Auto3D: Automatic Generation of the Low-energy 3D Structures with ANI Neural Network Potentials

Zhen Liu[1]; Tetiana Zubatiuk[1], Adrian Roitberg[2], Olexandr Isayev[1*]

[1]Department of Chemistry, Mellon College of Science, Carnegie Mellon University, Pittsburgh, Pennsylvania 15213, USA

[2] Department of Chemistry, University of Florida, Gainesville, Florida 32611, USA

[*] Correspondence: olexandr@olexandrisayev.com (O.I.)

Abstract

Computational programs accelerate the chemical discovery processes but often need proper 3-dimensional molecular information as part of the input. Getting optimal molecular structures is challenging because it requires enumerating and optimizing a huge space of stereoisomers and conformers. We developed the Python-based Auto3D package for generating the low-energy 3D structures using SMILES as the input. Auto3D is based on state-of-the-art algorithms and can automatize the isomer enumeration and duplicate filtering process, 3D building process, geometry optimization and ranking process. Tested on 50 molecules with multiple unspecified stereocenters, Auto3D is guaranteed to find the stereo-configuration that yields the lowest-energy conformer. With Auto3D we provide an extension of the ANI model. The new model, dubbed ANI-2xt, is trained on a tautomer-rich dataset. ANI-2xt is benchmarked with DFT methods on geometry optimization, electronic and Gibbs free energy calculations. Compared with ANI-2x, ANI-2xt provides a 42% error reduction for tautomeric reaction energy calculations when using the gold-standard coupled-cluster calculation as the reference. ANI-2xt can accurately predict the energies and is several orders of magnitude faster than DFT methods.

1. Introduction

One of the tasks of cheminformatics is to work with chemical structures and represent them for various downstream applications. However, molecules that many chemists would consider different in the context of the task, are often recorded identically across multiple databases. Many chemical standards and identifiers like Chemical Abstracts Service (CAS) system[1], InChI[2], UNICHEM[3] and PubChem[4] were developed. Database conversions, curation errors and information loss also contribute to this problem. It is often a big challenge to extract and curate complex stereochemistry. The effect of data quality on ML model performance and the importance of careful data curation have long been recognized[5,6]. Accumulation of database errors and incorrect processing of chemical structures could lead to significant losses in the model performance[7].

The definition of isomerism is simple: it is a "consequence of the fact that the atoms of a molecular formula can be arranged in different ways to give compounds, that differ in physical and chemical properties."[8] If the process of isomerization is well understood from a chemical point of view, then what is the complication of modeling the isomeric space and finding the proper isomeric configuration? The phenomenon of isomerism is more complex for cheminformatics than it might appear. Molecules in different protonation states, salts and mixtures are considered equivalent in some contexts and different in others. The lack of comprehensive standards and robust representations for digital chemical structures adds to the uncertainties of the algorithms that convert identifiers back into a three-dimensional (3D) molecular structure. Furthermore, additional computational efforts are required to find a "true" 3D configurational stereoisomer and its corresponding (single or multiple) conformations.

Many applications of machine learning (ML) algorithms for physical-chemical property prediction, are hindered by the well-known problem of getting "true" 3D molecular coordinates from the 2D structures collected in public databases. Many chemical identifiers easily encode *constitutional isomers* (**Fig. 1**) from these 2D structures. In contrast, *stereoisomers* result from the dynamic nature of molecules. Their atoms are arranged differently but are held together through the same chemical bonds, according to the key IUPAC terms. Encoding the stereoisomeric features of compounds is complicated since they are present in two subcategories: *configurational stereoisomers* (this is the case for E/Z isomers of alkenes, R/S for chiral molecules) which can be isolated under certain conditions through their (usual, but not necessarily) high energy barriers of

interconversion; and *conformational stereoisomers* which interconvert over low barriers through the rotatable single bonds, which makes their isolation difficult and their practical encoding with identifiers becomes infeasible.

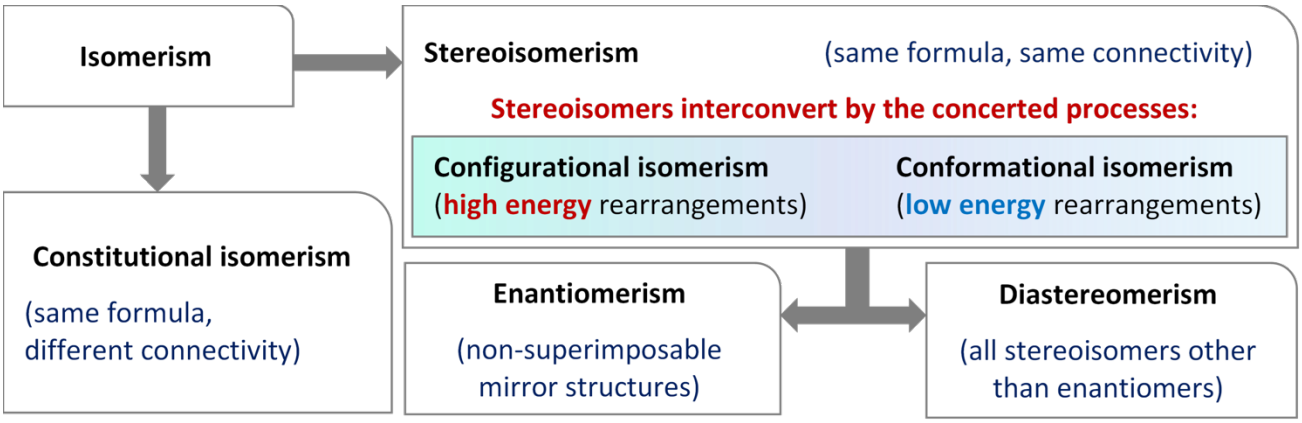


Figure 1. Isomerism hierarchy.

Thus, a large number of conformer sampling techniques (rule-based and/or data-based) have been developed and multiple studies have been carried out to compare the existing methods[9] and to explore the 3D conformational space for small molecules[10–12]. Several well-established commercial or open-source packages include Omega[13] from OpenEye[14], Catalyst[15], CREST[16], RDKit[17] and Molsoft[18].

Although the generation of the conformational ensembles is task-specific, the common assumption is that the configurational stereoisomer is properly encoded in the corresponding chemical identifier. However, that is not necessarily true. For example, the lack of information in the identifier becomes a serious concern if the chirality and/or *E/Z* configuration is not specified (**Fig.1**). Besides, most current algorithms suffer from the tradeoff between speed and accuracy. For example, Omega[13] optimizes 3D conformers with force fields for a fast speed but is compromised in accuracy. CREST[16] uses semiempirical quantum mechanical methods but is much slower.

In this article, we will only touch on some well-known algorithms for generating configurational and conformational spaces and searching for low-energy conformers. Our main goal is to create a practical and handy package for getting a suitable stereoisomeric 3D geometry from its 2D structure. It should be noted right away that the present approach does not yet consider

the physical environment and all calculations were performed in the gas phase. Although organic molecules are dynamic and could exist in several equivalent or interchangeable forms according to conditions, considering the physical environment in predictive algorithms is still a huge challenge for computational chemistry and ML. However, here we were guided by the generally accepted idea that calculations with the physical environment require at the first step, as input a suitable 3D molecular structure.

Here, we describe the open-source package, *Auto3D*, for getting the favorable 3D conformations of organic molecules from automatically generated stereoisomeric conformational space that is optimized by atomistic neural network potentials (NNPs) such as ANI[19] or AIMNet[20]. Auto3D serves us as a starting point, and we will extend it further toward a comprehensive analysis of molecular isomerism depending on environmental conditions. In this study, we first explained each component in Auto3D. Then we determined how the completeness of configurational and conformational spaces affects the molecular energy. The ANI methods and QM methods are benchmarked in terms of geometry optimization and tautomeric reaction energy calculations on the Nicklaus tautomer database[21,22]. In a final case study, we examined the behavior of Auto3D in the application for predicting tautomerization Gibbs free energy.

