# High Order Ab Initio Valence Force Field with Chemical Pattern Based Parameter Assignment

*Xudong Yang, Chengwen Liu, Pengyu Ren**

Department of Biomedical Engineering, The University of Texas at Austin, Austin, TX 78712, USA

* Corresponding author. Email: pren@utexas.edu

# Abstract

Bonded (or valence) interactions, which directly determine the local structures of the molecules, are fundamental parts of molecular mechanics force fields (FFs). Most popular classical FFs adopt the simple harmonic models for bond stretching and angle bending and ignore cross-coupling effects among the valence terms. This may lead to less accurate vibrational properties and configurations in molecular dynamics (MD) simulations. AMOEBA models utilize an MM3(MM4)-style bonded interaction model, in which the vibrational anharmonicity, the coupling effects among different energy terms, and the out-of-plane bending for *sp2*-hybridized atoms are considered. In this work, we report the development of bonded interaction parameters for a wide range of chemistry based on quantum mechanics (QM). About 270 atomic types defined by SMARTS strings were used to model the valence interactions. Our results indicate that the resulting valence parameters produce accurate vibrational frequencies (RMSD from QM is less than ~36.6 $cm^{-1}$) over a large set of molecules with diverse functional groups (445 molecules). By contrast, the harmonic models usually give an RMS error greater than 60 $cm^{-1}$. Meanwhile, this model accurately reflects the potential energy surface of the out-of-plane bending. Our model can generally be applied to the AMOEBA family and any MM3(MM4)-based molecular mechanics FFs.



**KEYWORDS:** AMOEBA potential, Force field, Bonded interactions, Normal mode frequencies

# 1. Introduction

Classical molecular dynamics (MD) simulations are widely applied by researchers to investigate various biomolecular systems. One of the essential elements determining the reliability of MD simulations is the underlying classical potentials, or force fields (FFs)[1-3], which describe the intra- and intermolecular interactions using empirical functional forms. Abundant studies focus on improving modeling the intermolecular (non-bonded) interactions as they are responsible for thermodynamic properties in protein-ligand binding[4-7], protein folding[8], enzymatic catalysis[9], and other critical biomolecular processes. Nonetheless, bonded interactions are equally important as they determine the local structural ensemble sampled during the MD simulations.[10]

Most popular FFs, such as AMBER[11-14], OPLS[15-17], and CHARMM[18-20], adopt simple harmonic functions to describe bond stretching and angle bending without considering the anharmonic nature of bonds and angles. AMBER omits cross-coupling effects between bond and angle terms, and it uses the Fourier style to describe out-of-plane effects. OPLS uses a similar functional form to AMBER. CHARMM employs a unique improper potential to represent the out-of-plane bending and a Urey-Bradley potential to account for the cross-coupling. Those methods are computationally efficient, but their accuracy is limited in complex molecular structures, especially in vibrational analysis. By contrast, AMOEBA-based models[21-23] use the sophisticated MM3(MM4)-style[24-25] functional forms to describe these bonded interactions, in which the anharmonicity, cross-coupling effects, and out-of-plane bending (opbend) have been included. First, higher-order polynomial terms were added to the bond stretching and angle bending terms to capture the anharmonicity while avoiding abnormal effects on strained bonds or angles due to the negative contributions from odd polynomial terms. Second, the cross-coupling terms were proven to be essential for capturing the coupling effects and improving the vibrational frequencies. For example, it is evidenced that the combination of a high-order polynomial model and cross-terms can greatly enhance the description of equilibrium structure and vibrational frequencies of formate ion.[26-27] Specifically, the average absolute deviation of vibrational frequencies ($|\Delta\nu|_{ave}$) plunges from 155.6 $cm^{-1}$ by the simple

harmonic model without any cross terms to 7.2 $cm^{-1}$ by a quartic model with third-order cross terms. Similar improvements have been found on systems including amides, peptides, and others.[28-29] Finally, the out-of-plane bending (opbend)[30] is a necessary and important energy term to maintain planar structure around trigonal centers as a significant pyramidalization can form without seriously distorting any of the three angles of bonds involved.

MM3/MM4 parameters were derived almost three decades ago. Quantum mechanics (QM) calculations have significantly improved over the years, with energy, structures, vibrational frequencies closely matching experimental measurements. In addition, a more comprehensive range of coverage in chemical space is needed, and better ways for parameter assignment and typing (e.g., chemical pattern based) are readily available. This work aims to provide an up-to-date database of valence parameters for common organic molecules using atomic types determined by SMARTS strings. These parameters will be used for AMOEBA[21] and AMOEBA+[22] FFs but are also compatible with other classical FFs that use MM3/MM4 style valence terms. We have developed an automated Python program that implements the Modified Seminario method[31] to generate bond and angle parameters from ab initio Hessian matrices. In addition, a reversed searching with the ranking tree was implemented inside this program to assign valence parameters for new molecules based on SMARTS patterns. Moreover, the anharmonicity, coupling effects, and out-of-plane bending have been analyzed on a large set of diverse molecules. The database of parameters will allow us to parameterize and simulate new biological and synthetic molecules, such as ligands and drug compounds by using AMOEBA and AMOEBA+ FFs.

# 2. Methodologies

## 2.1 Theoretical background

The bonded interactions of AMOEBA[21] and AMOEBA+[22-23] include five components: bond stretching, angle bending, out-of-plane bending (opbend), torsion, and cross-coupling terms for bond/angle/torsion, as shown in **Equation 1**.

$$E_{bonded} = E_{stretch} + E_{bend} + E_{torsion} + E_{opbend} + E_{cross} \quad (1)$$

This study focuses on bond stretching, angle bending, stretch-bend coupling, and opbend terms, which are the standard contributing elements to the total potential energy shared by almost all the molecules. The torsional energy involving four atoms in connection is usually coupled with other interactions such as electrostatics, van der Waals and polarization. The torsion term and related cross-coupling terms are treated as a residual correction to match the ab initio quantum mechanics reference energy after all the other terms of AMOEBA(+), including non-bonded terms are determined.

As mentioned above, the AMOEBA valence terms have been mostly inherited from MM3/MM4 model.[24-25] The advantage of using the higher-order polynomials is that the energy variation within a wider range of geometry changes can be accurately captured. Bond stretching term adopts the quartic form (Equation 2), which can generate reliable energy for bond variation in the range of ±0.3 Å from the equilibrium structure. Also, the sextet bending model is utilized, which can reliably reproduce ab initio energy for the angle variation in the range of ±70° and beyond.[32] The out-of-plane bending (opbend) term is used to keep the planar structures for *sp2* atoms at and around the trigonal centers such as the amide group. Without the out-of-plane or improper torsion term used by some force fields, large pyramidalization may occur during the MD simulations without seriously distorting the bond lengths and angles. In AMOEBA, the opbend term (Equation (4)) shares a similar functional form to the angle bending term. The $\chi$ is defined as the bending angle between the plane and the central atom (**Figure 1**), which means each out-of-plane structure has three components contributed by each peripheral atom.

$$E_{stretch} = k_b \Delta r^2 \left[1 - 2.55\Delta r + \frac{7}{12} \times 2.55^2 \Delta r^2\right] \quad (2)$$

$$E_{bend} = k_a \Delta\theta^2 [1 - 0.014\Delta\theta + 5.6(10)^{-5}\Delta\theta^2 - 7.0(10)^{-7}\Delta\theta^3 + 2.2(10)^{-8}\Delta\theta^4] \quad (3)$$

$$E_{oop} = k_{opbend} \Delta\chi^2 [1 - 0.014\Delta\chi + 5.6(10)^{-5}\Delta\chi^2 - 7.0(10)^{-7}\Delta\chi^3 + 2.2(10)^{-8}\Delta\chi^4] \quad (4)$$

Here $\Delta r = r - r_0$ is the difference between the real bond length and the reference $r_0$, also, we have $\Delta\theta = \theta - \theta_0$ and $\Delta\chi = \chi - \chi_0$. The cross-terms (Equation 5) describe the coupling effects among bond stretching, angle bending, and rotational torsion. In this study, we focus

on the stretch-bend cross term only. Torsion and related cross-terms are included in AMOEBA functional forms and will be considered later. These torsion-related cross-terms are needed to describe the hyperconjugation effects for conjugated molecules.[33] The stretch-bend cross-term is expressed in Equation (6): for any angle A-B-C, two parameters $k_{b,AB}$ and $k_{b,BC}$ denotes the coupling of the bond AB, BC respectively with angle ABC.

$$E_{cross} = E_{str-bnd} + E_{str-str} + E_{bnd-bnd} + E_{str-tors} + E_{bnd-tors} + E_{tors-tors} \quad (5)$$

$$E_{str-bnd} = [k_{ba,AB}\Delta r_{AB} + k_{ba,BC}\Delta r_{BC}]\Delta\theta_{ABC} \quad (6)$$

In this work, the distinct bond types are defined by the atomic types of 2 atoms forming the bond (A-B). Also, the different angle types have been classified by the atomic types of 3 atoms forming the angle (A-B-C).[23] Slightly simple types are used for the stretch-bend energy term, where any angles with the same central atom use the same parameters. Besides, the opbend terms have been classified by the type of central *sp2* atom with the type of one peripheral atom.

## 2.2 Initial bond/angle parameters from QM Hessian matrix

In this work, we adopt the modified Seminario method[31, 34] to generate our initial bond stretching and angle bending parameters directly from the QM Hessian matrix. The modified Seminario method is a convenient and efficient way to derive accurate stretching and bending parameters with all other contributions ruled out (as opposed to fitting).

The Seminario method was initially designed for the generation of OPLS-AA[35] style parameters and written in MATLAB. We have re-written the code in Python based on the atom types we defined here for organic molecules (see **Data availability**).[23] The OPLS-AA FF uses the classical harmonic model for stretching and bending, which means the modified Seminario cannot be directly applied to the parameterization based on our model. The resulting parameters from this method serve as excellent initial values for further optimization against experimental data, which is significantly improved over the previous simple estimates used by Tinker[36] or Poltype[37] tools.

The mathematical details and principles about the modified Seminario are described here. The Hessian matrix is written based on the Cartesian coordinates of each atom, which is further divided into many small 3×3 matrices belonging to different bonds, such as $\boldsymbol{H_{AB}}$ for the bond AB in Equation (7). Stretching parameters can be obtained straightforwardly by using Equation (7), where $\lambda_i^{AB}(i = 1,2,3)$ are the eigenvalues of the small Hessian for the bond AB and $\vec{v}_i^{AB}(i = 1,2,3)$ are the corresponding eigenvectors, $\vec{u}_{AB}$ is the bond vector (All the vectors mentioned here and in the next paragraph are normalized).

$$\boldsymbol{H_{AB}} = \begin{bmatrix} \frac{\partial^2 E}{\partial x_A \partial x_B} & \frac{\partial^2 E}{\partial x_A \partial y_B} & \frac{\partial^2 E}{\partial x_A \partial z_B} \\ \frac{\partial^2 E}{\partial y_A \partial x_B} & \frac{\partial^2 E}{\partial y_A \partial y_B} & \frac{\partial^2 E}{\partial y_A \partial z_B} \\ \frac{\partial^2 E}{\partial z_A \partial x_B} & \frac{\partial^2 E}{\partial y_A \partial z_B} & \frac{\partial^2 E}{\partial z_A \partial z_B} \end{bmatrix}, k_{AB} = \sum_{i=1}^{3} \lambda_i^{AB} \left| \vec{u}_{AB} \cdot \vec{v}_i^{AB} \right| \quad (7)$$

As to deriving angle bending parameters, additional work is involved. Inside the equations (8) below, we need to first derive the intermediate force constants in the direction perpendicular to the bonds forming the angle (vectors $\vec{u}_{PA}$ and $\vec{u}_{PC}$ with the orientation pointing to the interior in **Figure 2(a)**). Then, by considering the relation of those normal vectors and their contributions to the angle bending, the equations (9) and (10) are used to calculate the final bending force constants. More mathematical and analytical details about this method can be found elsewhere.[31] The *N* is the number of angles in the FF with a central atom B and involves the bond AB. Take oxalic acid as an example in **Figure 2(b)**; atom A is the peripheral atom for two coupled angles: ABC and ABE. To obtain the force constant for angle ABC $k_{\theta(ABC)}$, we can see the vector $\vec{u}_{PA,1}$ perpendicular to bond AB points to ABC and $\vec{u}_{PA,2}$ perpendicular to bond AB points to ABE. However, this method is not appropriate to deal with angles involving *sp*-hybridized atoms, because it cannot detect the orientation. *Sp* atoms are not common, so some more rough estimates are used here for initial values in the subsequent fitting to QM vibrational frequencies (VF).

$$k_{PA} = \sum_{i=1}^{3} \lambda_i^{AB} \left| \vec{u}_{PA} \cdot \vec{v}_i^{AB} \right|, \qquad k_{PC} = \sum_{i=1}^{3} \lambda_i^{CB} \left| \vec{u}_{PC} \cdot \vec{v}_i^{CB} \right| \quad (8)$$

$$s_{PA} = \begin{cases} 1 + \dfrac{\sum_{j=1}^{N} \left| \vec{u}_{PA,1} \cdot \vec{u}_{PA,j} \right|^2 - 1}{N-1} & (N > 1) \\ 1 & (N = 1) \end{cases} \tag{9}$$

$$\frac{1}{k_{\theta(ABC)}} = \frac{s_{PA}}{R_{AB}^2 k_{PA}} + \frac{s_{PC}}{R_{CB}^2 k_{PC}} \tag{10}$$

It is convenient to derive the initial bond/angle parameters using the Modified Seminario method from the QM Hessian Matrix of a given molecule. However, since this model is appropriate mainly for the harmonic model, the parameters need to be further fitted to the QM VF. The whole procedure has been illustrated in **Figure 5** and implemented in an automated program which will be described in Section 2.4.

