assembled for this proof-of-concept illustration follows the standard practice in evaluating virtual screening performance, whereby the known actives against the target of interest are added to a large library of decoys, which are presumed to be inactive against that target. This, however, does not accurately reflect the performance of a

virtual screening tool in prospective settings since databases screened for hits against a novel target rarely contain ligands that were optimized for potency against that specific target. Therefore, retrospective virtual screening, due to the composition of screening databases, tends to give a false impression of producing high hit rates.

Application 3: Refinement of Bound Ligand in Co-crystal Structures

The success of the structure-based design is predicated upon the availability of high-quality crystal structures with unambiguously determined atomic coordinates of the bound ligand and key binding site residues. However, incomplete electron density around the binding site region and other artifacts of crystal structure refinement can often result in PDB structures with poorly defined ligand coordinates^{52, 53}, even for structure with relatively high resolution⁵⁴. As the commonly used crystal structure refinement protocols mostly rely on molecular mechanics energy functions, the low quality of force field parameters often result in strained ligand geometries, especially when the electron density around the binding site region is not well characterized.

Here, we demonstrate that unreasonable or high-energy ligand conformation in PDB co-crystal structures can be corrected through a simple re-refinement procedure by utilizing TorsionNet profiles. Taking the initial crystal structure of the bound ligand and its associated electron density as the only input, our method simultaneously minimizes the total torsional strain by optimizing the dihedral angle of each rotatable bond and maximizes its electron density overlap through rigid rotation and translation of the entire molecule, thus, generating a lower strain structure that is also compatible with the experimental electron density. Optimization is done by minimizing the objective function, $E_{strain} - w \cdot O_{tanimoto}$, where $E_{strain} = \sum_{i=1}^{N_{\tau}} E_{\tau, strain}$ is the sum of strain energy from all N_{τ} constituent torsions, $O_{tanimoto}$ represents the Tanimoto overlap between the experimental electron density and a Gaussian representation of the molecule, and w is an adjustable weight parameter (default value 10). Tanimoto overlap, $O_{tanimoto}$, was calculated using OEShape toolkit (release 2.0.4, OpenEye Scientific Software, Santa Fe, NM). The optimization procedure was implemented as a Python script using OEFF toolkit (release 2.1.0, OpenEye Scientific Software, Santa Fe, NM). Our approach is conceptually similar to the

method underlying AFITT⁵⁵, a commercially available program. However, unlike AFITT, where internal strain is minimized using the full molecular mechanics force field, our method optimizes only the dihedral component of the internal strain using TorsionNet.