2. Methods

The high-level workflow for Auto3D is shown in Figure 2. For the input SMILES, Auto3D searches for the low-energy 3D structures by automatizing the tautomer enumeration, isomer enumeration, 3D building, geometry optimization, duplicate filtering and ranking step. Auto3D is accelerated with parallel 3D building and geometry optimization on GPUs. Currently, Auto3D supports two backends for isomer enumeration: one is commercial, and another is open-source. The licensed toolkit in Auto3D is based on the OpenEye[14] programming library. Specifically, Quacpac Toolkit[23] is used to generate tautomers; the Omega[13] Toolkit is used for stereoisomer enumeration and initial 3D conformer building with classic/macrocycle modes. Alternatively, RDKit[17] could be used for tautomer enumeration, stereoisomer enumeration, and 3D building, too. The optimization of the conformational space can be done by three NNPs: AIMNet[20], ANI-2x[24] or the new ANI-2xt. The ANI-2xt NNP has the same architecture as the ANI-2x but was trained on a tautomer-rich dataset. The conformer duplicates were removed through the OpenBabel[25] root mean standard deviation of atom positions (RMSD).

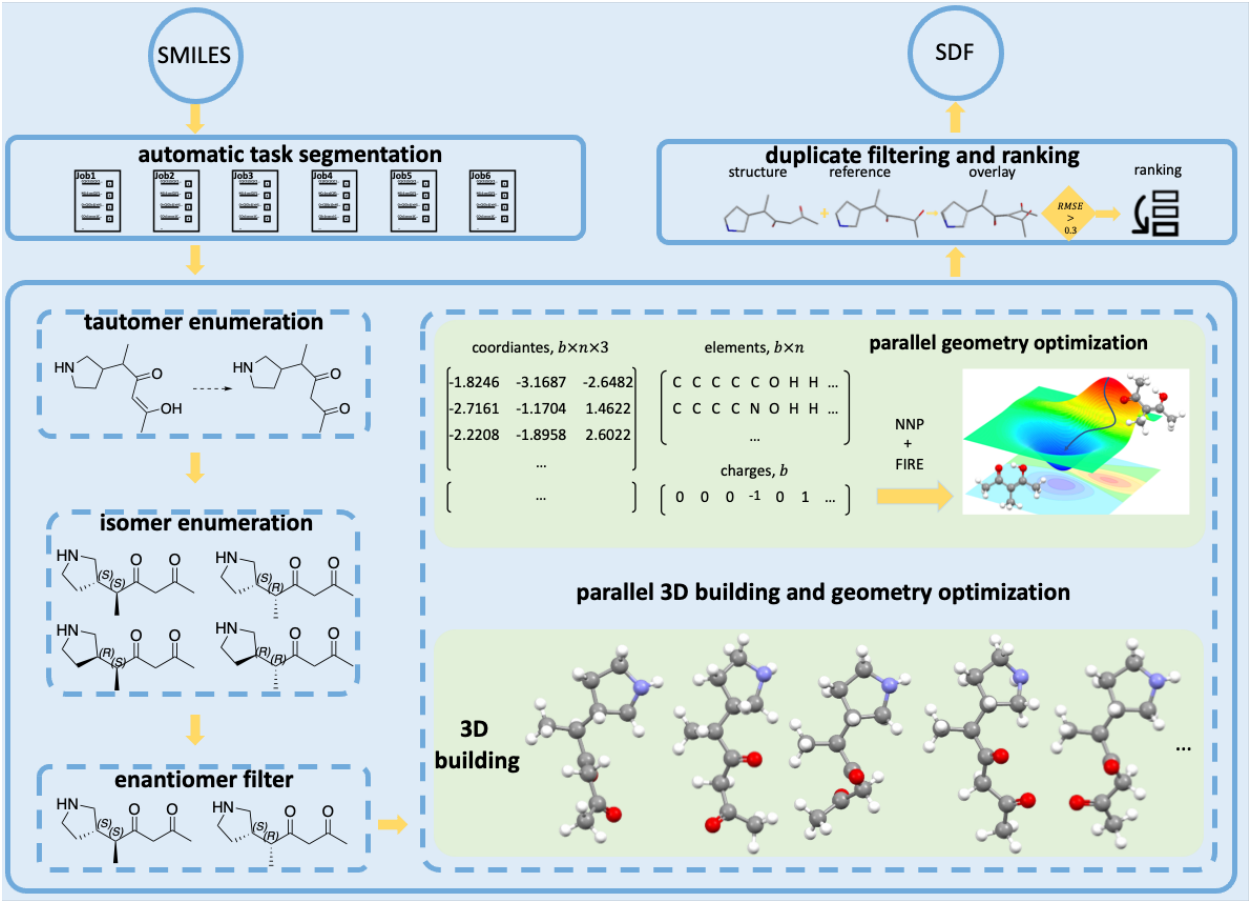


Figure 2. The main components in Auto3D. Tautomer and isomer enumeration are optional steps.

2.1. Auto3D Procedure and Engines

Auto3D contains the following modules: parallel task execution, isomer enumeration (including enantiomer filter), 3D building, geometry optimization, duplicate filtering and ranking (**Fig. 2**). It requires SMILES as the input format and returns the optimized low-energy 3D conformations in the SDF format. Firstly, if the input collection of molecules is large, it is automatically split into several small jobs based on the available CPU/GPU RAM size and core count. Per the most common usage scenario, tautomers are not enumerated in Auto3D's default settings. Then the *isomerization engine* enumerates stereoisomeric absolute configurations for a given molecule depending on the automatically detected number of the chiral centers (including stereocenters in the chain and the ring) and E/Z double bond configurations. The next step is to check the enantiomers. The *enantiomer filter engine* randomly removes inverted configurations and returns the diastereomers. Enantiomers have the same energy and are filtered to reduce the

time for the subsequential optimization step. The next essential step is to generate 3D conformers for every configuration. The *3D building engine* builds a combined pool of conformational ensembles by calling Omega or RDKit. Auto3D runs geometry optimization for the conformers with a *3D optimization engine* that can optimize thousands of conformers at the same time. Three NNPs (ANI2x, ANI2xt and AIMNet) are available for the geometry optimization engine, which finds the geometry that corresponds to a stationary point on the potential energy surface by computing the energy and analytic forces at each optimization step.

It should be noted that ANI-2x and ANI-2xt NNPs are currently parametrized for neutral organic molecules with seven elements: H, C, N, O, F, Cl, and S, while AIMNet can handle both neutral and charged molecules with most non-metal elements (**Table S1**). The user should specify an appropriate NNP for Auto3D based on their input elements and charges. In the final step, optimized conformers that correspond to the same input SMILES are grouped. They are then ranked by energies and filtered to remove duplicates. Top-k (k is a parameter from the user) conformer will be selected for each SMILES. All low-energy structures are combined and saved into a single SDF file.

2.2. Stereoisomer Enumeration and Filtering

The standard SMILES format encodes as a line of text both the connection table and the stereochemistry of a molecule.[26] While Weininger[27] did publish a canonicalization procedure (CANGEN) for SMILES, many data providers and data depositors ignore stereochemistry.[28] To showcase the magnitude of the problem, Table 1 shows the stereochemistry uncertainties in several databases.

Table 1. Percentage of unspecified atom stereochemistry in representative molecular libraries (the duplicates are preliminarily removed from the analyzed datasets).