## 2.3 Computational details

### (1). Selection of training set molecules

The training set consists of 445 molecules, with the distribution of the number of heavy atoms in each molecule shown in **Figure 3**. The whole range is 1 to 18 heavy atoms with most molecules having 5 to 9 heavy atoms. All the molecules have been classified into categories that span from alkane to bio-fragments, as illustrated in **Figure 4** and **Table 1**. In total there are grouped into seven categories: (1). alkanes and aliphatic rings; (2). molecules comprised of only single bonds; (3). molecules with double bonds or triple bonds; (4). benzene derivatives; (5). heterocyclics; (6). bio-fragments; (7). combinations of multiple aromatic cycles. The detailed list of molecules with their SMARTS and specific categories is in **Table S1**. The whole procedure of parametrization has been illustrated by the workflow in **Figure 5**.

### (2). Initial bond/angle parameters from Hessian matrix

The frequencies of the normal modes have been calculated by using Gaussian09[38] with MP2/6-31g** (6-311g** for iodine atom) after the tight optimization with the same method and basis set. The bond/angle equilibrium values at MP2/6-31g** are close to the experiment for the organic molecules.[39] In this work, the reference bond lengths and angles were extracted from the final optimized structures and averaged based on the pre-defined bond and angle

types. Then, the initial force constant parameters for bonds and angles were derived from the Hessian matrix by using the modified Seminario method.[31] These parameters were further refined in step (3), targeting the QM frequencies.

**(3). Bond/angle parameters refinement using QM frequency**

As suggested previously[40-41], the frequencies calculated by MP2/6-31g** (6-311g** for iodine) are scaled by a factor of 0.943 for better consistency with the experimental values. The vibrational analysis program (*vibrate.x*) in Tinker was used to generate normal mode frequencies for our models. Since we only focus on the interactions related to bonds and angles here and the frequencies of high range (usually large than 1000 $cm^{-1}$) are mainly contributed by those terms while the low range depends mainly on torsional terms, the reference data lower than 1000 $cm^{-1}$ are not used in the refinement. The least-square method from scipy[42] is used to optimize the parameters for bond, angle, and stretch-bend cross-terms, with the cost, root-mean-square-error (RMSE in VF), as given by Equation 11.

$$RMSE(frequency) = \sqrt{\frac{1}{n}\sum_{i}(\nu^{MM} - \nu^{QM,scaled})^2} \qquad (11)$$

**(4). Opbend parameters derivation from potential energy surface fitting**

The potential energy surfaces of the training set molecules were utilized to derive the out-of-plane bending (opbend) parameters (further details in **Table S2**). We focus on those molecules containing the *sp2* atoms that form trigonal planar structures. The central *sp2* atom was pulled out of the plane by 0.05, 0.1, 0.15 Å (Opbend-small subset hereafter), and 0.2, 0.25, 0.3 Å (Opbend-large subset hereafter), respectively, to create deformed configurations and corresponding energies. We obtained 4318 configurations for the training set, and the single point energy was calculated using Psi4[43] at MP2/aug-cc-pvtz (aug-cc-pvtz-pp for iodine) level of theory.

The least-square method was used to fit the opbend parameters by optimizing the weighted RMSE of the energy difference (shown by Equation 12). To ensure that our parameters are

not biased towards the high-energy (rare) structures far from equilibrium, we used a smaller weight (0.2) for configurations with energy 20 kcal/mol higher than that of equilibrium.

$$wRMSE(energy) = \sqrt{\frac{1}{n}\sum_{i}[(E^{MM,small} - E^{QM,small})^2 + 0.2 \times (E^{MM,large} - E^{QM,large})^2]} \quad (12)$$

**(5). Validation**

After obtaining the parameters by fitting frequencies and potential energy surface, we chose 120 new molecules as the validation set (see **Table S3 & S4** for details). The distribution of the number of heavy atoms in molecules of the validation set has been illustrated in **Figure 3**. Overall, the molecules in the validation set contain 4~14 heavy atoms. Those molecules are made of functional groups included in the training set molecules (**Table 1**). We mainly evaluated the bond/angle and coupling parameters by comparing the high-range frequencies of normal mode for these molecules. The opbend parameters were validated by comparing the energy differences between QM and our model for 829 configurations stemming from the 120 molecules in the validation set. This validation is intended to show the transferability of the current valence parameters.

## 2.4 Automated generation and assignment of valence parameters for new molecules

Although we have covered a range of chemistry (see **Table 1**) widely encountered in the practical simulations, there remains the possibility of failure to find the parameters for a certain molecule. For the molecules that the exact SMARTS patterns are missing in our database, we attempt to assign parameters based on the highest similarity to the existing chemical fragments.

Here we provide an automated program to determine the best matching atomic type for any new molecule and assign both valence parameters and our previously published charge-flux parameters[23] to each atom (see **Data availability**). It will attempt to find an *exact matching* from our database at first. When the *exact matching* fails, this program has implemented 2 methods to estimate those missing parameters: (1). Modified Seminario (in the red dash-line

rectangle of **Figure 5**); (2). Reversed searching by ranking tree (in cyan dash-line rectangle of **Figure 5**). In this work, we used the second method to generate the new parameters for the validation set.

The second method depends on the current parameter set to generate new parameters suitable for our model. We established a ranking tree where different types were placed in some positions in different levels determined by the complexity of the structure (**Figure 6**). For example, given a bond A-B which has no direct equivalent in the database, the program will check the position of types A and B in the ranking tree. It undergoes a reversed searching starting from the bottom (most complex), during which the types A and B are sequentially replaced by other types in the same level under the same root. If there is still no matching item, the searching will move to the upper level and continue this operation. The final parameter will be given by averaging the first 5 parameters found during the reversed searching.

For the common *sp3* carbon, we have introduced canonical types (*nsp3* C) based on the number of H atoms attached. They are designated as C30, C31, C32, C33, or C34 (See **Table S5**). Given a molecule, the detailed atom types of any *sp3* C atoms will be identified by comparing their local environment (SMARTS) with the database. If there is no complete match, these canonical types will be used to generate parameters. For instance, if we have a molecule CH3-CH(CH3)-CH2-CHO and find no angle parameter of CH-CH2-CHO, then we will collect all the known angle parameters for the structure as *nsp3* C – *nsp3* C – CHO in the database to compute the average value and assign its value to the above angle. If still no appropriate parameter is found, we will continue reversed searching on the carbon atoms which are not *sp3*. If high accuracy is required under this method, we can use the frequency fitting to improve all the valence terms except the opbend terms.

# 3. Results and discussion

## 3.1 Overview of the atomic types and valence parameters

Currently, we have determined the valence parameters based on about 270 atomic types.

All the parameters were derived by both fitting to QM data or the ranking tree. To be specific, we generated 768 bond terms (of which 42 terms from the ranking tree, with each term including $k_b$ and $b_0$), 1946 angle terms (of which 186 terms from the ranking tree, with each term involving $k_a$ and $\theta_0$), 109 stretch-bend terms (one term involves $k_{ba,AB}$ and $k_{ba,BC}$), and 565 opbend terms (of which 34 from the ranking tree, one term involves $k_{opbend}$).

The force constants have been determined for different structures by using our version of the modified Seminario program for bond and angle parameters. All the reference bond ($b_0$) and angle ($\theta_0$) are averaged over the same atomic types. The current atomic types can keep the overall RMSD of the reference bond length less than 0.05Å and angle degree less than 5° for our training set comparing to the QM optimized structures (see **Data availability**), which is acceptable in MD simulations[23]. As to cross terms, our current setting only turns on the stretching-bending coupling effects. The stretch-bend parameters are only distinguished by the atomic type of the central atom, i.e., all the angle structures with the same central atom share the same stretch-bend parameters. Next, the classification of out-of-plane bending terms (opbend) only considers the atomic type of the central atom and the side atom, which possibly pop out of the triangle plane. We will analyze this in detail in the following sections.

### 3.2 Analysis on parameters of the valence terms

We examined the force constants ($k_b$) and equilibrium bond lengths ($b_0$) between different *sp3* C and H. It is observed from **Table 2** that the *sp3* C with different numbers of the bonded hydrogens does not show a remarkable change in the bond length of C-H and corresponding force constant, especially the negligible difference among C(H3)-H, C(H2)-H, C(H1)-H. So, the bond type for terminal C and H can be equated with that for middle C and H. For most cases, the local functional groups have minimal effects on C-H because the RMSD of the lengths of all the matching bonds in the database is below 0.03 Å.

Next, we listed the possible pairs of the bonds between the canonical *sp3* C in **Table 2**. The bond lengths between *sp3* C and *sp3* C with the different number of connected hydrogens are mostly within the range of 1.52 Å~1.53 Å, but it is observed that less number of bonded

hydrogens lead to longer bond lengths and smaller force constants. The C(H3)-C(H1) has the largest force constant, but it involves a terminal carbon with 3 H atoms. For those carbons of the interior, this trend is mostly correct, suggesting a strong coupling between bonds and the chemical environment (other bonds and angles) inside the molecule.

As for other kinds of bonds and angles, the analysis is becoming complicated, and no clear trend is observed.

### 3.3 Normal mode frequency

The frequencies of the normal modes for all 445 molecules in the training set have been utilized for determining force constants after a tight structure optimization. The correlation between QM reference and our model is illustrated in **Figure 7(a)**. The RMSD of the frequency of the normal modes reaches 36.6 $cm^{-1}$ for the whole training set; at the same time, the relative error percentage is around 1.7% (

**Table 3**). By comparison, it was observed that OPLS combined with the modified Seminario method could generate frequencies with the relative errors of 7.4% (RMSD: 59.4 $cm^{-1}$) in general.[31] AMBER-type valence terms combined with a strategy of partial hessian fitting gives out RMSD ~81.4 $cm^{-1}$ from a much smaller database,[44] and GAFF would show frequency error mostly larger than 100 $cm^{-1}$.[1] Meanwhile, these previous works on vibration analysis used small databases limited in the number and the variety of molecules, which in turn demonstrates the reliability of the valence model and parameters derived in this work using a large and diverse set of molecules.

As seen in **Figure 7(a)**, most valence related vibration frequencies concentrate on two areas (1000~1800 $cm^{-1}$ and 2800~3600 $cm^{-1}$). The valence terms can capture those modes in the range 2800~3600 $cm^{-1}$, which are contributed mainly by the fast stretching between heavy atoms and hydrogen such as C-H, N-H, or O-H. These motions occur at the terminal of the molecule, indicating limited coupling effects with the local angle bending. It is a common practice to restrain the H related high frequency stretching during molecular simulation to allow larger time steps[45-46], while we have found reversible reference system propagator algorithms (RESPA) is a more effective approach – a small time step (e.g., 0.5 fs) can be used for valence forces and a large time step (e.g., 2 fs) is applied to the other contributions. The situation becomes more complicated when it comes to the range of 1000~1800 $cm^{-1}$. These involve double bond stretching, angle bending, and the mixture of several other motions. Because these motions directly affect the properties of interest in MD, the use of cross-terms and high-order polynomial models (for anharmonicity) becomes important. The detailed frequency comparisons for several typical molecules have been displayed in **Figure S1**, proving the reliability of our current model and parameters.

Some data points in **Figure 7(a)** with relatively large deviations arise from carbamate and guanine derivatives (several molecules having such functional groups). In carbamic acid and carbamate ion, we found that the normal mode contributed by the stretching of C=O is mixed with certain local angles bending (O-C-O, N-C-O, and more) lower frequency. The QM gives 1752 $cm^{-1}$ in carbamate ion and 1893 $cm^{-1}$ in carbamic acid for this coupled motion, while our

model gives the frequencies inside 2000~2100 $cm^{-1}$. For guanine, the deviation also happens on the C6=O stretching. It involves the angle bending N1-C6-C5 and both two ring structures show large deformation. Because we only utilized stretch-bend cross-terms in this work, there might be part of coupling effects that cannot be fully described. More coupling terms between different structures (such as those involving torsion) can be included to improve the performance for these complex situations further.

Here we assessed how much the cross-terms and the opbend terms contribute to the overall performance in **Table 4**. We re-classify Group (1)~(3) into 2 groups: Group A and Group B by the number of heavy atoms in each molecule (Group A: <10; Group B: >=10). The cross-terms impact the overall performance of vibration analysis, but opbend terms show minimal influence since the previous works about the simple quadratic model for stretching and bending reported the best RMSD up to nearly 60 $cm^{-1}$. It reveals that anharmonicity plays the most significant role in generating accurate frequencies. Besides, the results are improved consistently for all the groups except Group B when stretch-bend cross terms have been used in this work, indicating that cross terms can capture more details in structural vibrations. The molecules in Group B mostly include long carbon chain or large carbon ring, which demand more coupling effects considered. Notably, the improvement for those molecules with largely conjugated structures is remarkable: the RMSD for benzene derivatives is decreased by 27.3%, 16.1% for heterocyclics, 16.5% for bio-fragments, and 31.6% for the combinations of multiple aromatic cycles. These molecules possess complex structures, so the motion of interior atoms can easily impact the surrounding bonds and angles. In general, the RMSD of frequencies for the training set goes down by 13.6% by adding cross-terms. Opbend terms have very little effect on the frequencies and made them slightly worse for certain groups. These are likely vibration modes involving the motion in which some *sp2* atoms popping out of the plane and adding opbend helps keep the correct structure and energy while introducing noise to vibrational frequencies.