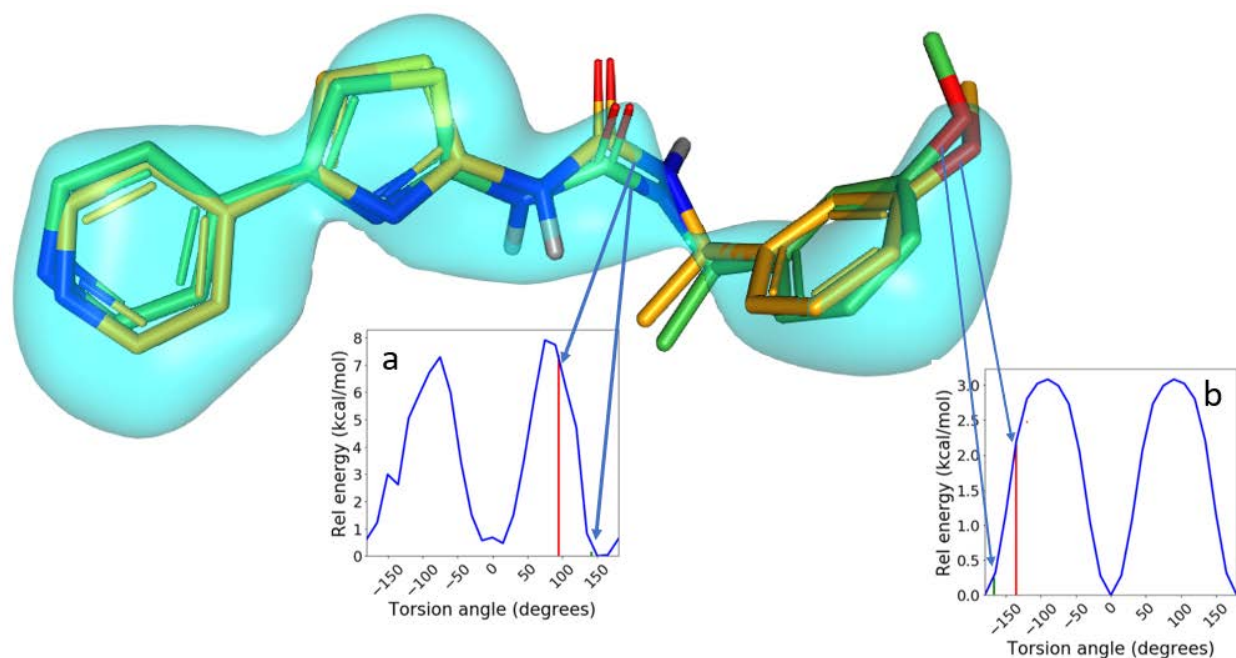


Figure 7. Overlay of the PDB ligand conformation (orange) and electron density (cyan) from PDB entry 3tv7 and the corresponding strain-minimized structure (green). The TorsionNet energy profile of two high-strain torsions are shown in panels a and b. The orange and green vertical lines on each plot represent the corresponding torsion angle in the initial and strain-minimized structures, respectively. For each torsion, the minimization process shifts the initial torsion angle (orange) to a new orientation (green) that is closer to the local minimum on the energy profile.

This re-refinement procedure is illustrated using the PDB example 3tv7, a 2.75 Å resolution co-crystal structure, with a bound ligand shown in Fig. 7 (PDB conformation depicted in orange) along with the experimental electron density (cyan) in the vicinity of the ligand. The ligand conformation in this PDB structure exhibits unusually high strain mainly because the central amide and the terminal methoxy torsions are significantly distorted from the energetically favorable planar orientation, likely an artifact of crystal structure refinement due to incomplete electron density around these motifs. Application of our strain minimization procedure to this

ligand shifts each dihedral angle closer to their nearest local minimum on the corresponding TorsionNet profile, yielding a conformation that has a significantly lower strain compared to the deposited PDB structure while maintaining a similar level of overlap with the experimental electron density. In particular, amide and methoxy torsions in the TorsionNet-optimized structure (Fig. 7, molecule depicted in green) adopt nearly planar orientations, compared to the somewhat perpendicular, high-energy orientation in the deposited structure, resulting in a substantial reduction in the strain energy of these two torsions, as depicted in the corresponding profiles shown in Fig. 7a and 7b.

CONCLUSIONS

In this work, we presented TorsionNet, a deep neural network model, that enables the rapid generation of small molecule torsion energy profiles with DFT accuracy. By combining prediction uncertainty estimates with active learning, we were able to automate chemical space sampling, allowing us to obtain a balanced coverage of the large medicinal chemistry space with a relatively small number of DFT calculations (50k fragments out of the available 3 million). This was, nonetheless, a major undertaking that required over 12 million CPU hours on the Cloud. This one-time computational investment allowed us to train a robust ML model that was shown to produce DFT-quality results with only a fraction of the time required by the corresponding QM calculations. TorsionNet achieves over two orders of magnitude speedup over DFT by replacing computationally expensive geometry optimization and energy evaluation steps in a typical QM calculation with a trained neural network model where inference is done using AEVs that are derived directly from a MM structure. In a large-scale evaluation, TorsionNet was able to accurately identify the preferred dihedral geometries of matching substructures in crystal structures of druglike molecules. The proof-of-concept application examples presented here demonstrate that incorporating DNN-based strain energy characterization can substantially improve the accuracy of molecular modeling workflows such as conformational analysis, virtual screening, and refinement of bioactive ligands in crystal structures.