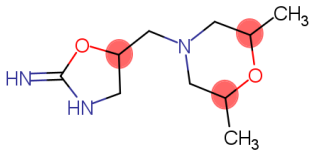
Molecular Database	Entry Molecules	Chiral molecules	Chiral molecules with unspecified atom chirality
Tautomer Database[22]	1773	17%	77%
DrugBank[29]	7633	58%	19%
ChEMBL[30]	2.1M	44%	52.5%

Even a highly curated database such as DrugBank has 19% of chiral molecules missing the stereocenter specification. In the Nicklaus tautomer database[21,22], 77% of chiral molecules contain

unspecified atomic chirality. Among the molecules with two and more atomic stereocenters, 12%, 8%, and 2% of them have two, three, and four unspecified atomic stereocenters, respectively.

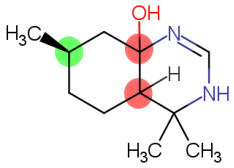
In Figure 3 we have collected some examples to show the unspecified, partially specified, and specified atom chirality (R/S) and bond stereochemistry (E/Z) in SMILES.

A. Unspecified R/S



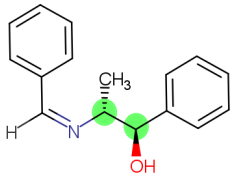
N=C1OC(CN2CC(C)OC(C)C2)CN1
Unspecified stereo units = 3
Total stereoisomers =8

B. Partially specified R/S



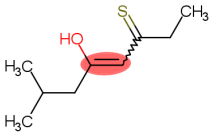
[H]C12CC[C@@H](C)CC1(O)N=CNC2(C)C
Unspecified stereo units = 2
Total stereoisomers =8

C. Specified R/S



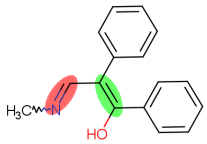
O[C@H](C1=CC=CC=C1)[C@@H](C)/N=C([H])\C2=CC=CC=C2
Unspecified stereo units = 0
Total stereoisomers =4

D. Unspecified E/Z



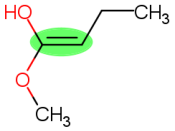
CCC(C=C(CC(C)C)O)=S
Unspecified stereo units = 1
Total stereoisomers =2

E. Partially specified E/Z



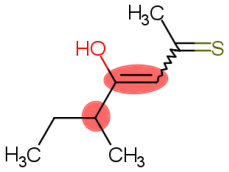
CN=C\C(=C(/O)C1=CC=CC=C1)C2=CC=CC=C2
Unspecified stereo units = 1
Total stereoisomers =4

F. Specified E/Z



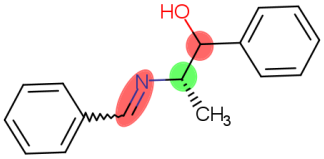
CC/C=C(OC)\O
Unspecified stereo units = 0
Total stereoisomers =2

G. Unspecified units



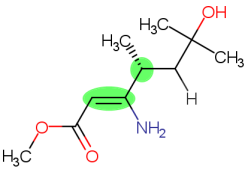
C(C=C(C(C)CC)O)(C)=S
Unspecified stereo units = 2
Total stereoisomers =4

H. Partially specified units



C1=CC(=CC=C1)C(O)[C@H](C)N=CC2=CC=C(C=C2)
Unspecified stereo units = 2
Total stereoisomers =8

I. Specified units



OC(C)(C)C([H])[C@@H](C)/C(N)=C/C(OC)=O
Unspecified stereo units = 0
Total stereoisomers =4

Figure 3. Examples of molecules from the Tautomer Database belonging to subsets with different levels of atom and bond stereochemistry uncertainties.

If the input SMILES defines all stereo information, Auto3D obeys these restrictions during the 3D building process. Otherwise, it enumerates all possible unspecified tetrahedral and double bond stereochemistry with the aid of the *Flipper* utility program in Omega[13] or RDKit[17], producing

a pool of *absolute configurations* of stereoisomers. During this process, specified stereocenters are preserved. The subsequent enantiomer filter removes one of the enantiomer pairs, returning the space of *relative configurations*. Thus, the final pool of stereoisomers contains only diastereomers that are not mirror images of each other. Figure 4 shows the strategy for enumerating stereoisomers and removing enantiomers for the molecule with two stereocenters (as an example) and different levels of uncertainties in molecules.

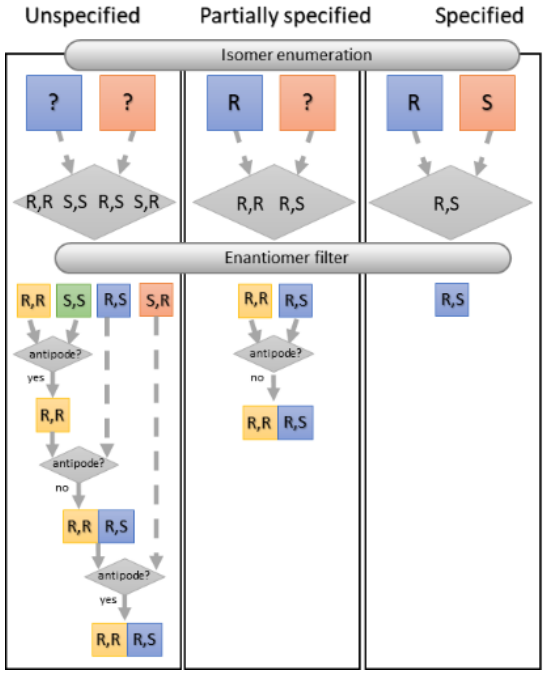


Figure 4. Strategy for generating stereoisomers based on stereochemistry uncertainties in molecular structure: unspecified, partially specified, and specified stereocenters; removing enantiomer entries of the same molecule based on stereochemistry flags: R (rectus) or S (sinister).

2.3. Generation of the 3D conformation

Auto3D by default runs the RDKit conformer generator for each configuration, which is based on the experimental-torsion-knowledge distance geometry (ETKDG)[17] method. The ETKDG was designed to reproduce torsional distributions in crystal conformations from Cambridge Structure Database (CSD) and protein−ligand complexes from Protein Data Bank (PDB). Auto3D by default generates up to 1000 conformations per configuration, which are then filtered using a 0.3 Å RMSD constraint to increase the structural diversity. Alternatively, Auto3D runs OpenEye's conformer generator, Omega. It is a knowledge-based approach that has been

broadly discussed elsewhere[31] and its performance has been widely validated and compared to other programs.[9,32–34] The algorithm is based on carefully curated torsion templates (torsion rules), rule-based generators, and knowledge of rigid 3D fragments. Molecular fragments are assembled with subsequent stability score ranking. Two Omega methods for conformer generation are available: *torsion driving* and *distance geometry*. The distance geometry method (*macrocycle mode*) works for all molecules but has been designed for molecules that contain flexible rings with more than eight atoms, while the torsion driving method (*classic mode*) is designed for molecules without macrocycles.

2.4. Geometry Optimization and Low-Energy Conformation Search

The development of reliable NNPs has seen tremendous progress[19,35]. Some methods achieved DFT accuracy at force field computational cost and showed promising application in molecular energy predictions and geometry optimizations[19]. In Auto3D we have implemented ANI-2x[24], AIMNet[36] and ANI-2xt (see next section) for the accurate and fast configurational space scan. The ANI-2x NNP was trained to mimic the energy and force output from the ωB97X/6-31G* method, while AIMNet and ANI-2xt were trained to mimic the energy and force output from the B97-3c method.