### 3.4 Potential energy surface for opbend parameters

Out-of-plane parameters are used to avoid the abnormal pyramidalization of the planar structures with *sp2* atoms in the center. The results of the current polynomial model with the

corresponding parameters of the training set for opbend fitting have been illustrated in **Figure 8**. In **Figure 8**. Most of the data points are well matched between the model and QM energy within a wide range up to ~90 kcal/mol, while the model collectively underestimates the QM reference values. This suggests that our model can reflect the shape of potential energy surface during pyramidalization with high accuracy, even when the deformation of the structure is very large. The energy RMSD for the training set is 0.8375 kcal/mol for the Opbend-small set (maximum QM energy 22.9956 kcal/mol with average 2.3185 kcal/mol), 3.5226 kcal/mol for the Opbend-large set (maximum QM energy 91.0101 kcal/mol with average 12.9017 kcal/mol) and 2.5582 kcal/mol for the whole set (

**Table 5**). Although Allinger's model can improve the description of potential energy surface in the out-of-plane bending, the simple model cannot capture the effect fully. As a result, the model energy tends to be smaller than expected from overall observation. Nonetheless, this model separates the out-of-plane motion into three components from each side atoms, which makes the parameters more transferable and flexible in applying in many distinct planar structures.

### 3.5 Validation

To test the accuracy and transferability of the valence parameters derived in this work, we applied these parameters to all the 120 molecules in the validation set. All the parameters for the validation set molecules are automatically assigned or generated via the ranking tree method described in Methodologies. Frequency validation results are shown in **Figure 7(b)** and

**Table 3**. The RMSD in frequencies for each group of validation set is mostly consistent with the results from the training set, ca. ~34 $cm^{-1}$. In addition, we compared the equilibrium $b_0$ and $\theta_0$ parameters assigned based on atom type with QM-optimized geometry for the 120 molecules. The overall RMSD is less than 0.05Å for $b_0$ and less than 5° for $\theta_0$ (see **Data Availability**). For the out-of-plane effects, the performance of the opbend parameters is satisfying as shown in **Figure 8(b)** and

**Table 5**. The RMSD between QM and MM energy for all the 829 configurations from 120 molecules is ~2.8 kcal/mol. Overall the validation results demonstrated high accuracy and transferability of our parameters for a wide range of small organic molecules.

# 4. Conclusion

The classical harmonic equations have been used for decades to calculate the potential energy of bonded interactions by many FFs. While computationally attractive, it shows deficiencies in describing the potential energy surface and vibrational frequencies of molecules compared to quantum calculations. AMOEBA-family FFs apply the MM3(MM4)-style, anharmonic valence interaction model to improve the accuracy of structure and spectroscopic properties. The model employs high-order polynomial functionals in bond stretching and angle bending, the cross-terms that couple different valence motions, and out-of-plane bending terms. The benefits of these high-order and coupling terms are evidenced in this work. They noticeably improve the accuracy of normal mode frequencies, indicating a strong anharmonicity in bonds and angles. The cross-terms capture more details of the intramolecular motions and significantly improve normal mode frequencies with consistently high accuracy, especially for large structures and conjugation molecules. In addition, our results indicate that opbend terms generate correct potential energy surface around the planar structure for those *sp2* atoms.

The current parameter set can be applied to a wide range of distinct functional groups, and it is expandable to tackle additional chemistry in the future. The modified Seminario method has been combined with the current atom types of the AMOEBA(+) model, providing reliable initial force constant parameters of bond stretching and angle bending as reference values from QM Hessian. Refinement of bond, angle, and bond-angle coupling explicitly targeting the QM frequency data further increases the quality of these force constant parameters. The opbend parameters are derived by fitting to the potential energy surface of distorted structures composing *sp2* atoms. It is interesting to examine how good our model can describe the overall geometry of simple compounds comparing to high-level QM or experimental structures.

However, this will not be practical until all the other energy components are parameterized, as both the valence and non-boned interactions are responsible for the accurate prediction of the molecular geometry. An automated program to assign the parameters has also been presented in this work, which is accessible to the public (see Data availability). This program as well as the parameters, will be further implemented in the canonical Tinker [36] and Poltype software packages [37].

# Acknowledgments

This work was supported by the National Institutes of Health (R01GM106137 and R01GM114237), and National Science Foundation (CHE-1856173).

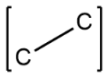
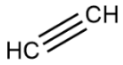
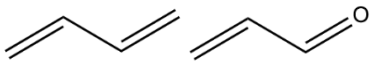
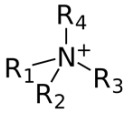
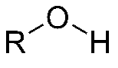
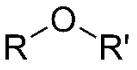
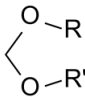
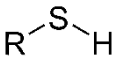
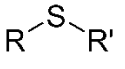
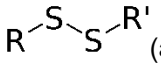
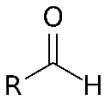
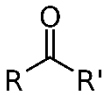
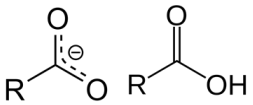
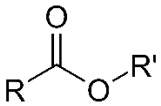
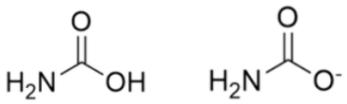
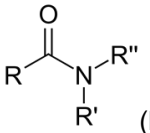
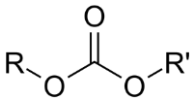
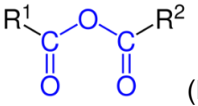
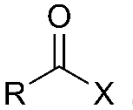
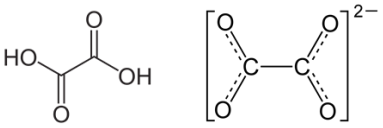
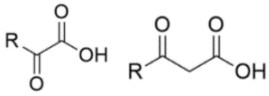
# Data availability

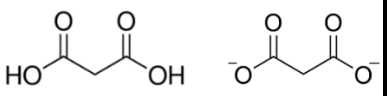
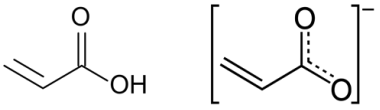
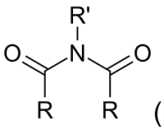
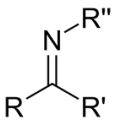
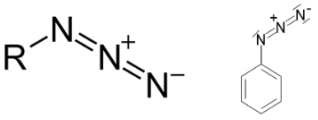
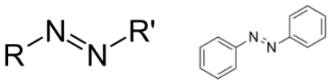
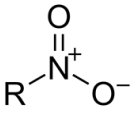
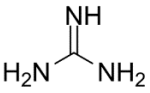
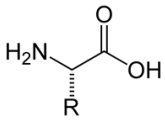
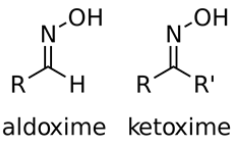
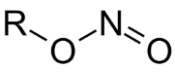
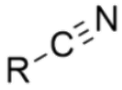
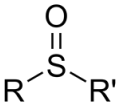
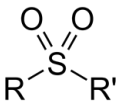
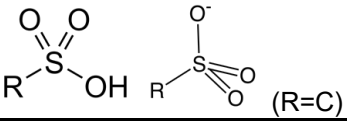
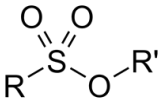
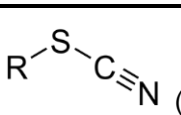
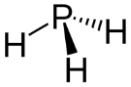
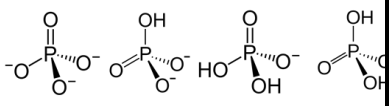
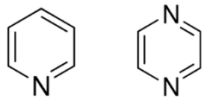
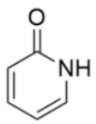
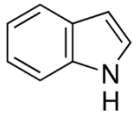
1. The molecules in tinker format and final bonded parameters can be accessed at: https://github.com/prenlab/amoebaplus_data/tree/master/valence2021

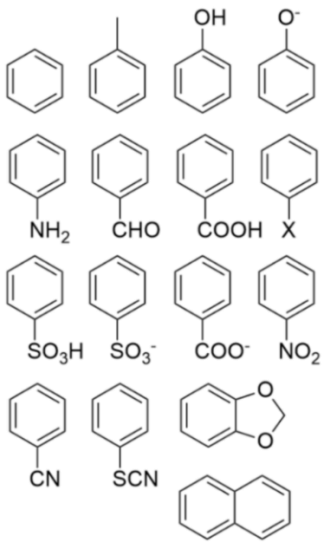
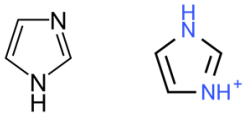
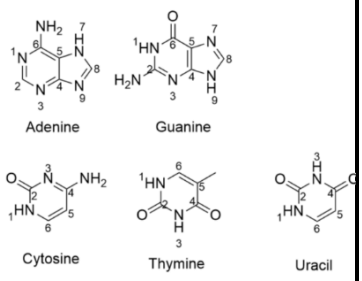
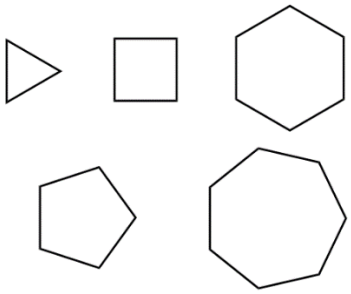
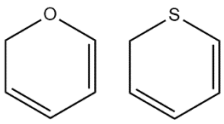
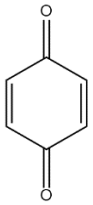
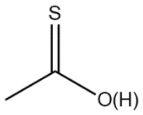
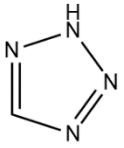
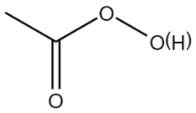
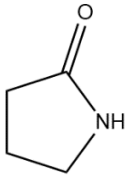
2. The programs developed in this work are accessible through the GitHub repository at: https://github.com/prenlab/amoebaplus_parameter

# Tables

**Table 1.** A general illustration of the functional groups and distinct chemical structures covered in the current study.

| Group | Chemical structure | Group | Chemical structure |
|---|---|---|---|
| Alkane |  (chain & cyclic) | Alkene |  |
| Alkyne |  | Conjugated molecules |  |
| Amine |  (all R = H, C) | Ammonium |  (all R = H, C) |
| Alcohol |  (R = C) | Ether |  (all R = C) |
| Di-oxyl |  (all R = C) | Thiol |  (R = C) |
| Sulfide |  (all R = C) | Disulfide |  (all R = C) |
| Halide | R-X (R = C; X=F, Cl, Br, I) | Aldehyde |  (R=C, H) |
| Ketone |  (R=C) | Carboxylic acid / Carboxylate |  (R=C, H) |
| Carboxylate ester |  (R=C, H; R'=C) | Carbamic acid / Carbamate |  |
| Amide |  (R, R', R''=C, H) | Carbonate / Bicarbonate |  (R=C, H; R'=C) |
| Anhydride |  ($R^1$, $R^2$=C) | Acyl halide |  (R=C; X=F, Cl, Br) |
| Oxalic acid Oxalate |  | α-keto acetate / β-keto acetate |  (R=C) |

| | | | |
|---|---|---|---|
| Malonic acid / Malonate |  | Acrylic acid/ Acrylate |  |
| Imide |  (R, R'=C) | Imine (Aldimine/ Ketimine) |  (R, R', R" remain 1 C) |
| Azide |  | Azo-compound |  |
| Nitro-compound |  (R=C) | Guanidine |  |
| Amino acid |  | Oxime |  (R=C, H; R' =C) |
| Nitrite |  (R=C) | Nitrile |  (R=C) |
| Sulfoxide |  (R, R'=C) | Sulfone |  (R, R'=C) |
| Sulfonic acid / Sulfonate |  (R=C) | Sulfonate ester |  (R, R'=C) |
| Thiocyanate |  (R=C) | Phosphine |  |
| Phosphoric acid / Phosphate |  | Pyridine / Pyrazine |  |
| Benzene Derivatives | | Conjugated lactam |  |
| | | Indole |  |

| | | | |
|---|---|---|---|
| |  | Imidazole / Imidazolium |  |
| NA Base |  | Aliphatic cycles and derivatives |  |
| Pyran |  | Benzoquinone |  |
| Thioacetate |  | Tetrazole |  |
| Peracetic acid/ Peroxoate |  | Pyrrolidinone |  |

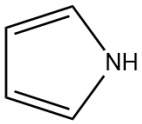
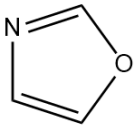
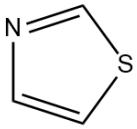
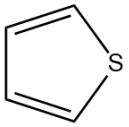
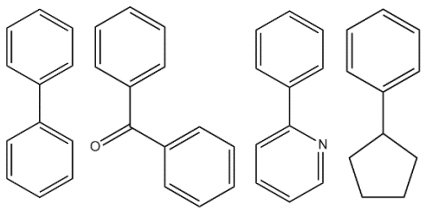
| | | | |
|---|---|---|---|
| Pyrrole |  | Oxazole |  |
| Thiazole |  | Thiophene |  |
| Connected structure of double heterocyclics, benzenes or aliphatic cycles |  | | |