Unlike molecular mechanics force fields, TorsionNet is not constrained by a fixed functional form or a pre-specified set of parameters. This flexibility allows TorsionNet to capture the underlying complexities of the high dimensional potential energy surface with high accuracy. Just as with any other machine learning model, the performance of TorsionNet is expected to degrade when it is applied to molecules containing novel chemistries that are not adequately represented in the training set. However, such examples can be readily identified using the proposed offset-based uncertainty measure, another advantageous feature of our workflow. Unlike molecular mechanics force fields, where any update to the existing model would require significant manual effort, TorsionNet can be re-trained simply by augmenting the existing training set with new data.

In contrast to earlier works (e.g. ANI, SchNet), we trained our model to predict one specific and important energy component, instead of attempting to learn a general potential energy function for any given small molecule, which remains a challenging task, despite recently enabled capabilities for generating large amounts of reference QM data and training machine learning models on large datasets⁵⁶.

The methodology described here can be extended to other energy terms that are important in determining small molecule conformations and protein-ligand binding. Our approach is particularly well-suited to describe the energetics of saturated and partially saturated ring systems, where active learning can be applied in a similar fashion to capture the huge chemical diversity of rings and their substitution patterns found in druglike molecules with a relatively small number of reference QM calculations. We envision modeling non-bonded interactions such as hydrogen bond, pi interactions etc. as other potential extensions of our approach. Here again, high-level QM calculations (e.g. MP2) along with active learning could be leveraged to efficiently characterize the chemical and configurational space of structural motifs relevant to protein-ligand binding. Compared to the first principle and other theoretically rigorous but computationally prohibitive methods (e.g. FEP), well-constructed and trained ML models are likely to offer the best tradeoff between speed and accuracy for various high-throughput drug discovery applications.

Corresponding Author

*brajesh.rai@pfizer.com

Notes

The authors declare no competing finance interests.

Supporting Information

DFT-optimized geometry and energy of 12k TorsionNet500 conformers. Figures showing the key physicochemical properties of the torsion fragments and molecules in the test datasets, the performance of TorsionNet at different offset levels, and the performance of ANI-2x on the cross-validation dataset.

Python scripts implementing various components of TorsionNet are publicly available on <http://github.com/PfizerRD/torsion-strain>.

ACKNOWLEDGMENT

We thank Geoff Skillman and the entire Orion development and support team for their help with running QM torsion scan calculations on the AWS. Scott Parker and Dawn Wawrzynowicz are thanked for their help with the logistics of access to Orion compute environment on the AWS. We gratefully acknowledge Enoch Huang, Sandy Farmer, Tamara Mansfeld, and Pfizer Innovation Operating Council for funding the cloud computing expenses. Frank Lovering is thanked for his feedback on the manuscript. We thank Jacquelyn Klug-McLeod for her help with generating the TorsionNet500 benchmark dataset.