Auto3D features a batch-optimizer that can optimize thousands of 3D structures simultaneously. The optimizer supports all three NNPs. For this, we implemented a Python-based Fast Inertial Relaxation Engine (FIRE[37]) method for geometry optimization. The forces are calculated by taking the derivative of energy with respect to coordinates. This process is based on PyTorch's automatic differentiation engine (Autograd). Autograd requires inputs to have the same dimensions. But molecules naturally contain different numbers of atoms, making the matrices different in size. To tackle this, we padded short molecules with dummy atoms so that all molecules can be processed in batches. According to our benchmark (**Table S2**), it took just one second to optimize ~ 16,000 conformers for one step on an Nvidia RTX 3090 GPU for molecules containing up to 100 heavy atoms. Optimized conformers are then ranked by their energies.

2.5. ANI-2xt Dataset Compilation and Training

This work is focused on tautomer-related predictions. Meanwhile, the accuracy of an NNP critically depends on the size and molecule types of its training dataset[38]. The initial ANI models

were trained to predict the energies of general organic molecules, without an emphasis on tautomerism. Since the size of the conformational space scales exponentially with the molecule size, the first generation of ANI (ANI-1x[38]) was limited to the dataset of organic molecules with only four atom types: H, C, O and N; the second generation of ANI (ANI-2x[24]) was trained with three additional chemical elements: F, S and Cl, which altogether make up ~90% of drug-like molecules.

The new ANI-2xt potential was specifically trained to bring better performance for tautomer-related tasks. In addition to the original ANI-1x and ANI-2x training datasets, ANI-2xt was augmented with a dominant tautomer dataset generated with ChemAxon.[39]. The data preparation protocol is similar to the preparation of the ANI-2x dataset[24]. Specifically, this additional dataset contains a diverse set of bioactive molecules with up to 20 non-H atoms (~82k molecules from ChEMBL[30,40]). We used ChemAxon Tautomer Generation Plugin to enumerate potential tautomeric forms. This database (like other public libraries) covers only configurational space but not conformational space of the molecules, while the latter is a critical requirement for the NNP training procedure. Therefore, we carried out the non-equilibrium conformation generation process using GFN2-XTB[41] molecular dynamics at 400K for 20ps. The optimized structures were selected for energy calculations with the B97-3c composite scheme[42] in ORCA[43,44]. This created 3.4M new training data points. Combined with the ANI-1x and ANI-2x datasets, the final ANI-2xt training data includes ~13 million non-equilibrium molecular conformations. The train/test split was 90/10 (**Fig. S1**). ANI-2xt implementation is similar to ANI-2x, which has been described elsewhere[45].

2.6. Tautomerization Dataset and Tautomerization Energies

1412 tautomer pairs have been selected from the Tautomer Database[21,22]. It consists of different types of tautomer reactions, number of atoms and functional groups. We consider the tautomeric interconversion as a chemical reaction of the type R⇌P, where R represents the ground state of the reactant and P represents the ground state of the tautomeric product. The prototropic tautomerism tested in this paper includes the following R/P reactions: keto / enol, amide/ imidol, thioamide / thioimidol, thioketo / enethiol, lactam / lactim, oxo-imine / enol-imine, thioketo-enol / keto-enethiol, paraquinonoid / orthoquinonoid, and nitrile / keteneimine (**Fig. S2**). Our first goal

is to validate NNPs in Auto3D by comparing the tautomeric reaction energy ΔE_{taut}^{el} (Eq.1) from DFT and ANI with the "gold-standard" coupled cluster calculations.

$$\Delta E_{taut}^{el} = E_{sp}^{el}(P) - E_{sp}^{el}(R)\ . \quad (1)$$

For this goal, we used Auto3D (isomerization engine was Omega, and optimization engine was ANI-2x) to find the lowest energy conformers for the (R/P) tautomers. These structures are further optimized with ωB97M-V/def2-TZVP[46] using the ORCA 4.1 package to obtain the best possible geometry. Then we performed single-point calculations in DFT, ANI and DLPNO-CCSD(T)/CBS[47,48] methods. Specifically, we used the ωB97 family of functionals (ωB97X/6-31G*[49], ωB97M-V/def2-TZVP[46]), B97-3c[42] and ANI family of NNPs (ANI-2x, ANI-2xt). The coupled cluster was used as a very accurate reference for DFT and ANI methods. A random subset of 370 tautomeric reactions was selected for DLPNO-CCSD(T) coupled cluster calculations due to the high computational cost. To assess the method's accuracy, the tautomeric reaction energy is compared between different methods.

The thermodynamical calculations have been performed for tautomeric interconversions R⇌P in the gas phase. We used the rigid-rotor harmonic oscillator (RRHO) approach in quantum chemical calculations. The tautomerization Gibbs free energy difference between tautomers is defined as:

$$\Delta G_{taut}^{gas} = \Delta E_{taut}^{el} + \Delta E(ZPE) + \Delta E^{t} - T\Delta S, \quad (2)$$

where the difference in total energy from the electronic structure calculation ΔE_{taut}^{el} is obtained at the same level method as zero-temperature vibrational energy $\Delta E(ZPE)$, thermal corrections ΔE^{t} and entropy contributions $T\Delta S$ are from corresponding frequency calculations. Specifically, we used Auto3D (isomerization engine was Omega, and optimization engine was ANI-2xt) to find the lowest energy conformers for the tautomeric pairs (R/P) followed by vibrational frequencies calculation with ANI-2xt and B97-3c, respectively. For ANI-2xt, the thermodynamic functions (Eq.2) were calculated with the ideal gas approximation by the Atomic Simulation Environment (ASE) using ANI-2xt as the calculator. The reference thermodynamics calculations were done with B97-3c composite scheme using ORCA.

3. Results and Discussion

In this section, we systemically assessed how the accounting for the stereoisomeric configurational space affects the low-energy conformer search. We also benchmarked ANI-2xt on

geometry optimization and energy prediction with several DFT and ML methods. In a final case study, we benchmarked Auto3D for Gibbs free energy calculations starting from SMILES.

3.1. Stereoisomer Enumeration

To show that the isomer enumeration step in Auto3D is essential in the process of finding the low-energy conformer, we compared the final conformers from two modes: Auto3D with isomer enumeration and Auto3D without enumeration. The isomerization engine was Omega and the optimization engine was ANI-2xt for both modes. We sampled 50 molecules from the ChEMBL database. Each molecule contains at least 3 unspecified chiral centers (**Fig. S3**). With isomer enumeration, 287,182 conformers were generated for the 50 molecules. In contrast, only 40,212 conformers were generated without isomer enumeration.

With isomer enumeration, almost all final structures have lower energies compared to the output without isomer enumeration (**Fig. 5**). A similar trend was observed when RDKit was used as the isomerization engine (**Fig. S4**). We used molecule CHEMBL1596382 (**Fig. 5** panel B and C) to understand the energy difference between the two modes. This was one of the two extreme examples with the maximal difference between the obtained energies. For CHEMBL1596382, the energy of the conformer with isomer enumeration is 11.7 kcal/mol lower than the conformer from the other mode. The 3 stereocenters are assigned with different chiral symbols, which results in very different 3D structures. With the SSR configuration, the molecule is organized in a way that enables more favorable van der Waals interactions, which makes the conformer more stable compared to other configurations. Very rarely, the conformer whose configuration was randomly assigned has lower energy than the conformer from the isomer enumeration mode (**Fig. 5** Panel D). This is because the random configuration happens to be the same or the enantiomer of the low-energy configuration. During the geometry optimization process, they are optimized into two conformers with slightly different local geometries.