**Table 2.** The comparison of equilibrium bond lengths and corresponding force constants among *sp3* C and H. [a)]

| bond [b)] | Equilibrium length (Database) [c)] | Force constant (Database) [c)] |
|---|---|---|
| **C[H4]-H** | 1.0897 (4) 0.0000 | 383.57 |
| **C[H3]-H** | 1.0940 (1297) 0.0297 | 345.98 |
| **C[H2]-H** | 1.0957 (448) 0.0028 | 340.53 |
| **C[H1]-H** | 1.0980 (52) 0.0034 | 338.72 |
| **C[H3]-C[H3]** | 1.5255 (1) 0.0000 | 266.49 |
| **C[H3]-C[H2]** | 1.5225 (153) 0.0065 | 203.31 |
| **C[H3]-C[H1]** | 1.5240 (82) 0.0052 | 280.35 |
| **C[H3]-C[H0]** | 1.5306 (16) 0.0044 | 210.15 |
| **C[H2]-C[H2]** | 1.5236 (30) 0.0046 | 205.76 |
| **C[H2]-C[H1]** | 1.5231 (9) 0.0112 | 214.73 |
| **C[H2]-C[H0]** | 1.5352 (3) 0.0094 | 190.84 |
| **C[H1]-C[H1]** | 1.5420 (1) 0.0000 | 172.75 |
| **C[H0]-C[H0]** | 1.5697 (1) 0.0000 | 133.12 |

a) Statistics are reported from all the configurations in our database.

b) Here, C[H*n*] means the *sp3* C with *n* hydrogens bonded.

c) The numbers in the parenthesis represent the number of matching structures in the database, and the underlined number means the RMSD: root mean square derivation. (Unit: length – Å, force constant – kcal·mol$^{-1}$·Å$^{-1}$))

**Table 3.** Statistical comparison of QM scaled frequencies of normal modes and the values calculated by MM model.

| **Statistics** [a] | **Training set** [b] | | | **Validation set** [b] | | |
|---|---|---|---|---|---|---|
| | RMSE | MRE | Num. | RMSE | MRE | Num. |
| **Group A** | 38.4279 | 1.81% | 4846 | 37.9629 | 1.82% | 1181 |
| **Group B** | 34.2120 | 1.72% | 354 | 35.2745 | 1.59% | 146 |
| **Benzene derivatives** | 28.5765 | 1.45% | 1433 | 32.1555 | 1.48% | 982 |
| **Heterocyclics** | 41.0277 | 1.91% | 1122 | 33.3864 | 1.71% | 541 |
| **Bio-fragments** | 39.0039 | 1.84% | 831 | 38.6137 | 1.94% | 81 |
| **Combination of multiple aromatic cycles** | 28.8252 | 1.41% | 752 | 26.0445 | 1.30% | 449 |
| **Total** | 36.6139 | 1.73% | 9338 | 34.0284 | 1.62% | 3410 |
| | | **$R^2$**=0.9981 | | | **$R^2$**=0.9983 | |

a) Group A includes those molecules with less than 10 heavy atoms (non-hydrogen) and Group B means those molecules with $\geq 10$ heavy atoms. Both Group A and Group B contain those molecules which are not involved in other categories listed in the table.

b) RMSE: root mean square error (unit: $cm^{-1}$); MRE: mean relative error ($\frac{1}{N}\sum_{i=1}^{N}\frac{|\nu(QM)-\nu(MM)|}{\nu(QM)}\times 100\%$). The number of respective unique data points is marked with Num. $R^2$: correlation coefficient.

**Table 4.** Statistical comparison of QM scaled frequency of normal modes and the values calculated by MM model with or without coupling terms and opbend terms in the training set.

| Group name [a] | w. coupling w. opbend [b] | | w. coupling w.o. opbend [b] | | w.o. coupling w. opbend [b] | |
|---|---|---|---|---|---|---|
| | RMSE | MRE | RMSE | MRE | RMSE | MRE |
| **Group A** | 38.4279 | 1.81% | 39.0842 | 1.83% | 41.9415 | 1.95% |
| **Group B** | 34.2120 | 1.72% | 33.4050 | 1.69% | 33.2874 | 1.73% |
| **Benzene derivatives** | 28.5765 | 1.45% | 28.5561 | 1.45% | 39.3057 | 1.93% |
| **Heterocyclics** | 41.0277 | 1.91% | 39.8838 | 1.89% | 48.8816 | 2.38% |
| **Bio-fragments** | 39.0039 | 1.84% | 38.7749 | 1.88% | 46.6939 | 2.33% |
| **Combination of multiple cycles** | 28.8252 | 1.41% | 28.8945 | 1.42% | 42.4678 | 2.11% |
| **Total** | 36.6139 | 1.73% | 36.7788 | 1.75% | 42.6315 | 2.04% |

a) Group A means those molecules with less than 10 heavy atoms (non-hydrogen) and Group B means those molecules with $\geq 10$ heavy atoms. Both Group A and Group B contain those molecules which are not involved in other categories listed in the table.

b) RMSE: root mean square error (unit: $cm^{-1}$); MRE: mean relative error ($\frac{1}{N}\sum_{i=1}^{N}\frac{|\nu(QM)-\nu(MM)|}{\nu(QM)}\times 100\%$) has been included in the parenthesis. $R^2$: correlation coefficient.

**Table 5.** Statistical comparison of QM energy differences and the values calculated by the MM model.

| Group name [b] | | Training set [c] | Validation set [d] |
|---|---|---|---|
| Opbend-small [a] | RMSE | 0.8375 | 1.3453 |
| | QM_Max | 22.9956 | 17.6173 |
| | QM_Ave | 2.3185 | 3.4736 |
| Opbend-large [a] | RMSE | 3.5226 | 4.1265 |
| | QM_Max | 91.0101 | 53.3861 |
| | QM_Ave | 12.9017 | 15.4029 |
| Total | RMSE | 2.5582 | 2.8082 |
| | $R^2$ | 0.9743 | 0.9471 |
| | QM_Ave | 7.6003 | 8.2367 |

a) Opbend-small and Opbend-large groups rely on the scale of deformation of the specified structures (see section 2.2).

b) RMSE: root mean square error; QM_Max: maximum of the absolute QM $\Delta E$; QM_Ave: average of the absolute QM $\Delta E$. (Unit: kcal·mol$^{-1}$).

c) $R^2$: correlation coefficient. The total number of configurations for the training set is 4318.

d) $R^2$: correlation coefficient. The total number of configurations for the validation set is 829.

## Figures

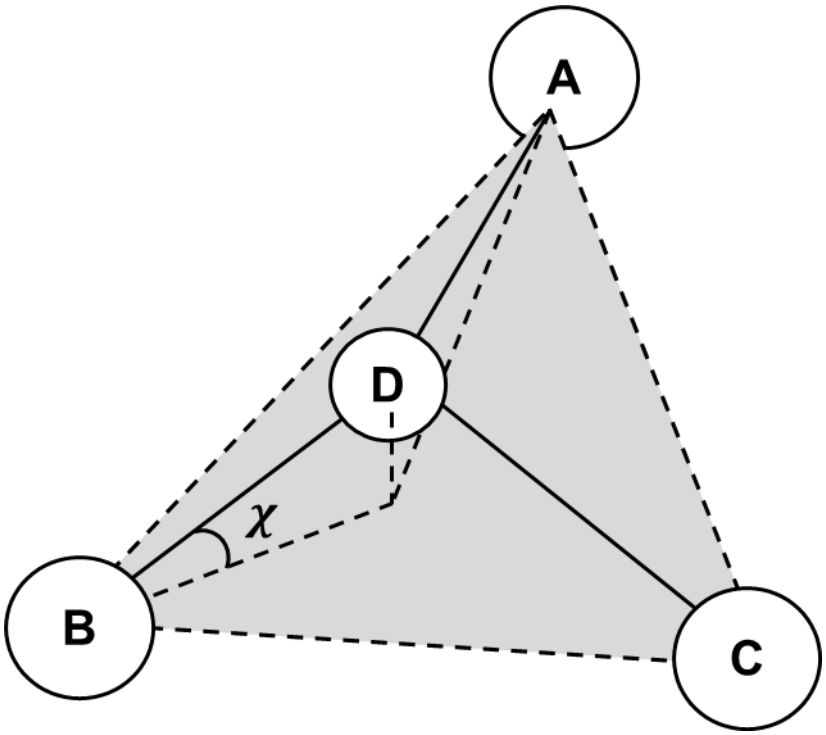



**Figure 1.** An illustration of a trigonal center D pyramidalizing from plane ABC. The out-of-plane angle $\chi$ has been defined as illustrated.

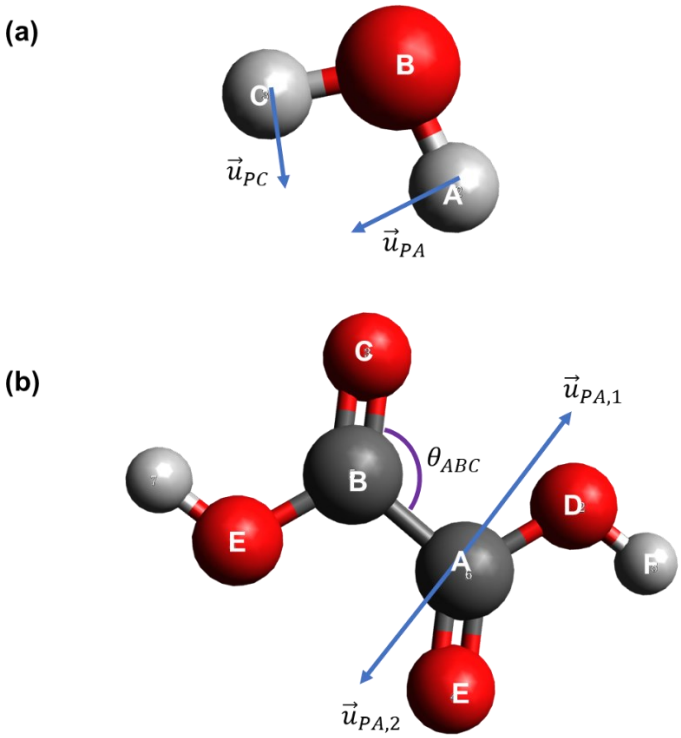



**Figure 2.** An illustration of the vectors used in the Seminario method. (a) The 2 vectors $\vec{u}_{PA}$ and $\vec{u}_{PC}$ contributing to the force constant of angle ABC in water molecule; (b). The 2 vectors pointing to 2 coupling angles with A as a side atom in oxalic acid.

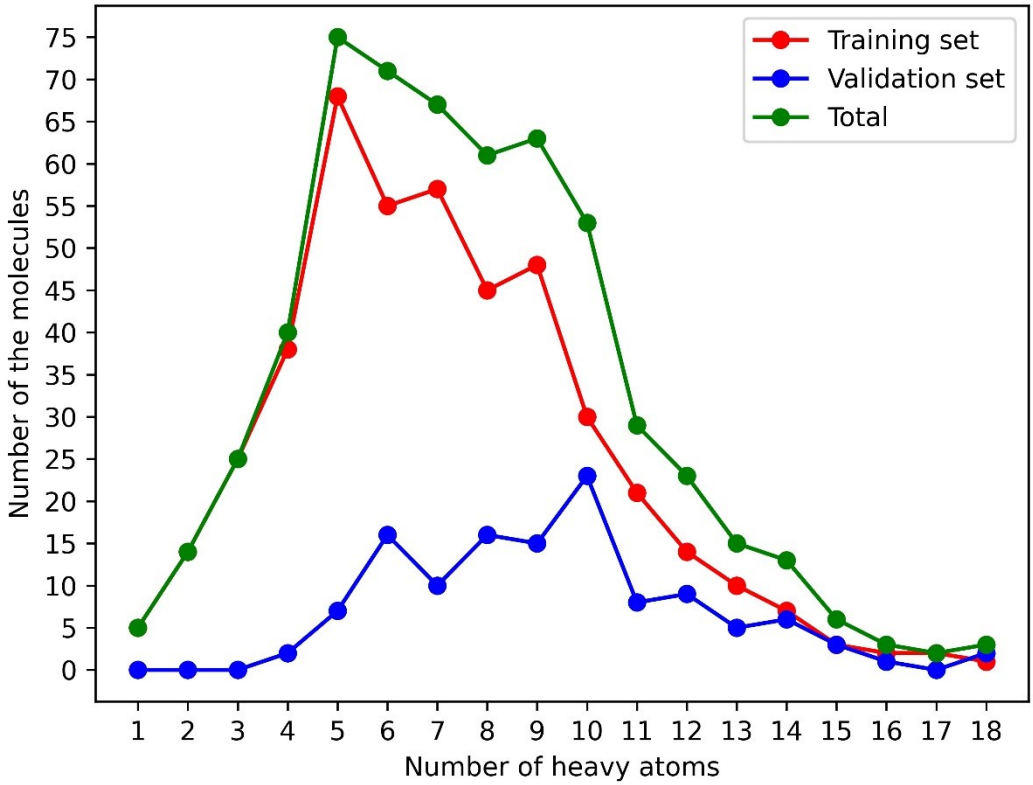



**Figure 3.** Distribution of the number of heavy atoms in different molecules. (Red: training set; Blue: validation set; Green: total.)