REFERENCES

1. Hawkins, P. C. D., Conformation Generation: The State of the Art. *Journal of Chemical Information and Modeling* **2017**, 57 (8), 1747-1756.
2. Irwin, J. J.; Shoichet, B. K., Docking Screens for Novel Ligands Conferring New Biology. *Journal of Medicinal Chemistry* **2016**, 59 (9), 4103-4120.
3. Ripphausen, P.; Nisius, B.; Bajorath, J., State-of-the-art in ligand-based virtual screening. *Drug Discovery Today* **2011**, 16 (9), 372-376.
4. Groom, C. R.; Bruno, I. J.; Lightfoot, M. P.; Ward, S. C., The Cambridge Structural Database. *Acta Crystallographica Section B* **2016**, 72 (2), 171-179.
5. Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E., The Protein Data Bank. *Nucleic Acids Research* **2000**, 28 (1), 235-242.
6. Schärfer, C.; Schulz-Gasch, T.; Ehrlich, H.-C.; Guba, W.; Rarey, M.; Stahl, M., Torsion Angle Preferences in Druglike Chemical Space: A Comprehensive Guide. *Journal of Medicinal Chemistry* **2013**, 56 (5), 2016-2028.
7. Brameld, K. A.; Kuhn, B.; Reuter, D. C.; Stahl, M., Small Molecule Conformational Preferences Derived from Crystal Structure Data. A Medicinal Chemistry Focused Analysis. *Journal of Chemical Information and Modeling* **2008**, 48 (1), 1-24.
8. Bruno, I. J.; Cole, J. C.; Kessler, M.; Luo, J.; Motherwell, W. D. S.; Purkis, L. H.; Smith, B. R.; Taylor, R.; Cooper, R. I.; Harris, S. E.; Orpen, A. G., Retrieval of Crystallographically-Derived Molecular Geometry Information. *Journal of Chemical Information and Computer Sciences* **2004**, 44 (6), 2133-2144.
9. Li, J.; Ehlers, T.; Sutter, J.; Varma-O'Brien, S.; Kirchmair, J., CAESAR: A New Conformer Generation Algorithm Based on Recursive Buildup and Local Rotational Symmetry Consideration. *Journal of Chemical Information and Modeling* **2007**, 47 (5), 1923-1932.
10. Hawkins, P. C. D.; Skillman, A. G.; Warren, G. L.; Ellingson, B. A.; Stahl, M. T., Conformer Generation with OMEGA: Algorithm and Validation Using High Quality Structures from the Protein Databank and Cambridge Structural Database. *Journal of Chemical Information and Modeling* **2010**, 50 (4), 572-584.
11. Watts, K. S.; Dalal, P.; Murphy, R. B.; Sherman, W.; Friesner, R. A.; Shelley, J. C., ConfGen: A Conformational Search Method for Efficient Generation of Bioactive Conformers. *Journal of Chemical Information and Modeling* **2010**, 50 (4), 534-546.
12. Riniker, S.; Landrum, G. A., Better Informed Distance Geometry: Using What We Know To Improve Conformation Generation. *Journal of Chemical Information and Modeling* **2015**, 55 (12), 2562-2574.
13. Friedrich, N.-O.; Flachsenberg, F.; Meyder, A.; Sommer, K.; Kirchmair, J.; Rarey, M., Conformer: A Novel Method for the Generation of Conformer Ensembles. *Journal of Chemical Information and Modeling* **2019**, 59 (2), 731-742.
14. Halgren, T. A., MMFF VII. Characterization of MMFF94, MMFF94s, and other widely available force fields for conformational energies and for intermolecular-interaction energies and geometries. *Journal of Computational Chemistry* **1999**, 20 (7), 730-748.
15. Vanommeslaeghe, K.; Hatcher, E.; Acharya, C.; Kundu, S.; Zhong, S.; Shim, J.; Darian, E.; Guvench, O.; Lopes, P.; Vorobyov, I.; Mackerell Jr., A. D., CHARMM general force field: A force field for drug-like molecules compatible with the CHARMM all-atom additive biological force fields. *Journal of Computational Chemistry* **2010**, 31 (4), 671-690.
16. Harder, E.; Damm, W.; Maple, J.; Wu, C.; Reboul, M.; Xiang, J. Y.; Wang, L.; Lupyan, D.; Dahlgren, M. K.; Knight, J. L.; Kaus, J. W.; Cerutti, D. S.; Krilov, G.; Jorgensen, W. L.; Abel, R.; Friesner,