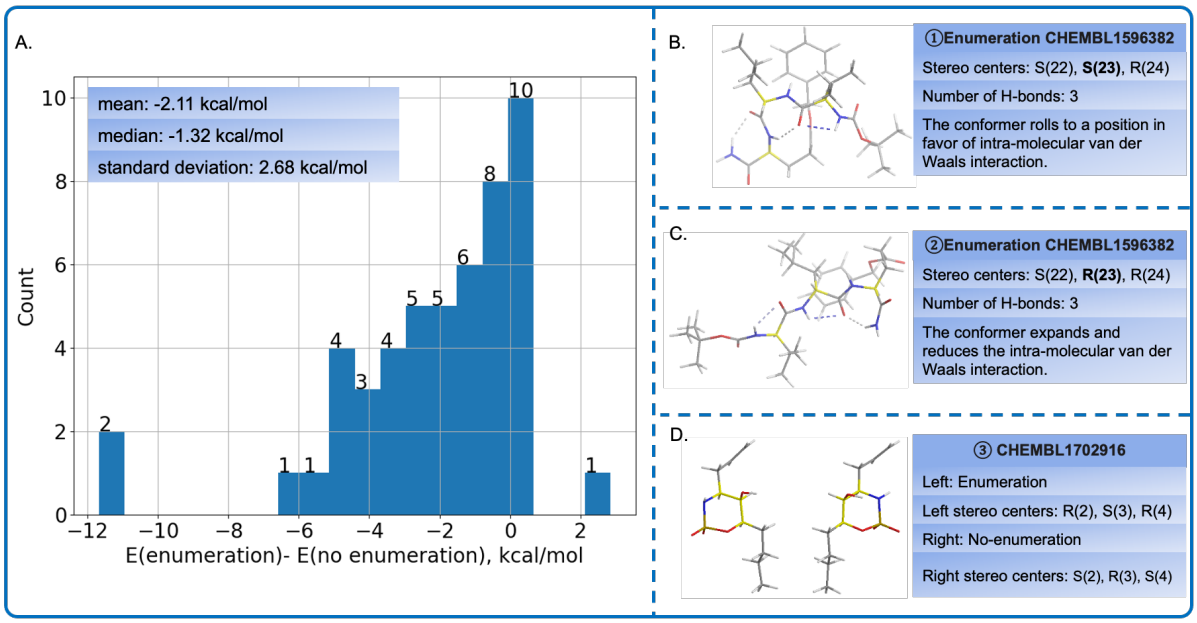


Figure 5. Comparing the lowest energy structure from two Auto3D modes (with isomer enumeration and without isomer enumeration). A: Energy difference distributions for the two modes. Panel B and C are the low-energy conformers for molecule CHEMBL 1596382 from Auto3D with isomer enumeration and without isomer enumeration, respectively. The energy in B is 11.7 kcal/mol lower than that in C. Hydrogen bonds are shown with dashed lines. Panel D is an example where two modes chose a pair of enantiomers as the low-energy structures. They are the right-most data point in the histogram. All stereocenters are in yellow.

3.2. Geometry Optimization

To show the reliability of geometry optimization with ANI-2xt, the geometry optimization results from ANI-2x and ANI-2xt were compared against the ωB97X/6-31G* results. Firstly, we ran geometry optimization for 2810 molecules from the Tautomer database[22] with ωB97X/6-31G*. These structures were then reoptimized by ANI-2x and ANI-2xt, respectively. Then we compared the geometry differences from ANI-2x/ANI-2xt against the DFT reference using three metrics: RMSD, maximum torsion angle difference and mean bond length difference (Please see SI for the definitions). Representative molecular alignments are depicted in panel A. The first pair of structures has an RMSD of 0.20 Å. For the second pair, the RMSD is around the mean value of 0.13 Å, and the two conformers are almost identical.

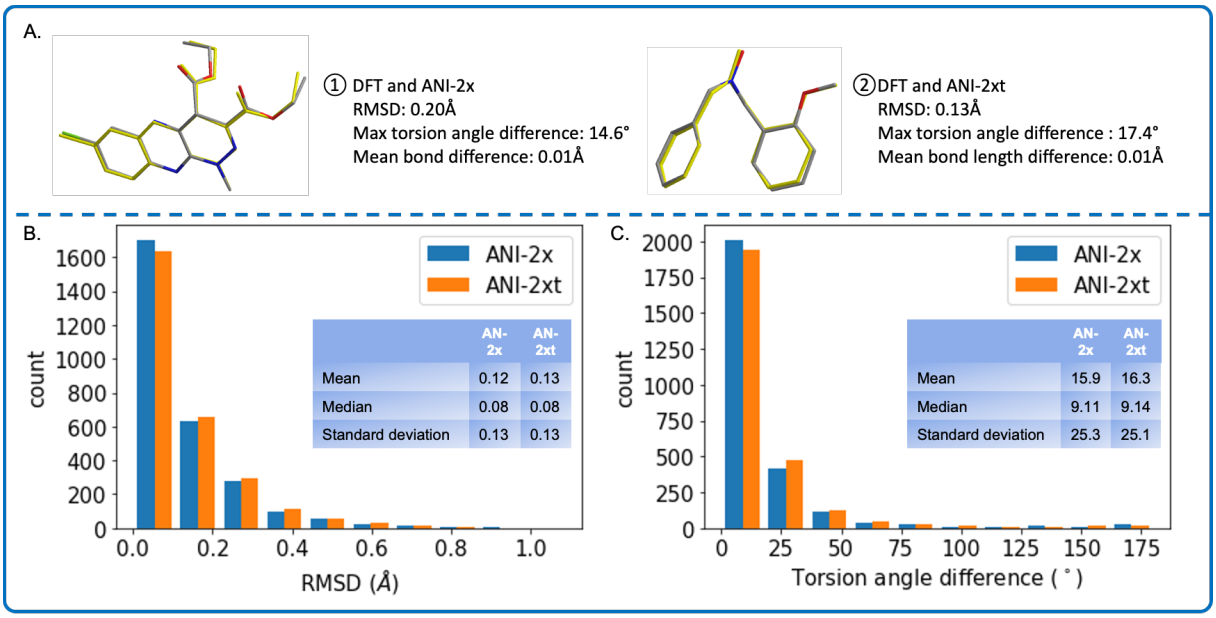


Figure 6. The structural difference between the molecules optimized by ωB97X/6-31G* and ANI-2x/ANI-2xt method. For each pair of molecules in panel A, the yellow structure is from ANI-2x or ANI-2xt and the other is from ωB97X/6-31G*. Panel B shows the distribution of RMSD values from ANI-2x/ANI-2xt against ωB97X/6-31G*. Panel C shows the distribution of maximum torsion angle differences from ANI-2x/ANI-2xt against ωB97X/6-31G*.

Compared with ωB97X/6-31G* geometries, 78.8% of ANI-2xt structures are within the RMSD value of 0.20 Å. The mean and median RMSD values are 0.13 Å and 0.08 Å, respectively. In general, a smaller RMSD value means a better superposition between the two structures. In our observation, an RMSD that is below 0.20 Å usually represents a good alignment. In addition to RMSD metrics, we used maximum torsion angle difference. It can catch the differences in the relative orientation of two groups within a structure. Torsion angles naturally have a larger magnitude compared with other metrics. In our cases, structures within a small difference in the torsion angle (around 15°) usually align well. Values at around 90° or 180° are interesting because this indicates that two groups of atoms are pointing in perpendicular or reversed directions in the two conformers. This usually happens when multiple local minima exist, and the optimization algorithm converged into different local minima during the geometry optimization process. We also used mean bond length differences as the third metric. For both methods, almost all structures have mean bond length differences within 0.02 Å, indicating a high consistency with the reference structure (**Fig. S5**).

As expected, the ANI geometries are very consistent with the ωB97X/6-31G* during the geometry optimization process, although there is a possible space for improvements. The ANI-2xt model gave very similar performance to ANI-2x in terms of geometry optimization, which has been widely adopted in different applications and has shown reliable performance[50–52]. Therefore, ANI-2xt could also be used as a reliable optimizing potential in Auto3D.