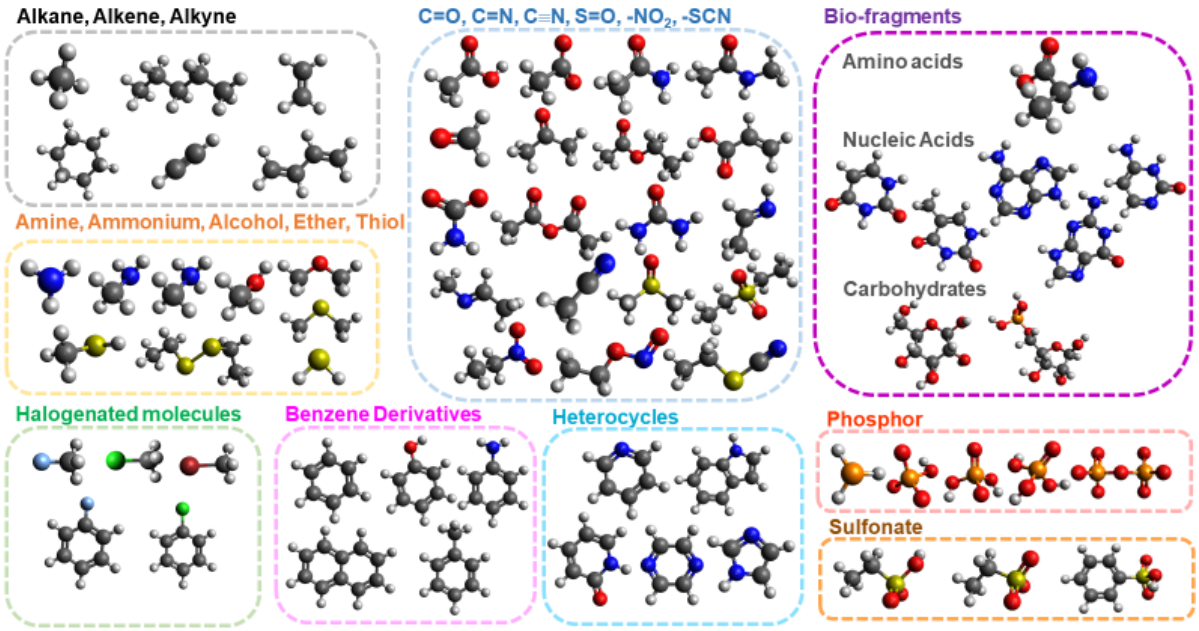



**Figure 4.** Representatives of the organic molecules covered in this study.

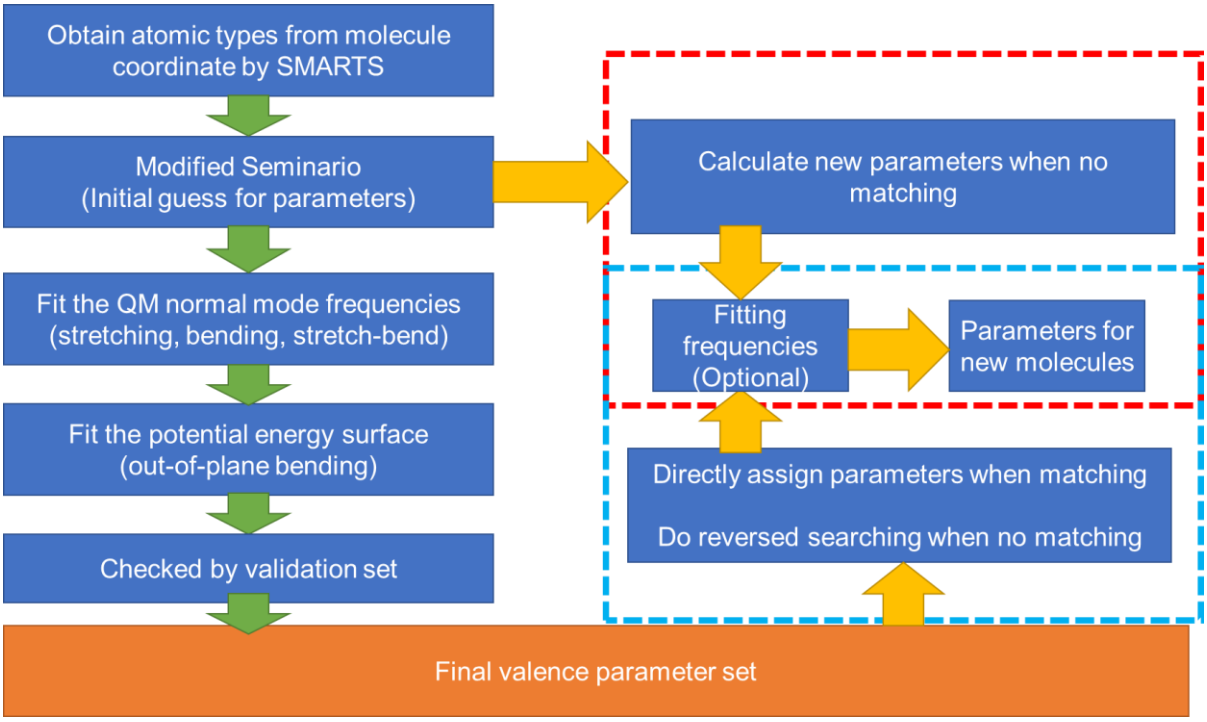



**Figure 5.** The procedure to generate the final valence parameter set (green arrows) and the methods to obtain and assign parameters for any new molecule (yellow arrows): assignment by Modified Seminario is shown inside the red dash box, and assignment by reversed searching is shown inside the blue dash box.

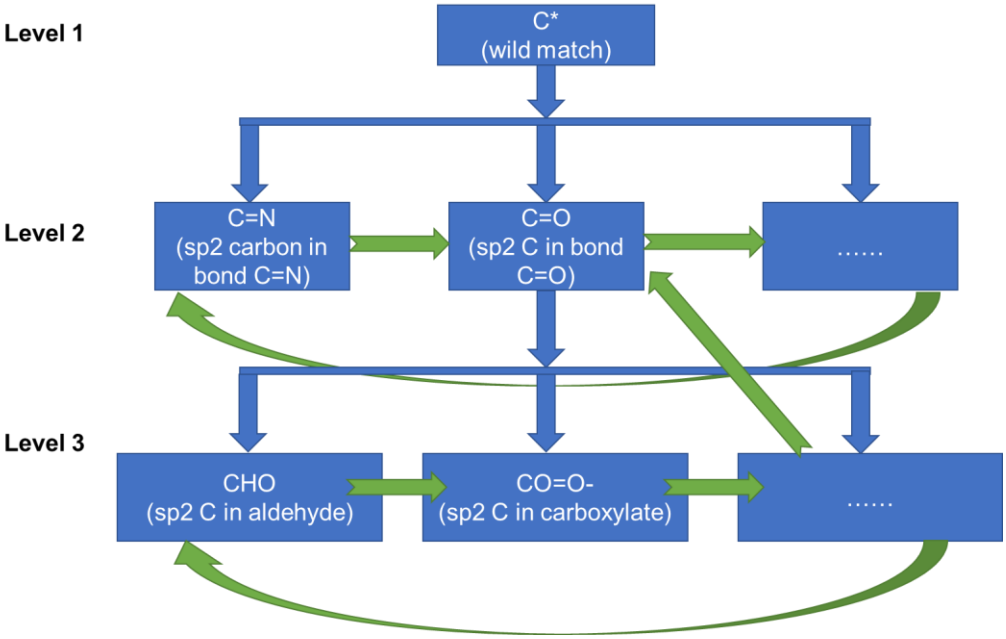



**Figure 6.** The basic structure of the ranking tree (simplified). The green arrows show the sequence of searching in the same level and reversely to the upper levels.

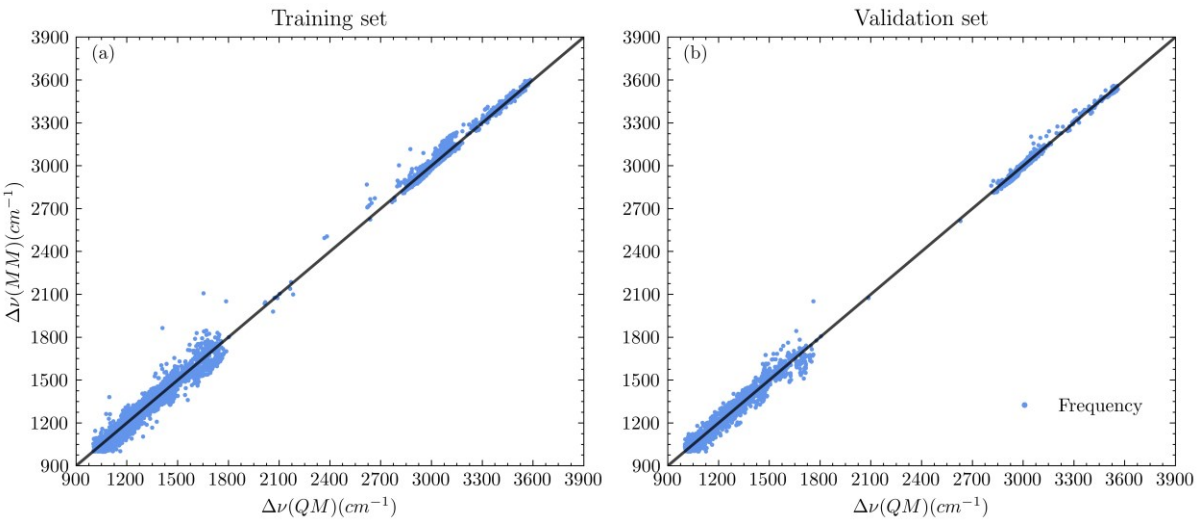



**Figure 7.** The correlation between the QM and MM normal mode frequencies($cm^{-1}$) for the molecules in **(a)** the training set and **(b)** the validation set.

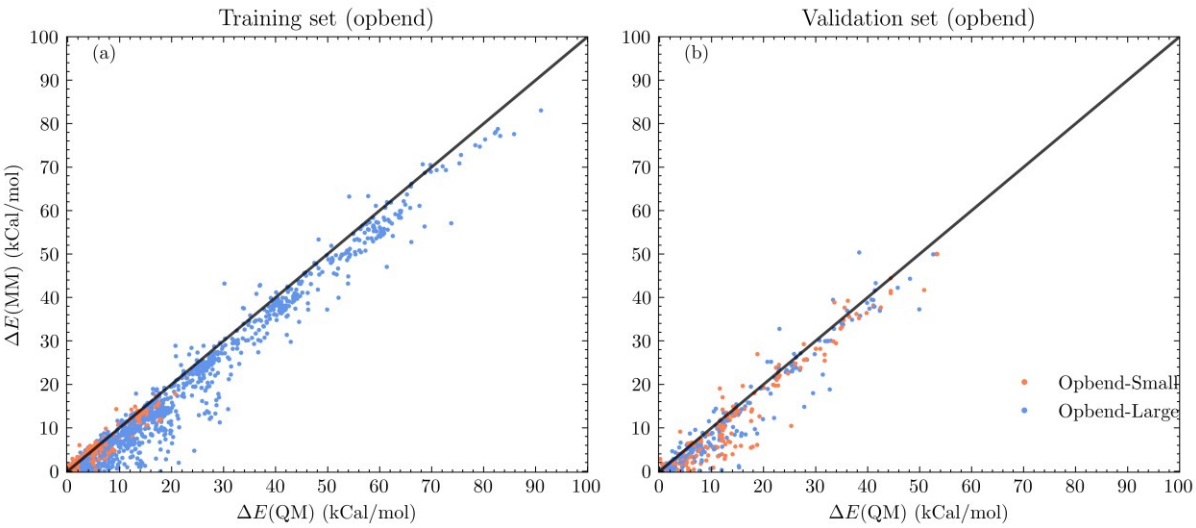



**Figure 8.** The correlation between the energy difference calculated by QM and MM for all the configurations in **(a)** the training set and **(b)** the validation set. Opbend-small refers to small deformation and opbend-large set refers to large deformation of the specified structures, as described in Section 2.3(4).