- R. A., OPLS3: A Force Field Providing Broad Coverage of Drug-like Small Molecules and Proteins. *Journal of Chemical Theory and Computation* **2016**, *12* (1), 281-296.
17. Horta, B. A. C.; Merz, P. T.; Fuchs, P. F. J.; Dolenc, J.; Riniker, S.; Hünenberger, P. H., A GROMOS-Compatible Force Field for Small Organic Molecules in the Condensed Phase: The 2016H66 Parameter Set. *Journal of Chemical Theory and Computation* **2016**, *12* (8), 3825-3850.
18. Roos, K.; Wu, C.; Damm, W.; Reboul, M.; Stevenson, J. M.; Lu, C.; Dahlgren, M. K.; Mondal, S.; Chen, W.; Wang, L.; Abel, R.; Friesner, R. A.; Harder, E. D., OPLS3e: Extending Force Field Coverage for Drug-Like Small Molecules. *Journal of Chemical Theory and Computation* **2019**, *15* (3), 1863-1874.
19. Kaminski, G. A.; Friesner, R. A.; Tirado-Rives, J.; Jorgensen, W. L., Evaluation and Reparametrization of the OPLS-AA Force Field for Proteins via Comparison with Accurate Quantum Chemical Calculations on Peptides. *The Journal of Physical Chemistry B* **2001**, *105* (28), 6474-6487.
20. Wang, L.; Wu, Y.; Deng, Y.; Kim, B.; Pierce, L.; Krilov, G.; Lupyan, D.; Robinson, S.; Dahlgren, M. K.; Greenwood, J.; Romero, D. L.; Masse, C.; Knight, J. L.; Steinbrecher, T.; Beuming, T.; Damm, W.; Harder, E.; Sherman, W.; Brewer, M.; Wester, R.; Murcko, M.; Frye, L.; Farid, R.; Lin, T.; Mobley, D. L.; Jorgensen, W. L.; Berne, B. J.; Friesner, R. A.; Abel, R., Accurate and Reliable Prediction of Relative Ligand Binding Potency in Prospective Drug Discovery by Way of a Modern Free-Energy Calculation Protocol and Force Field. *Journal of the American Chemical Society* **2015**, *137* (7), 2695-2703.
21. Allen, F. H.; Harris, S. E.; Taylor, R., Comparison of conformer distributions in the crystalline state with conformational energies calculated by ab initio techniques. *Journal of Computer-Aided Molecular Design* **1996**, *10* (3), 247-254.
22. Hao, M.-H.; Haq, O.; Muegge, I., Torsion Angle Preference and Energetics of Small-Molecule Ligands Bound to Proteins. *Journal of Chemical Information and Modeling* **2007**, *47* (6), 2242-2252.
23. Rai, B. K.; Sresht, V.; Yang, Q.; Unwalla, R.; Tu, M.; Mathiowetz, A. M.; Bakken, G. A., Comprehensive Assessment of Torsional Strain in Crystal Structures of Small Molecules and Protein-Ligand Complexes using ab Initio Calculations. *Journal of Chemical Information and Modeling* **2019**, *59* (10), 4195-4208.
24. Gutiérrez Sanfeliciano, S. M.; Schaus, J. M., Rapid assessment of conformational preferences in biaryl and aryl carbonyl fragments. *PLOS ONE* **2018**, *13* (3), e0192974.
25. Dahlgren, M. K.; Schyman, P.; Tirado-Rives, J.; Jorgensen, W. L., Characterization of Biaryl Torsional Energetics and its Treatment in OPLS All-Atom Force Fields. *Journal of Chemical Information and Modeling* **2013**, *53* (5), 1191-1199.
26. Butler, K. T.; Davies, D. W.; Cartwright, H.; Isayev, O.; Walsh, A., Machine learning for molecular and materials science. *Nature* **2018**, *559* (7715), 547-555.
27. von Lilienfeld, O. A., Quantum Machine Learning in Chemical Compound Space. **2018**, *57* (16), 4164-4169.
28. Rupp, M.; Tkatchenko, A.; Müller, K.-R.; von Lilienfeld, O. A., Fast and Accurate Modeling of Molecular Atomization Energies with Machine Learning. *Physical Review Letters* **2012**, *108* (5), 058301.
29. Blum, L. C.; Raymond, J.-L., 970 Million Druglike Small Molecules for Virtual Screening in the Chemical Universe Database GDB-13. *Journal of the American Chemical Society* **2009**, *131* (25), 8732-8733.
30. Rai, B. K.; Bakken, G. A., Fast and accurate generation of ab initio quality atomic charges using nonparametric statistical regression. **2013**, *34* (19), 1661-1671.
31. Pereira, F.; Xiao, K.; Latino, D. A. R. S.; Wu, C.; Zhang, Q.; Aires-de-Sousa, J., Machine Learning Methods to Predict Density Functional Theory B3LYP Energies of HOMO and LUMO Orbitals. *Journal of Chemical Information and Modeling* **2017**, *57* (1), 11-21.
32. Bleiziffer, P.; Schaller, K.; Riniker, S., Machine Learning of Partial Charges Derived from High-Quality Quantum-Mechanical Calculations. *Journal of Chemical Information and Modeling* **2018**, *58* (3), 579-590.