3.3. Tautomerization Energy

To validate the accuracy of ANI-2x and ANI-2xt models on the Nicklaus tautomer dataset, both were benchmarked against DFT methods and DLPNO-CCSD(T)/CBS[47,48]. DLPNO-CCSD(T)/CBS is usually considered the gold standard in computational chemistry, so we used it as the ground truth reference. Both QM methods and ANI methods got reasonable predictions for tautomeric reaction energies (**Fig. 7**). Compared with DLPNO-CCSD(T)/CBS results, the ANI-2x method has a mean absolute error (MAE) of 4.32 kcal/mol. This error was reduced by 42% to 2.52 kcal/mol with our ANI-2xt model. This improvement is believed to be due to the substantial number of tautomers in the ANI-2xt training data. The *ωB97*M-V/def2-TZVP method shows excellent performance with respect to DLPNO-CCSD(T)/CBS with an MAE of just 0.64 kcal/mol. It is worth noting that ANI-2xt achieved similar accuracy compared to *ωB97*X/6-31G*, when using DLPNO-CCSD(T)/CBS or *ωB97*M-V/def2-TZVP results as the reference. In addition to accuracy, the ANI-2xt model is about six orders of magnitude faster than the DFT methods.

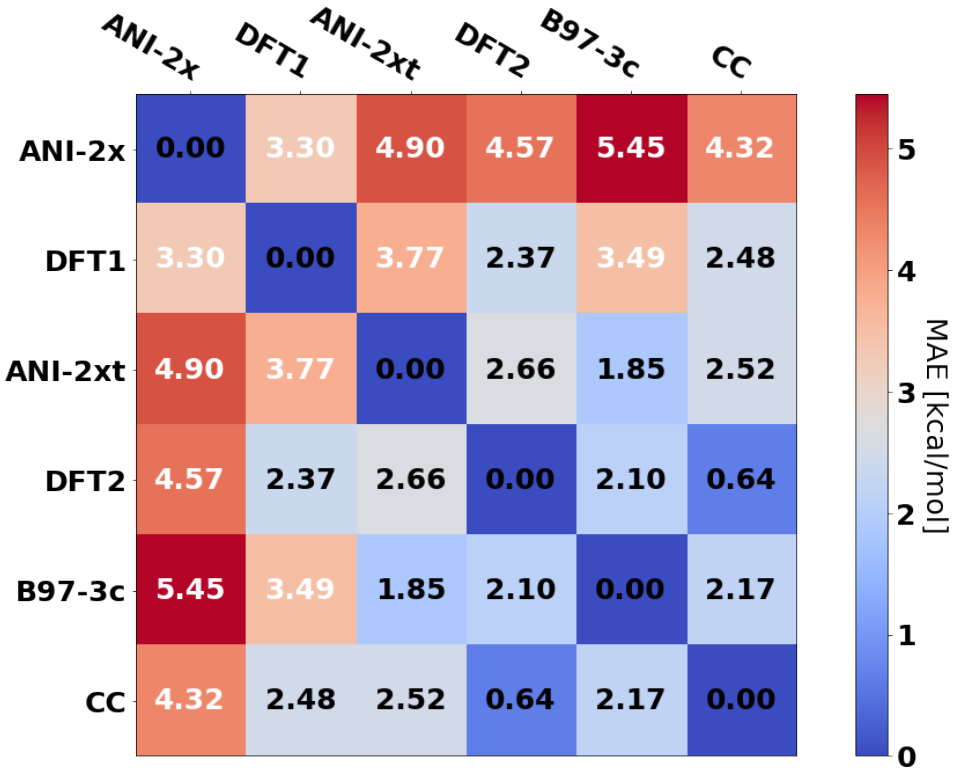


Figure 7. Comparing the tautomeric reaction energies calculated from different methods. DFT1 represents *ωB97*X/6-31G*; DFT2 represents *ωB97*M-V/def2-TZVP; CC represents DLPNO-CCSD(T)/CBS. For ANI methods and DFT methods, 1412 tautomerization reactions were collected. We randomly sampled 370 of the tautomerization reaction for DLPNO-CCSD(T)/CBS calculations.

Recently Chodera and co-workers benchmarked ANI-1x potential to compute tautomer ratios in vacuum and solvent[53]. Despite not being trained on tautomers, the reported RMSE was 1.5 kcal/mol compared with the reference QM data. Similarly, Meuwly[54] and co-workers concluded that the ANI-1x model was the best ML potential among the five benchmarked models on the Tautobase dataset[55] with an RMSE of 2.85 kcal/mol.

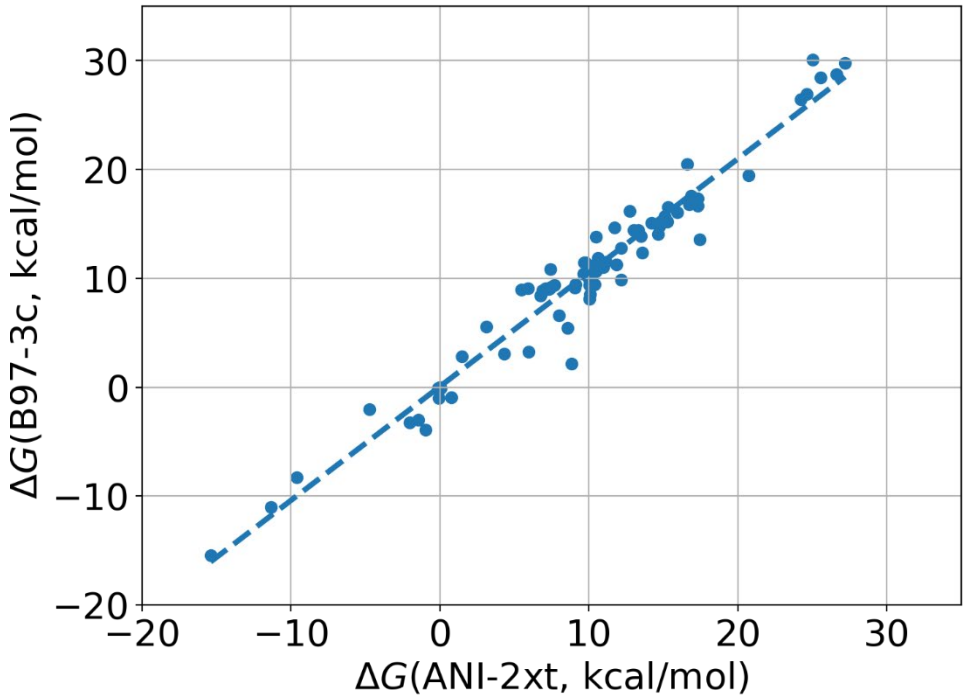


Figure 8. Comparing the tautomerization Gibbs free energy between B97-3c and ANI-2xt.

Here we benchmarked ANI-2xt for a Gibbs free energy calculation task with B97-3c (**Fig. 8**). From the Nicklaus tautomer dataset[22], we randomly sampled gas-phase tautomerization reactions that only involve 2 tautomers and H, C, N, O, F, S, Cl elements. This resulted in 81 tautomerization reactions. Comparing the tautomerization Gibbs free energy from ANI-2xt and B97-3c, the R^2 is 0.95. The RMSE and MAE are 1.96 and 1.48 kcal/mol, respectively. It demonstrates that ANI-2xt is consistent with the DFT methods in terms of electronic and free energies calculations. Still, we need to note that the current ANI-2xt model assumes a vacuum environment, whereas tautomerization is strongly dependent on the solvent. There could be a gap between the theoretical and real-world results. We are working on continuum solvent models to take account of the solvent effects.