# References


1. Dauber-Osguthorpe, P.; Hagler, A. T., Biomolecular force fields: where have we been, where are we now, where do we need to go and how do we get there? *J Comput Aided Mol Des* **2019,** *33* (2), 133-203.
2. Hagler, A. T., Force field development phase II: Relaxation of physics-based criteria... or inclusion of more rigorous physics into the representation of molecular energetics. *J Comput Aided Mol Des* **2019,** *33* (2), 205-264.
3. Jing, Z.; Liu, C.; Cheng, S. Y.; Qi, R.; Walker, B. D.; Piquemal, J. P.; Ren, P., Polarizable Force Fields for Biomolecular Simulations: Recent Advances and Applications. *Annu Rev Biophys* **2019,** *48*, 371-394.
4. Gilson, M. K.; Zhou, H. X., Calculation of protein-ligand binding affinities. *Annu Rev Biophys Biomol Struct* **2007,** *36*, 21-42.
5. Gallicchio, E.; Levy, R. M., Recent theoretical and computational advances for modeling protein-ligand binding affinities. *Adv Protein Chem Struct Biol* **2011,** *85*, 27-80.
6. Harger, M.; Lee, J. H.; Walker, B.; Taliaferro, J. M.; Edupuganti, R.; Dalby, K. N.; Ren, P., Computational insights into the binding of IN17 inhibitors to MELK. *J Mol Model* **2019,** *25* (6), 151.
7. Kaoud, T. S.; Johnson, W. H.; Ebelt, N. D.; Piserchio, A.; Zamora-Olivares, D.; Van Ravenstein, S. X.; Pridgen, J. R.; Edupuganti, R.; Sammons, R.; Cano, M.; Warthaka, M.; Harger, M.; Tavares, C. D. J.; Park, J.; Radwan, M. F.; Ren, P.; Anslyn, E. V.; Tsai, K. Y.; Ghose, R.; Dalby, K. N., Modulating multi-functional ERK complexes by covalent targeting of a recruitment site in vivo. *Nat Commun* **2019,** *10* (1), 5232.
8. Neupane, K.; Manuel, A. P.; Woodside, M. T., Protein folding trajectories can be described quantitatively by one-dimensional diffusion over measured energy landscapes. *Nature Physics* **2016,** *12* (7), 700-703.
9. Lonsdale, R.; Ranaghan, K. E.; Mulholland, A. J., Computational enzymology. *Chemical Communications* **2010,** *46* (14), 2354-2372.
10. Qiu, Y.; Smith, D.; Boothroyd, S.; Jang, H.; Wagner, J.; Bannan, C. C.; Gokey, T.; Lim, V. T.; Stern, C.; Rizzi, A.; Lucas, X.; Tjanaka, B.; Shirts, M. R.; Gilson, M.; Chodera, J.; Bayly, C. I.; Mobley, D.; Wang, L.-P., Development and benchmarking of open force field v1. 0.0, the parsley small molecule force field. *ChemRxiv* **2021**.
11. Maier, J. A.; Martinez, C.; Kasavajhala, K.; Wickstrom, L.; Hauser, K. E.; Simmerling, C., ff14SB: Improving the Accuracy of Protein Side Chain and Backbone Parameters from ff99SB. *Journal of Chemical Theory and Computation* **2015,** *11* (8), 3696-3713.
12. Debiec, K. T.; Cerutti, D. S.; Baker, L. R.; Gronenborn, A. M.; Case, D. A.; Chong, L. T., Further along the Road Less Traveled: AMBER ff15ipq, an Original Protein Force Field Built on a Self-Consistent Physical Model. *Journal of Chemical Theory and Computation* **2016,** *12* (8), 3926-3947.
13. Wang, L.-P.; McKiernan, K. A.; Gomes, J.; Beauchamp, K. A.; Head-Gordon, T.; Rice, J. E.; Swope, W. C.; Martínez, T. J.; Pande, V. S., Building a More Predictive Protein Force Field: A Systematic and Reproducible Route to AMBER-FB15. *The Journal of Physical Chemistry B* **2017,** *121* (16), 4023-4039.
14. Salomon-Ferrer, R.; Case, D. A.; Walker, R. C., An overview of the Amber biomolecular simulation package. *WIREs Computational Molecular Science* **2013,** *3* (2), 198-210.

15. Jorgensen, W. L.; Maxwell, D. S.; Tirado-Rives, J., Development and Testing of the OPLS All-Atom Force Field on Conformational Energetics and Properties of Organic Liquids. *Journal of the American Chemical Society* **1996,** *118* (45), 11225-11236.
16. 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.
17. Jorgensen, W. L.; Tirado-Rives, J., The OPLS [optimized potentials for liquid simulations] potential functions for proteins, energy minimizations for crystals of cyclic peptides and crambin. *Journal of the American Chemical Society* **1988,** *110* (6), 1657-1666.
18. MacKerell, A. D.; Bashford, D.; Bellott, M.; Dunbrack, R. L.; Evanseck, J. D.; Field, M. J.; Fischer, S.; Gao, J.; Guo, H.; Ha, S.; Joseph-McCarthy, D.; Kuchnir, L.; Kuczera, K.; Lau, F. T. K.; Mattos, C.; Michnick, S.; Ngo, T.; Nguyen, D. T.; Prodhom, B.; Reiher, W. E.; Roux, B.; Schlenkrich, M.; Smith, J. C.; Stote, R.; Straub, J.; Watanabe, M.; Wiórkiewicz-Kuczera, J.; Yin, D.; Karplus, M., All-Atom Empirical Potential for Molecular Modeling and Dynamics Studies of Proteins. *J. Phys. Chem. B* **1998,** *102* (18), 3586-3616.
19. Vanommeslaeghe, K.; Hatcher, E.; Acharya, C.; Kundu, S.; Zhong, S.; Shim, J.; Darian, E.; Guvench, O.; Lopes, P.; Vorobyov, I.; Mackerell, A. D., Jr., 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.
20. Lopes, P. E.; Guvench, O.; MacKerell, A. D., Jr., Current status of protein force fields for molecular dynamics simulations. *Methods Mol Biol* **2015,** *1215*, 47-71.
21. Ren, P.; Ponder, J. W., Polarizable Atomic Multipole Water Model for Molecular Mechanics Simulation. *The Journal of Physical Chemistry B* **2003,** *107* (24), 5933-5947.
22. Liu, C.; Piquemal, J.-P.; Ren, P., AMOEBA+ Classical Potential for Modeling Molecular Interactions. *J. Chem. Theory Comput.* **2019,** *15* (7), 4122-4139.
23. Yang, X.; Liu, C.; Walker, B. D.; Ren, P., Accurate description of molecular dipole surface with charge flux implemented for molecular mechanics. *The Journal of Chemical Physics* **2020,** *153* (6), 064103.
24. Allinger, N. L.; Yuh, Y. H.; Lii, J. H., Molecular mechanics. The MM3 force field for hydrocarbons. 1. *Journal of the American Chemical Society* **1989,** *111* (23), 8551-8566.
25. Allinger, N. L.; Chen, K.; Lii, J.-H., An improved force field (MM4) for saturated hydrocarbons. *Journal of Computational Chemistry* **1996,** *17* (5-6), 642-668.
26. Maple, J. R.; Dinur, U.; Hagler, A. T., Derivation of force fields for molecular mechanics and dynamics from ab initio energy surfaces. *Proc Natl Acad Sci U S A* **1988,** *85* (15), 5350-5354.
27. Lii, J. H.; Allinger, N. L., Molecular mechanics. The MM3 force field for hydrocarbons. 2. Vibrational frequencies and thermodynamics. *Journal of the American Chemical Society* **1989,** *111* (23), 8566-8575.
28. Maple, J. R.; Hwang, M. J.; Jalkanen, K. J.; Stockfisch, T. P.; Hagler, A. T., Derivation of class II force fields: V. Quantum force field for amides, peptides, and related compounds. *Journal of Computational Chemistry* **1998,** *19* (4), 430-458.
29. Palmo, K.; Mirkin, N. G.; Pietila, L. O.; Krimm, S., Spectroscopically determined force fields for macromolecules. 1. n-Alkane chains. *Macromolecules* **1993,** *26* (25), 6831-6840.

30. Hagler, A. T., Quantum Derivative Fitting and Biomolecular Force Fields: Functional Form, Coupling Terms, Charge Flux, Nonbond Anharmonicity, and Individual Dihedral Potentials. *Journal of Chemical Theory and Computation* **2015,** *11* (12), 5555-5572.
31. Allen, A. E. A.; Payne, M. C.; Cole, D. J., Harmonic Force Constants for Molecular Mechanics Force Fields via Hessian Matrix Projection. *Journal of Chemical Theory and Computation* **2018,** *14* (1), 274-281.
32. Jensen, F., Introduction to Computational Chemistry. *Wiley* **2017**, 28.
33. Nevins, N.; Lii, J.-H.; Allinger, N. L., Molecular mechanics (MM4) calculations on conjugated hydrocarbons. *Journal of Computational Chemistry* **1996,** *17* (5-6), 695-729.
34. Seminario, J. M., Calculation of intramolecular force fields from second-derivative tensors. *International Journal of Quantum Chemistry* **1996,** *60* (7), 1271-1277.
35. Robertson, M. J.; Tirado-Rives, J.; Jorgensen, W. L., Improved Peptide and Protein Torsional Energetics with the OPLS-AA Force Field. *Journal of Chemical Theory and Computation* **2015,** *11* (7), 3499-3509.
36. Rackers, J. A.; Wang, Z.; Lu, C.; Laury, M. L.; Lagardère, L.; Schnieders, M. J.; Piquemal, J.-P.; Ren, P.; Ponder, J. W., Tinker 8: Software Tools for Molecular Design. *Journal of Chemical Theory and Computation* **2018,** *14* (10), 5273-5289.
37. Wu, J. C.; Chattree, G.; Ren, P., Automation of AMOEBA polarizable force field parameterization for small molecules. *Theor. Chem. Acc.* **2012,** *131* (3), 1138.
38. Frisch, M. J. T., G. W.; Schlegel, H. B.; Scuseria, G. E.;; Robb, M. A. C., J. R.; Scalmani, G.; Barone, V.; Mennucci,; B.; Petersson, G. A. N., H.; Caricato, M.; Li, X.; Hratchian, H.; P.; Izmaylov, A. F. B., J.; Zheng, G.; Sonnenberg, J. L.; Hada, M.;; Ehara, M. T., K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima,; T.; Honda, Y. K., O.; Nakai, H.; Vreven, T.; Montgomery, J. A., Jr.;; Peralta, J. E. O., F.; Bearpark, M.; Heyd, J. J.; Brothers, E.; Kudin,; K. N.; Staroverov, V. N. K., R.; Normand, J.; Raghavachari, K.;; Rendell, A. B., J. C.; Iyengar, S. S.; Tomasi, J.; Cossi, M.; Rega,; N.; Millam, J. M. K., M.; Knox, J. E.; Cross, J. B.; Bakken, V.;; Adamo, C. J., J.; Gomperts, R.; Stratmann, R. E.; Yazyev, O.;; Austin, A. J. C., R.; Pomelli, C.; Ochterski, J. W.; Martin, R. L.;; Morokuma, K. Z., V. G.; Voth, G. A.; Salvador, P.;; Dannenberg, J. J. D., S.; Daniels, A. D.; Farkas, Ö.;; Foresman, J. B. O., J. V.; Cioslowski, J.; Fox, D. J., Gaussian 09. *Gaussian, Inc., Wallingford CT, 2009* **2009**.
39. Hehre, W. J., *A guide to molecular mechanics and quantum chemical calculations*. Wavefunction Irvine, CA: 2003; Vol. 2.
40. Johnson, R. D., III. NIST Computational Chemistry Comparison and Benchmark Database, accessed 2020. *http://cccbdb.nist.gov/*.
41. Irikura, K. K.; Johnson, R. D.; Kacker, R. N., Uncertainties in Scaling Factors for ab Initio Vibrational Frequencies. *The Journal of Physical Chemistry A* **2005,** *109* (37), 8430-8437.
42. Virtanen, P.; Gommers, R.; Oliphant, T. E.; Haberland, M.; Reddy, T.; Cournapeau, D.; Burovski, E.; Peterson, P.; Weckesser, W.; Bright, J.; van der Walt, S. J.; Brett, M.; Wilson, J.; Millman, K. J.; Mayorov, N.; Nelson, A. R. J.; Jones, E.; Kern, R.; Larson, E.; Carey, C. J.; Polat, İ.; Feng, Y.; Moore, E. W.; VanderPlas, J.; Laxalde, D.; Perktold, J.; Cimrman, R.; Henriksen, I.; Quintero, E. A.; Harris, C. R.; Archibald, A. M.; Ribeiro, A. H.; Pedregosa, F.; van Mulbregt, P.; Vijaykumar, A.; Bardelli, A. P.; Rothberg, A.; Hilboll, A.; Kloeckner, A.; Scopatz, A.; Lee, A.; Rokem, A.; Woods, C. N.; Fulton, C.; Masson, C.; Häggström, C.; Fitzgerald, C.; Nicholson, D. A.; Hagen, D. R.; Pasechnik, D. V.; Olivetti, E.; Martin, E.; Wieser, E.; Silva, F.; Lenders, F.; Wilhelm, F.; Young, G.; Price, G. A.; Ingold, G.-L.; Allen, G. E.; Lee, G. R.; Audren, H.; Probst, I.; Dietrich, J. P.; Silterra, J.; Webber, J. T.; Slavič, J.;

Nothman, J.; Buchner, J.; Kulick, J.; Schönberger, J. L.; de Miranda Cardoso, J. V.; Reimer, J.; Harrington, J.; Rodríguez, J. L. C.; Nunez-Iglesias, J.; Kuczynski, J.; Tritz, K.; Thoma, M.; Newville, M.; Kümmerer, M.; Bolingbroke, M.; Tartre, M.; Pak, M.; Smith, N. J.; Nowaczyk, N.; Shebanov, N.; Pavlyk, O.; Brodtkorb, P. A.; Lee, P.; McGibbon, R. T.; Feldbauer, R.; Lewis, S.; Tygier, S.; Sievert, S.; Vigna, S.; Peterson, S.; More, S.; Pudlik, T.; Oshima, T.; Pingel, T. J.; Robitaille, T. P.; Spura, T.; Jones, T. R.; Cera, T.; Leslie, T.; Zito, T.; Krauss, T.; Upadhyay, U.; Halchenko, Y. O.; Vázquez-Baeza, Y.; SciPy, C., SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nature Methods* **2020,** *17* (3), 261-272.