33. Unke, O. T.; Muwly, M., PhysNet: A Neural Network for Predicting Energies, Forces, Dipole Moments, and Partial Charges. *Journal of Chemical Theory and Computation* **2019**, *15* (6), 3678-3693.
34. Feinberg, E. N.; Sur, D.; Wu, Z.; Husic, B. E.; Mai, H.; Li, Y.; Sun, S.; Yang, J.; Ramsundar, B.; Pande, V. S., PotentialNet for Molecular Property Prediction. *ACS Central Science* **2018**, *4* (11), 1520-1530.
35. Gastegger, M.; Behler, J.; Marquetand, P., Machine learning molecular dynamics for the simulation of infrared spectra. *Chemical Science* **2017**, *8* (10), 6924-6935.
36. Smith, J. S.; Isayev, O.; Roitberg, A. E., ANI-1: an extensible neural network potential with DFT accuracy at force field computational cost. *Chemical Science* **2017**, *8* (4), 3192-3203.
37. Schütt, K. T.; Sauceda, H. E.; Kindermans, P.-J.; Tkatchenko, A.; Müller, K.-R., SchNet – A deep learning architecture for molecules and materials. **2018**, *148* (24), 241722.
38. Smith, J. S.; Nebgen, B. T.; Zubatyuk, R.; Lubbers, N.; Devereux, C.; Barros, K.; Tretiak, S.; Isayev, O.; Roitberg, A. E., Approaching coupled cluster accuracy with a general-purpose neural network potential through transfer learning. *Nature Communications* **2019**, *10* (1), 2903.
39. Chmiela, S.; Tkatchenko, A.; Sauceda, H. E.; Poltavsky, I.; Schütt, K. T.; Müller, K.-R., Machine learning of accurate energy-conserving molecular force fields. **2017**, *3* (5), e1603015.
40. Yao, K.; Herr, J. E.; Toth, David W.; McKintyre, R.; Parkhill, J., The TensorMol-0.1 model chemistry: a neural network augmented with long-range physics. *Chemical Science* **2018**, *9* (8), 2261-2269.
41. Halgren, T. A., MMFF VI. MMFF94s option for energy minimization studies. **1999**, *20* (7), 720-729.
42. Parrish, R. M.; Burns, L. A.; Smith, D. G. A.; Simmonett, A. C.; DePrince, A. E.; Hohenstein, E. G.; Bozkaya, U.; Sokolov, A. Y.; Di Remigio, R.; Richard, R. M.; Gonthier, J. F.; James, A. M.; McAlexander, H. R.; Kumar, A.; Saitow, M.; Wang, X.; Pritchard, B. P.; Verma, P.; Schaefer, H. F.; Patkowski, K.; King, R. A.; Valeev, E. F.; Evangelista, F. A.; Turney, J. M.; Crawford, T. D.; Sherrill, C. D., Psi4 1.1: An Open-Source Electronic Structure Program Emphasizing Automation, Advanced Libraries, and Interoperability. *Journal of Chemical Theory and Computation* **2017**, *13* (7), 3185-3197.
43. Behler, J.; Parrinello, M., Generalized Neural-Network Representation of High-Dimensional Potential-Energy Surfaces. *Physical Review Letters* **2007**, *98* (14), 146401.
44. Behler, J., Constructing high-dimensional neural network potentials: A tutorial review. **2015**, *115* (16), 1032-1050.
45. Eric Jones, T. O., and Pearu Peterson, SciPy: Open source scientific tools for Python. **2001**.
46. Devereux, C.; Smith, J. S.; Davis, K. K.; Barros, K.; Zubatyuk, R.; Isayev, O.; Roitberg, A. E., Extending the Applicability of the ANI Deep Learning Molecular Potential to Sulfur and Halogens. *Journal of Chemical Theory and Computation* **2020**, *16* (7), 4192-4202.
47. Schärfer, C.; Schulz-Gasch, T.; Hert, J.; Heinzerling, L.; Schulz, B.; Inhester, T.; Stahl, M.; Rarey, M., CONFECT: Conformations from an Expert Collection of Torsion Patterns. *ChemMedChem* **2013**, *8* (10), 1690-1700.
48. Cole, J. C.; Korb, O.; McCabe, P.; Read, M. G.; Taylor, R., Knowledge-Based Conformer Generation Using the Cambridge Structural Database. *Journal of Chemical Information and Modeling* **2018**, *58* (3), 615-629.
49. Friedrich, N.-O.; Meyder, A.; de Bruyn Kops, C.; Sommer, K.; Flachsenberg, F.; Rarey, M.; Kirchmair, J., High-Quality Dataset of Protein-Bound Ligand Conformations and Its Application to Benchmarking Conformer Ensemble Generators. *Journal of Chemical Information and Modeling* **2017**, *57* (3), 529-539.
50. Murphy, R. B.; Repasky, M. P.; Greenwood, J. R.; Tubert-Brohman, I.; Jerome, S.; Annabhimoju, R.; Boyles, N. A.; Schmitz, C. D.; Abel, R.; Farid, R.; Friesner, R. A., WScore: A Flexible and