4. Conclusions

Generating 3D structures of flexible small organic molecules is not a trivial task[56,57]. First of all, it requires exploring the space of configurational stereoisomers due to the possible

uncertainties in chemical databases. Additionally, it requires the evaluation of a conformational ensemble that covers an enormous conformer space to find the low-energy conformers[17]. Therefore, we developed the open-source Auto3D package for generating low-energy 3D structures from SMILES. Auto3D is independent of operating systems and accelerated through multicore and GPU calculations capabilities. It integrates state-of-the-art isomer generation programs and reliable neural network potentials. Auto3D automatizes the isomer enumeration and duplicates filtering process, 3D building process, geometry optimizing and ranking process.

The isomerization engine and optimization engine in Auto3D were tested to be reliable during three benchmarking studies. By enumerating unspecified chiral centers, Auto3D finds the configuration that gives the lowest energy. In conjunction with Auto3D, we also developed ANI-2xt, an incremental improvement to the popular ANI-2x NNP. During a geometry optimization benchmarking test, ANI-2xt showed high consistency to the ωB97X/6-31G* method. We used Auto3D to calculate tautomerization Gibbs free energy change from scratch, and the results are within 2 kcal/mol from the target DFT accuracy.

There are some limitations and potential risks that require future work. Firstly, some chiral molecules are used in the racemic form or contain underdefined stereo centers in real applications, so assigning specific stereo information could result in over-curation. The experimental results using an underdefined chiral molecule might be different from that of a fully defined low-energy stereoisomer. Secondly, the current NNPs of Auto3D optimize geometries in a vacuum condition. We need to include continuum solvent models in the future.

Supplementary Information

CHEMBL.smi contains 50 molecules that were used for validating the isomer enumeration step.
Geometry.smi contains 2810 molecules that were used for benchmarking geometry optimization.
tautomerization_E.smi contains 2824 molecules that were used to calculate tautomeric reaction energies.
tautomerization_G.smi contains 162 molecules that were used to calculate tautomerization Gibbs free energies.

Acknowledgments

The authors acknowledge Dr. Roman Zubatiuk for helping with the training dataset for ANI-2xt and the batch optimizer in Auto3D. The work performed by Z.L., T.Z., and O.I. (PI) was made possible by the Office of Naval Research (ONR) through support provided by the Energetic Materials Program (MURI grant no. N00014-21-1-2476). A.R. acknowledges National Science Foundation (NSF) CHE-1802831 award. We also acknowledge the Extreme Science and Engineering Discovery Environment (XSEDE) award CHE200122, which is supported by NSF grant number ACI-1053575. We gratefully acknowledge the support and hardware donation from NVIDIA Corporation and express our special gratitude to Jonathan Lefman. We also would like to thank the Armed Forces of Ukraine and dedicate this paper to all brave defenders of Ukraine against Russian Invasion.

Data and Software Availability

All data, code, user documentation and examples can be found on our GitHub page at https://github.com/isayevlab/Auto3D_pkg and supplementary information to this paper.

References

(1) CAS Registry https://www.cas.org/cas-data/cas-registry (accessed May 9, 2022).

(2) InChI https://iupac.org/who-we-are/divisions/division-details/inchi/ (accessed May 9, 2022).

(3) UNICHEM https://www.unichem.com/ (accessed May 9, 2022).

(4) PubChem https://pubchem.ncbi.nlm.nih.gov/ (accessed May 9, 2022).

(5) Fourches, D.; Muratov, E.; Tropsha, A. Trust, but Verify: On the Importance of Chemical Structure Curation in Cheminformatics and QSAR Modeling Research. *J. Chem. Inf. Model.* **2010**, *50*, 1189–1204.

(6) Fourches, D.; Muratov, E.; Tropsha, A. Trust, but Verify II: A Practical Guide to Chemogenomics Data Curation. *J. Chem. Inf. Model.* **2016**, *56*, 1243–1252.

(7) Young, D.; Martin, T.; Venkatapathy, R.; Harten, P. Are the Chemical Structures in Your QSAR Correct? *QSAR Comb. Sci.* **2008**, *27*, 1337–1345.

(8) Graham, S. T. W.; Fryhle, C. B.; Snyder, S. A. *Organic Chemistry*, 12th ed.; John Wiley & Sons Incorporated, 2017.

(9) Friedrich, N.-O.; De, C.; Kops, B.; Flachsenberg, F.; Sommer, K.; Rarey, M.; Kirchmair,

J. Benchmarking Commercial Conformer Ensemble Generators. *J. Chem. Inf. Model.* **2017**, *57*, 2719–2728.

(10) Brameld, K. A.; Kuhn, B.; Reuter, D. C.; Stahl, M. Small Molecule Conformational Preferences Derived from Crystal Structure Data. A Medicinal Chemistry Focused Analysis. *J. Chem. Inf. Model.* **2008**, *48*, 1–24.

(11) Kirchmair, J.; Wolber, G.; Laggner, C.; Langer, T. Comparative Performance Assessment of the Conformational Model Generators Omega and Catalyst: A Large-Scale Survey on the Retrieval of Protein-Bound Ligand Conformations. *J. Chem. Inf. Model.* **2006**, *46*, 1848–1861.

(12) Boström, J.; Greenwood, J. R.; Gottfries, J. Assessing the Performance of OMEGA with Respect to Retrieving Bioactive Conformations. *J. Mol. Graph. Model.* **2003**, *21*, 449–462.

(13) OMEGA 4.1.1.1: OpenEye Scientific Software https://docs.eyesopen.com/toolkits/python/omegatk/index.html (accessed June 6, 2021).

(14) Open Eye Scientific Software. http://www.eyesopen.com/ (accessed June 6, 2021).

(15) Kirchmair, J.; Laggner, C.; Wolber, G.; Langer, T. Comparative Analysis of Protein-Bound Ligand Conformations with Respect to Catalyst's Conformational Space Subsampling Algorithms. *J. Chem. Inf. Model.* **2005**, *45*, 422–430.

(16) Pracht, P.; Bohle, F.; Grimme, S. Automated Exploration of the Low-Energy Chemical Space with Fast Quantum Chemical Methods. *Phys. Chem. Chem. Phys.* **2020**, *22*, 7169–7192.

(17) Riniker, S.; Landrum, G. A. Better Informed Distance Geometry: Using What We Know To Improve Conformation Generation. *J. Chem. Inf. Model.* **2015**, *55*, 2562–2574.

(18) Molsoft. http://www.molsoft.com/2dto3d.html (accessed July 8, 2021).

(19) Zubatiuk, T.; Isayev, O. Development of Multimodal Machine Learning Potentials: Toward a Physics-Aware Artificial Intelligence. *Acc. Chem. Res.* **2021**, *54*, 1575–1585.

(20) Zubatyuk, R.; Smith, J. S.; Leszczynski, J.; Isayev, O. Accurate and Transferable Multitask Prediction of Chemical Properties with an Atoms-in-Molecules Neural Network. *Sci. Adv.* **2019**, *5*, eaav6490.

(21) Tautomer Structures Extracted from Experimental Literature https://cactus.nci.nih.gov/download/tautomer/ (accessed Nov 18, 2020).

(22) Dhaked, D. K.; Guasch, L.; Nicklaus, M. C. Tautomer Database: A Comprehensive Resource for Tautomerism Analyses. *J. Chem. Inf. Model.* **2020**, *60*, 1090–1100.

(23) Quacpac ToolKit https://docs.eyesopen.com/toolkits/python/quacpactk/index.html (accessed June 6, 2021)

(24) 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. *J. Chem. Theory Comput.* **2020**, *16*, 4192–4202.