43. 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. *J. Chem. Theory Comput.* **2017,** *13* (7), 3185-3197.

44. Wang, R.; Ozhgibesov, M.; Hirao, H., Partial hessian fitting for determining force constant parameters in molecular mechanics. *J Comput Chem* **2016,** *37* (26), 2349-59.

45. Ryckaert, J.-P.; Ciccotti, G.; Berendsen, H. J. C., Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes. *Journal of Computational Physics* **1977,** *23* (3), 327-341.

46. Ciccotti, G.; Ryckaert, J. P., Molecular dynamics simulation of rigid molecules. *Computer Physics Reports* **1986,** *4* (6), 346-392.

# Supplementary Information

## High Order Ab Initio Valence Force Field with Chemical Pattern Based Parameter Assignment

*Xudong Yang, Chengwen Liu, Pengyu Ren*[*]

Department of Biomedical Engineering, The University of Texas at Austin, Austin, TX 78712, USA

* Corresponding author. Email: pren@utexas.edu

**Table S1. The molecules in the training set**

The training set has been shown with their name, their classified group, and the SMILE strings (445 molecules). We have 7 groups: (1).alkanes and aliphatic rings; (2).molecules comprised of only single bonds; (3).molecules with double bonds or triple bonds; (4).benzene derivatives; (5).heterocyclics; (6).bio-fragments; (7).combinations of multiple aromatic cycles.

| Name | Group | SMILE |
|---|---|---|
| Water | 0 | O |
| Methane | 1 | C |
| Ammonia | 2 | N |
| Methylamine | 2 | CN |
| Methanol | 2 | CO |
| N-Methylacetamide | 3 | CC(=O)NC |
| Acetic acid | 3 | CC(=O)O |
| Acetamide | 3 | CC(=O)N |
| Uracil | 6 | C1=CNC(=O)NC1=O |
| Pyridine | 5 | C1=CC=NC=C1 |
| Benzene | 4 | C1=CC=CC=C1 |
| Ethene | 1 | C=C |
| Pentane | 1 | CCCCC |
| Neopentane | 1 | CC(C)(C)C |
| Cyclopentane | 1 | C1CCCC1 |
| Methane thiol | 2 | CS |
| DMSO | 3 | CS(=O)C |
| Fluorobenzene | 4 | C1=CC=C(C=C1)F |
| Chlorobenzene | 4 | C1=CC=C(C=C1)Cl |
| Bromobenzene | 4 | C1=CC=C(C=C1)Br |
| DimethylSulfide | 2 | CSC |
| Imidazole | 5 | C1=CN=CN1 |
| Pyrrolidine | 5 | C1CCNC1 |
| Phenol | 4 | C1=CC=C(C=C1)O |
| Indole | 5 | C1=CC=C2C(=C1)C=CN2 |
| Difluoromethane | 2 | C(F)F |
| Dichloromethane | 2 | C(Cl)Cl |
| Dibromomethane | 2 | C(Br)Br |
| Fluoromethane | 2 | CF |
| Chloromethane | 2 | CCl |
| Bromomethane | 2 | CBr |
| Hydrogenphosphate | 6 | OP(=O)([O-])[O-] |
| Dihydrogen phosphate | 6 | OP(=O)(O)[O-] |
| Phosphorous acid | 6 | OP(O)O |
| Methylammonium | 2 | C[NH3+] |
| 1H-imidazol-3-ium | 5 | C1=C[NH+]=CN1 |

| | | |
|---|---|---|
| Guanidine | 2 | C(=N)(N)N |
| Acetate | 3 | CC(=O)[O-] |
| Formimidamide | 3 | C(=N)N |
| Formamide | 3 | C(=O)N |
| Formicacid | 3 | C(=O)O |
| Ethane | 1 | CC |
| Ethanol | 2 | CCO |
| 7-Methylindole | 5 | CC1=C2C(=CC=C1)C=CN2 |
| Methylguanidine | 2 | CN=C(N)N |
| 4-Methylimidazole | 5 | CC1=CN=CN1 |
| 2-Propanol | 2 | CC(C)O |
| Ethyne | 1 | C#C |
| 2-Aminopyridine | 5 | C1=CC=NC(=C1)N |
| 2-Pyridone | 5 | C1=CC(=O)NC=C1 |
| Hydrogen sulfide | 2 | S |
| Pyrazine | 5 | C1=CN=CC=N1 |
| Thymine | 6 | CC1=CNC(=O)NC1=O |
| Adenine | 6 | C1=NC2=NC=NC(=C2N1)N |
| Formaldehyde | 3 | C=O |
| Acetonitrile | 3 | CC#N |
| Dimethylamine | 2 | CNC |
| Isobutane | 1 | CC(C)C |
| Trimethylamine | 2 | CN(C)C |
| Trimethylammonium | 2 | C[NH+](C)C |
| Ethane thiol | 2 | CCS |
| Methyl benzene | 4 | CC1=CC=CC=C1 |
| Dimethylammonium | 2 | C[NH2+]C |
| Acetone | 3 | CC(=O)C |
| Cyclohexane | 1 | C1CCCCC1 |
| Dimethylether | 2 | COC |
| Naphthalene | 4 | C1=CC=C2C=CC=CC2=C1 |
| Guanine | 6 | C1=NC2=C(N1)C(=O)NC(=N2)N |
| Diethylether | 2 | CCOCC |
| Acetaldehyde | 3 | CC=O |
| Cytosine | 6 | C1=C(NC(=O)N=C1)N |
| Phenolate | 4 | C1=CC=C(C=C1)[O-] |
| Aniline | 4 | C1=CC=C(C=C1)N |
| Tetrafluoromethane | 2 | C(F)(F)(F)F |
| Tetrachloromethane | 2 | C(Cl)(Cl)(Cl)Cl |
| Trifluoromethane | 2 | C(F)(F)F |
| Trichloromethane | 2 | C(Cl)(Cl)Cl |
| Tribromomethane | 2 | C(Br)(Br)Br |
| Acetic Anhydride | 3 | CC(=O)OC(=O)C |

| | | |
|---|---|---|
| Carbamic acid | 3 | C(=O)(N)O |
| 1-Butylamine | 2 | CCCCN |
| 3-Methylindole | 5 | CC1=CNC2=CC=CC=C12 |
| 4-Methylimidazole | 5 | CC1=CN=CN1 |
| Ethyl methyl sulfide | 2 | CCSC |
| 1-Propylguanidinium | 2 | CCC[NH+]=C(N)N |
| Propionamide | 3 | CCC(=O)N |
| Propionic acid | 3 | CCC(=O)O |
| Diethyl sulfone | 3 | CCS(=O)(=O)CC |
| Diethyl disulfide | 2 | CCSSCC |
| Formate | 3 | C(=O)[O-] |
| Propionyl chloride | 3 | CCC(=O)Cl |
| Propionyl fluoride | 3 | CCC(=O)F |
| Propionyl bromide | 3 | CCC(=O)Br |
| Diethyl carbonate | 3 | CCOC(=O)OCC |
| Ethyl hydrogen carbonate | 3 | CCOC(=O)O |
| Diethoxymethane | 2 | CCOCOCC |
| Benzonitrile | 4 | C1=CC=C(C=C1)C#N |
| Phosphane | 6 | P |
| Propionate | 3 | CCC(=O)[O-] |
| Propionaldehyde | 3 | CCC=O |
| 2-Butanone | 3 | CCC(=O)C |
| N-Ethylformamide | 3 | CCNC=O |
| Dimethyl formamide | 3 | CN(C)C=O |
| 1,3-Benzodioxole | 4 | O1c2c(OC1)cccc2 |
| Dioxolane | 2 | C1COCO1 |
| Ethanamine | 3 | CCN |
| Ethanimine | 3 | CC=N |
| Methanamine | 3 | CN |
| 2-Propanimine | 3 | CC(=N)C |
| Ethyl Acetoacetate | 3 | CCOC(=O)CC(=O)C |
| Oxalic acid | 3 | C(=O)(C(=O)O)O |
| Hydrogen oxalate | 3 | C(=O)(C(=O)[O-])O |
| Oxalate | 3 | C(=O)(C(=O)[O-])[O-] |
| Malonic acid | 3 | C(C(=O)O)C(=O)O |
| Hydrogen malonate | 3 | C(C(=O)O)C(=O)[O-] |
| Malonate | 3 | C(C(=O)[O-])C(=O)[O-] |
| 2-Oxobutanoic Acid | 3 | CCC(=O)C(=O)O |
| 2-Oxobutanoate | 3 | CCC(=O)C(=O)[O-] |
| Dipropionamide | 3 | CCC(=O)NC(=O)CC |
| Nitroethane | 3 | CC[N+](=O)[O-] |
| Nitrobenzene | 4 | C1=CC=C(C=C1)[N+](=O)[O-] |
| Ethyl N-ethyl-N- | 3 | CCN(C)C(=O)OCC |

| | | |
|---|---|---|
| methylcarbamate | | |
| Methylcarbamic acid | 3 | CNC(=O)O |
| Carbamate | 3 | C(=O)(N)[O-] |
| Ethyl nitrite | 3 | CCON=O |
| Benzenediazonium | 4 | C1=CC=C(C=C1)[N+]#N |
| Azobenzene | 4 | C1=CC=C(C=C1)N=NC2=CC=CC=C2 |
| 1-[(S)-Methylsulfinyl]ethane | 3 | [S@@](=O)(CC)C |
| Ethane sulfonic acid | 7 | CCS(=O)(=O)O |
| Ethane sulfonate | 7 | CCS(=O)(=O)O |
| Benzene sulfonic acid | 7 | C1=CC=C(C=C1)S(=O)(=O)O |
| Benzene sulfonate | 7 | C1=CC=C(C=C1)S(=O)(=O)[O-] |
| Ethyl acetate | 3 | CCOC(=O)C |
| Ethyl methanesulfonate | 7 | CCOS(=O)(=O)C |
| Ethyl thiocyanate | 3 | CCSC#N |
| Phenyl thiocyanate | 4 | C1=CC=C(C=C1)SC#N |
| Butadiene | 1 | C=CC=C |
| Acrolein | 3 | C=CC=O |
| Propionaldehyde oxime | 3 | CCC=NO |
| Acetone oxime | 3 | CC(=NO)C |
| Benzyl alcohol | 4 | C1=CC=C(C=C1)CO |
| Ethylbenzene | 4 | CCC1=CC=CC=C1 |
| Benzylamine | 4 | C1=CC=C(C=C1)CN |
| Benzaldehyde | 4 | C1=CC=C(C=C1)C=O |
| Benzoic acid | 4 | C1=CC=C(C=C1)C(=O)O |
| Benzoate | 4 | C1=CC=C(C=C1)C(=O)[O-] |
| AcrylicAcid | 3 | C=CC(=O)O |
| Acrylate | 3 | C=CC(=O)[O-] |
| L-Alanine | 6 | CC(C(=O)O)N |
| Pyrophosphate ion | 6 | [O-]P(=O)([O-])OP(=O)([O-])[O-] |
| Phenylazanium | 4 | C1=CC=C(C=C1)[NH3+] |
| Glucose | 6 | C(C1C(C(C(C(O1)O)O)O)O)O |
| Ribose phosphate | 6 | C(C(C(C(C=O)O)O)O)OP(=O)(O)O |
| Glyoxal | 3 | C(=O)C=O |
| 2-Methyl-1-butene | 1 | CCC(=C)C |
| (E)-but-2-ene | 1 | CC=CC |
| Glycolic acid | 3 | C(C(=O)O)O |
| Glycolate | 3 | C(C(=O)[O-])O |
| Malonaldehyde | 3 | C(C=O)C=O |
| Urea | 3 | C(=O)(N)N |
| Fluoroethane | 2 | CCF |
| Chloroethane | 2 | CCCl |
| Bromoethane | 2 | CCBr |
| 2,3-Dimethylbutane | 1 | CC(C)C(C)C |