Accurate Treatment of Explicit Water Molecules in Ligand–Receptor Docking. *Journal of Medicinal Chemistry* **2016**, 59 (9), 4364-4384.

51. Friesner, R. A.; Murphy, R. B.; Repasky, M. P.; Frye, L. L.; Greenwood, J. R.; Halgren, T. A.; Sanschagrin, P. C.; Mainz, D. T., Extra Precision Glide: Docking and Scoring Incorporating a Model of Hydrophobic Enclosure for Protein–Ligand Complexes. *Journal of Medicinal Chemistry* **2006**, 49 (21), 6177-6196.

52. Liebeschuetz, J.; Hennemann, J.; Olsson, T.; Groom, C. R., The good, the bad and the twisted: a survey of ligand geometry in protein crystal structures. *Journal of Computer-Aided Molecular Design* **2012**, 26 (2), 169-183.

53. Deller, M. C.; Rupp, B., Models of protein–ligand crystal structures: trust, but verify. *Journal of Computer-Aided Molecular Design* **2015**, 29 (9), 817-836.

54. Reynolds, C. H., Protein–Ligand Cocrystal Structures: We Can Do Better. *ACS Medicinal Chemistry Letters* **2014**, 5 (7), 727-729.

55. Janowski, P. A.; Moriarty, N. W.; Kelley, B. P.; Case, D. A.; York, D. M.; Adams, P. D.; Warren, G. L., Improved ligand geometries in crystallographic refinement using AFITT in PHENIX. *Acta Crystallographica Section D* **2016**, 72 (9), 1062-1072.

56. Folmsbee, D.; Hutchison, G., Assessing conformer energies using electronic structure and machine learning methods. *n/a* (n/a), e26381.

TOC Figure