(25) O'Boyle, N. M.; Banck, M.; James, C. A.; Morley, C.; Vandermeersch, T.; Hutchison, G. R. Open Babel: An open chemical toolbox. *J. Cheminf.* **2011**, *3*, 1-14.

(26) Warr, W. A. Representation of Chemical Structures. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **2011**, *1*, 557–579.

(27) Weininger, D.; Weininger, A.; Weininger, J. L. SMILES. 2. Algorithm for Generation of Unique SMILES Notation. *J. Chem. Inf. Comput. Sci.* **1989**, *29*, 97–101.

(28) O'Boyle, N. M. Towards a Universal SMILES Representation - A Standard Method to Generate Canonical SMILES Based on the InChI. *J. Cheminf.* **2012**, *4*, 22.

(29) Wishart, D. S.; Feunang, Y. D.; Guo, A. C.; Lo, E. J.; Marcu, A.; Grant, J. R.; Sajed, T.; Johnson, D.; Li, C.; Sayeeda, Z.; Assempour, N.; Iynkkaran, I.; Liu, Y.; Maciejewski, A.; Gale, N.; Wilson, A.; Chin, L.; Cummings, R.; Le, D.; Pon, A.; Knox, C.; Wilson, M. DrugBank 5.0: A Major Update to the DrugBank Database for 2018. *Nucleic Acids Res.* **2018**, *46*, D1074–D1082.

(30) Mendez, D.; Gaulton, A.; Bento, A. P.; Chambers, J.; De Veij, M.; Félix, E.; Magariños, M. P.; Mosquera, J. F.; Mutowo, P.; Nowotka, M.; Gordillo-Marañón, M.; Hunter, F.; Junco, L.; Mugumbate, G.; Rodriguez-Lopez, M.; Atkinson, F.; Bosc, N.; Radoux, C. J.; Segura-Cabrera, A.; Hersey, A.; Leach, A. R. ChEMBL: Towards Direct Deposition of Bioassay Data. *Nucleic Acids Res.* **2019**, *47*, D930–D940.

(31) 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. *J. Chem. Inf. Model.* **2010**, *50*, 572–584.

(32) Poli, G.; Seidel, T.; Langer, T. Conformational Sampling of Small Molecules With ICon: Performance Assessment in Comparison With OMEGA. *Front. Chem.* **2018**, *6*.

(33) Jonas Boström. Reproducing the Conformations of Protein-Bound Ligands: A Critical Evaluation of Several Popular Conformational Searching Tools. *J. Comput.Aided Mol. Des.* **2001**, *15*, 1137–1152.

(34) Hawkins, P. C. D.; Nicholls, A. Conformer Generation with OMEGA: Learning from the Data Set and the Analysis of Failures. *J. Chem. Inf. Model.* **2012**, *52*, 2919–2936.

(35) Ko, T. W.; Finkler, J. A.; Goedecker, S.; Behler, J. General-Purpose Machine Learning Potentials Capturing Nonlocal Charge Transfer. *Acc. Chem. Res.* **2021**, *54*, 808–817.

(36) 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.* **2021**, *12*, 1–11.

(37) Bitzek, E.; Koskinen, P.; Gähler, F.; Moseler, M.; Gumbsch, P. Structural Relaxation Made Simple. *Phy. Rev. Lett.* **2006**, *97*, 1-4.

(38) Smith, J. S.; Isayev, O.; Roitberg, A. E. ANI-1: An Extensible Neural Network Potential with DFT Accuracy at Force Field Computational Cost. *Chem. Sci.* **2017**, *8*, 3192–3203.

(39) CXCALC 5.7.1.:ChemAxon Software. https://www.chemaxon.com (accessed June 30, 2020).

(40) Davies, M.; Nowotka, M.; Papadatos, G.; Dedman, N.; Gaulton, A.; Atkinson, F.; Bellis, L.; Overington, J. P. ChEMBL Web Services: Streamlining Access to Drug Discovery Data and Utilities. *Nucleic Acids Res.* **2015**, *43*, W612–W620.

(41) Bannwarth, C.; Ehlert, S.; Grimme, S. GFN2-XTB—An Accurate and Broadly Parametrized Self-Consistent Tight-Binding Quantum Chemical Method with Multipole Electrostatics and Density-Dependent Dispersion Contributions. *J. Chem. Theory Comput.* **2019**, *15*, 1652–1671.

(42) Brandenburg, J. G.; Bannwarth, C.; Hansen, A.; Grimme, S. B97-3c: A Revised Low-Cost Variant of the B97-D Density Functional Method. *J. Chem. Phys.* **2018**, *148*, 064104.

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

(44) Neese, F. Software Update: The ORCA Program System, Version 4.0. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **2017**, *8*, e1327.

(45) Gao, X.; Ramezanghorbani, F.; Isayev, O.; Smith, J. S.; Roitberg, A. E. TorchANI: A Free and Open Source PyTorch-Based Deep Learning Implementation of the ANI Neural Network Potentials. *J. Chem. Inf. Model.* **2020**, *60*, 3408–3415.

(46) Mardirossian, N.; Head-Gordon, M. ΩB97M-V: A Combinatorially Optimized, Range-Separated Hybrid, Meta-GGA Density Functional with VV10 Nonlocal Correlation. *J. Chem. Phys.* **2016**, *144*, 214110.

(47) Riplinger, C.; Neese, F. An Efficient and near Linear Scaling Pair Natural Orbital Based Local Coupled Cluster Method. *J. Chem. Phys.* **2013**, *138*, 1–18.

(48) 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. *Nat. Commun.* **2019**, *10*, 1–8.

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

(50) Kulichenko, M.; Smith, J. S.; Nebgen, B.; Li, Y. W.; Fedik, N.; Boldyrev, A. I.; Lubbers, N.; Barros, K.; Tretiak, S. The Rise of Neural Networks for Materials and Chemical Dynamics. *J. Phys. Chem. Lett.* **2021**, *12*, 6227–6243.

(51) Gupta, A.; Zhou, H. X. Machine Learning-Enabled Pipeline for Large-Scale Virtual Drug Screening. *J. Chem. Inf. Model.* **2021**, *61*, 4236–4244.

(52) Rosenberger, D.; Smith, J. S.; Garcia, A. E. Modeling of Peptides with Classical and Novel Machine Learning Force Fields: A Comparison. *J. Phys. Chem. B* **2021**, *125*, 3598–3612.

(53) Wieder, M.; Fass, J.; Chodera, J. D. Fitting Quantum Machine Learning Potentials to Experimental Free Energy Data: Predicting Tautomer Ratios in Solution. *Chem. Sci.* **2021**, *12*, 11364–11381.

(54) Vazquez-Salazar, L. I.; Boittier, E. D.; Unke, O. T.; Meuwly, M. Impact of the Characteristics of Quantum Chemical Databases on Machine Learning Prediction of Tautomerization Energies. *J. Chem. Theory Comput.* **2021**, *17*, 4769–4785.

(55) Wahl, O.; Sander, T. Tautobase: An Open Tautomer Database. *J.Chem. Inf. Model*. **2020**, *60*, 1085–1089.

(56) Sadowski, J.; Gasteiger, J. From Atoms and Bonds to Three-Dimensional Atomic Coordinates: Automatic Model Builders. *Chem. Rev.* **1993**, *93*, 2567–2581.

(57) Lagorce, D.; Pencheva, T.; Villoutreix, B. O.; Miteva, M. A. DG-AMMOS: A New Tool to Generate 3D Conformation of Small Molecules Using D Istance G Eometry and A

Utomated M Olecular M Echanics O Ptimization for in Silico S Creening. *BMC Chem. Biol.* **2009**, *9*.