| | | |
|---|---|---|
| 2,2-Dimethylbutane | 1 | CCC(C)(C)C |
| 3-Ethylpentane | 1 | CCC(CC)CC |
| 2,2,3,3-Tetramethylbutane | 1 | CC(C)(C)C(C)(C)C |
| Propane | 1 | CCC |
| Glycerol | 2 | C(C(CO)O)O |
| 1,3-Diaminopropane | 2 | C(CN)CN |
| 1,3-Propanedithiol | 2 | C(CS)CS |
| 1,2,3-Trifluoropropane | 2 | C(C(CF)F)F |
| 1,2,3-Trichloropropane | 2 | C(C(CCl)Cl)Cl |
| 1,2,3-Tribromopropane | 2 | C(C(CBr)Br)Br |
| 3-Aminopropanol | 2 | C(CN)CO |
| 3-Mercaptopropanol | 2 | C(CO)CS |
| 3-Fluoropropanol | 2 | C(CO)CF |
| 3-Chloropropanol | 2 | C(CO)CCl |
| Diethylammonium | 2 | CC[NH2+]CC |
| Diethylamine | 2 | CCNCC |
| Diisopropylether | 2 | CC(C)OC(C)C |
| Diisopropyl sulfide | 3 | CC(C)SC(C)C |
| Dipropyl disulfide | 3 | CCCSSCCC |
| Citric acid | 3 | C(C(=O)O)C(CC(=O)O)(C(=O)O)O |
| Alpha-Ketoglutarate | 3 | C(CC(=O)[O-])C(=O)C(=O)[O-] |
| Succinate | 3 | C(CC(=O)[O-])C(=O)[O-] |
| Malic acid | 3 | C(C(C(=O)O)O)C(=O)O |
| Cysteine | 6 | C(C(C(=O)O)N)S |
| Ethyl propionate | 3 | CCC(=O)OCC |
| Arginine | 6 | C(CC(C(=O)O)N)CN=C(N)N |
| Glutamine | 6 | C(CC(=O)N)C(C(=O)O)N |
| Diisopropyl sulfoxide | 3 | CC(C)S(=O)C(C)C |
| Isopropyl sulfone | 3 | CC(C)S(=O)(=O)C(C)C |
| Isobutyric anhydride | 3 | CC(C)C(=O)OC(=O)C(C)C |
| Diisopropyl carbonate | 3 | CC(C)OC(=O)OC(C)C |
| N-Methyl-1-butanimine | 3 | CCCC=NC |
| 2-Methyl-1-nitropropane | 3 | CC(C)C[N+](=O)[O-] |
| Fumaric acid | 3 | C(=CC(=O)O)C(=O)O |
| Hexa-1,3,5-triene | 1 | C=CC=CC=C |
| 3-Pentanone | 3 | CCC(=O)CC |
| Butyraldehyde | 3 | CCCC=O |
| 2-Butyne | 1 | CC#CC |
| Butyronitrile | 3 | CCCC#N |
| Hexamethylbenzene | 4 | CC1=C(C(=C(C(=C1C)C)C)C)C |
| Salicylic acid | 3 | C1=CC=C(C(=C1)C(=O)O)O |
| Cumene | 4 | CC(C)C1=CC=CC=C1 |
| p-Cresol | 4 | CC1=CC=C(C=C1)O |

| | | |
|---|---|---|
| 3-Methylanilinium | 4 | CC1=CC(=CC=C1)[NH3+] |
| 3-Methylbenzaldehyde | 4 | CC1=CC(=CC=C1)C=O |
| p-Toluidine | 4 | CC1=CC=C(C=C1)N |
| 3-Nitrotoluene | 4 | CC1=CC(=CC=C1)[N+](=O)[O-] |
| 4-Fluorophenol | 4 | C1=CC(=CC=C1O)F |
| 1,3,5-Trichlorobenzene | 4 | C1=C(C=C(C=C1Cl)Cl)Cl |
| 1,3-Dibromobenzene | 4 | C1=CC(=CC(=C1)Br)Br |
| P-dioxane | 5 | C1COCCO1 |
| D-Galactose | 6 | C(C1C(C(C(C(O1)O)O)O)O)O |
| THF | 5 | C1CCOC1 |
| 4-Aminopyridine | 5 | C1=CN=CC=C1N |
| 1-Propanol | 2 | CCCO |
| 1,3-Propanediol | 2 | C(O)CC(O) |
| Ethylamine | 2 | CCN |
| Propylamine | 2 | CCCN |
| 1-Methylcytosine | 6 | CN1C=CC(=NC1=O)N |
| 1-Methylthymine | 6 | CC1=CN(C(=O)NC1=O)C |
| 1-Methyluracil | 6 | CN1C=CC(=O)NC1=O |
| 9-Methyladenine | 6 | CN1C=NC2=C(N=CN=C21)N |
| 9-Methylguanine | 6 | CN1C=NC2=C1N=C(NC2=O)N |
| Cytidine | 6 | C1=CN(C(=O)N=C1N)C2C(C(C(O2)CO)O)O |
| Deoxycytidine | 6 | C1C(C(OC1N2C=CC(=NC2=O)N)CO)O |
| Dimethylphosphate | 6 | COP(=O)(O)OC |
| Methylphosphate | 6 | COP(=O)([O-])[O-] |
| Ribose | 6 | C1C(C(C(C(O1)O)O)O)O |
| Trimethylphosphate | 6 | COP(=O)(OC)OC |
| Propan-2-amine | 2 | CC(C)N |
| 2-chloropropane | 2 | CC(C)Cl |
| 2-methylpropan-2-ol | 2 | CC(C)(C)O |
| 2-methoxypropane | 2 | CC(C)OC |
| 1,1,1-trifluoroethane | 2 | CC(F)(F)F |
| 2-methylpropanal | 3 | CC(C)C=O |
| ethyl(trimethyl)azanium | 2 | CC[N+](C)(C)C |
| (2R)-2-aminopropanoic acid | 3 | CC(C(=O)O)N |
| 2-methylpropanoic acid | 3 | CC(C)C(=O)O |
| methyl propanoate | 3 | CCC(=O)OC |
| hex-3-yne | 1 | CCC#CCC |
| 2-ethoxypropane | 2 | CCOC(C)C |
| (2R)-2-hydroxypropanoic acid | 3 | CC(C(=O)O)O |
| (2S)-2-hydroxypropanoic | 3 | CC(C(=O)O)O |

| | | |
|---|---|---|
| acid | | |
| 3-methylbutan-2-one | 3 | CC(C)C(=O)C |
| (E)-pent-2-enal | 3 | CCC=CC=O |
| (2S)-3-methylbut-3-en-2-ol | 3 | CC(C(=C)C)O |
| (E,2R)-hex-3-en-2-ol | 3 | CCC=CC(C)O |
| (E,2S)-hex-3-en-2-ol | 3 | CCC=CC(C)O |
| Ethyl dihydrogen phosphate | 6 | CCOP(=O)(O)O |
| (Z)-1-ethoxybut-1-ene | 3 | CCC=COCC |
| N-ethylpropanamide | 3 | CCC(=O)NCC |
| S-ethyl propanethioate | 3 | CCC(=O)SCC |
| propan-2-yl acetate | 3 | CC(C)OC(=O)C |
| (E)-pent-2-enoic acid | 3 | CCC=CC(=O)O |
| 2-methylpentan-3-one | 3 | CCC(=O)C(C)C |
| (Z)-pent-2-enoic acid | 3 | CCC=CC(=O)O |
| propan-2-yl propanoate | 3 | CCC(=O)OC(C)C |
| N-propan-2-ylpropanamide | 3 | CCC(=O)NC(C)C |
| (2S)-N-ethyl-2-hydroxypropanamide | 3 | CCNC(=O)C(C)O |
| (2R)-N-ethyl-2-hydroxypropanamide | 3 | CCNC(=O)C(C)O |
| ethyl 2-methylpropanoate | 3 | CCOC(=O)C(C)C |
| (E)-4-oxohex-2-enal | 3 | CCC(=O)C=CC=O |
| 3-oxopentanoic acid | 3 | CCC(=O)CC(=O)O |
| octa-3,5-diyne | 1 | CCC#CC#CCC |
| (E)-hex-3-enoic acid | 3 | CCC=CCC(=O)O |
| diethyl hydrogen phosphate | 6 | CCOP(=O)(O)OCC |
| (2S)-2-hydroxy-N-propan-2-ylpropanamide | 3 | CC(C)NC(=O)C(C)O |
| (2R)-2-hydroxy-N-propan-2-ylpropanamide | 3 | CC(C)NC(=O)C(C)O |
| 4-ethylphenol | 3 | CCC1=CC=C(C=C1)O |
| (2S,3E,5E)-octa-3,5-dien-2-ol | 3 | CCC=CC=CC(C)O |
| (2R,3E,5E)-octa-3,5-dien-2-ol | 3 | CCC=CC=CC(C)O |
| (E)-4-oxohex-2-enoic acid | 3 | CCC(=O)C=CC(=O)O |
| 2-(propanoylamino)acetic acid | 3 | CCC(=O)NCC(=O)O |
| 5-ethylbenzene-1,3-diol | 4 | CCC1=CC(=CC(=C1)O)O |
| (2R)-2-(propanoylamino)propanoic acid | 3 | CCC(=O)NC(C)C(=O)O |
| (2S)-2- | 3 | CCC(=O)NC(C)C(=O)O |

| | | |
|---|---|---|
| (propanoylamino)propanoic acid | | |
| S-ethyl 3-oxopentanethioate | 3 | CCC(=O)CC(=O)SCC |
| 2,5-diethyl-3,4-dimethylfuran | 5 | CCC1=C(C(=C(O1)CC)C)C |
| [(E)-3-methylpent-2-enyl] propanoate | 3 | CCC(=CCOC(=O)CC)C |
| ethyl [(2R,3R,5S,6R)-2,3,4,5,6-pentahydroxycyclohexyl] hydrogen phosphate | 6 | CCOP(=O)(O)OC1C(C(C(C(C1O)O)O)O)O |
| Cyclopropane | 1 | C1CC1 |
| Cyclobutane | 1 | C1CCC1 |
| Methyldiphosphate | 6 | CO[P](O[P](=O)([O-])[O-])(=O)[O-] |
| Methyltriphosphate | 6 | CO[P](O[P](O[P]([O-])(=O)[O-])(=O)[O-])(=O)[O-] |
| D-Alaninate | 6 | CC(C(=O)[O-])N |
| D-Alaninium | 6 | CC(C(=O)O)[N+] |
| D-Alanine zwitterion | 6 | CC(C(=O)[O-])[N+] |
| 2-methylpropanoate | 3 | CC(C)C(=O)[O-] |
| (2R)-2-hydroxypropanoate | 3 | CC(C(=O)[O-])O |
| (2S)-2-hydroxypropanoate | 3 | CC(C(=O)[O-])O |
| Ethyl hydrogen phosphate | 6 | CCOP(=O)([O-])O |
| Ethyl phosphate | 6 | CCOP(=O)([O-])[O-] |
| (E)-pent-2-enoate | 3 | CCC=CC(=O)[O-] |
| (Z)-pent-2-enoate | 3 | CCC=CC(=O)[O-] |
| 3-oxopentanoate | 3 | CCC(=O)CC(=O)[O-] |
| (E)-hex-3-enoate | 3 | CCC=CCC(=O)[O-] |
| Diethyl phosphate | 6 | CCOP(=O)([O-])OCC |
| 2-(propanoylamino)acetate | 3 | CCC(=O)NCC(=O)[O-] |
| (2R)-2-(propanoylamino)propanoate | 3 | CCC(=O)NC(C)C(=O)[O-] |
| Methanethiolate | 2 | C[S-] |
| Thioethoxide | 2 | CC[S-] |
| Propane-2-thiolate | 2 | CC([S-])C |
| N,N-Diethylethanesulfonamide | 3 | CCN(CC)S(=O)(=O)CC |
| Methanesulfonamide | 3 | CS(=O)(=O)N |
| Propan-2-aminius | 2 | CC(C)[N+] |
| D-deoxyribose | 6 | O1[C@@H]([C@@H](O[H])C[C@@H]1O[H])CO[H] |

| | | |
|---|---|---|
| CH3I | 2 | IC |
| CH2I2 | 2 | ICI |
| CHI3 | 2 | IC(I)I |
| CI4 | 2 | IC(I)(I)I |
| CBr4 | 2 | BrC(Br)(Br)Br |
| Benzotrifluoride | 4 | C(C1=CC=CC=C1)(F)(F)F |
| Benzotrichloride | 4 | C(C1=CC=CC=C1)(Cl)(Cl)Cl |
| 111-trichloropropane | 2 | ClC(Cl)(Cl)CC |
| 111-trifluoropropane | 2 | FC(F)(F)CC |
| 222-trifluoroethylbenzene | 4 | FC(F)(F)CC1=CC=CC=C1 |
| 222-trichloroethylbenzene | 4 | ClC(Cl)(Cl)CC1=CC=CC=C1 |
| Tribromoethylbenzene | 4 | C1(=CC=CC=C1)CC(Br)(Br)Br |
| Benzyliodide | 4 | C1(=CC=CC=C1)CI |
| Iodobenzene | 4 | C1(=CC=CC=C1)I |
| 1,1,1-Tribromopropane | 2 | C(CC)(Br)(Br)Br |
| 1,1,1,3,3,3-Hexafluoropropane | 2 | C(CC(F)(F)F)(F)(F)F |
| 1,1,1,3,3,3-Hexachloropropane | 2 | C(CC(Cl)(Cl)Cl)(Cl)(Cl)Cl |
| 1-Phenylpiperazinium | 4 | C1=CC=C(C=C1)N2CC[N+]CC2 |
| 1-Ethyl-1-methylpiperidin-1-ium | 2 | C1CCCC[N+]1(C)CC |
| Tribromomethylbenzene | 4 | C(C1=CC=CC=C1)(Br)(Br)Br |
| Biphenyl | 7 | C1=CC=CC=C1C2=CC=CC=C2 |
| Diphenylmethane | 7 | C1=CC=CC=C1CC2=CC=CC=C2 |
| 1,1-Diphenylethane | 7 | C1=CC=CC=C1C(C2=CC=CC=C2)C |
| 2-Phenoxypyridine | 7 | C1=CC=CC=C1OC2=CC=CC=N2 |
| 2-Anilinopyridine | 7 | C1=CC=CC=C1NC2=NC=CC=C2 |
| Phenyl-2-pyridyl-methylamine | 7 | C1=CC=CC=C1N(C)C2=NC=CC=C2 |
| Diphenylamine | 7 | C1=CC=CC=C1NC2=CC=CC=C2 |
| Caprolactam | 3 | C1(CCCCCN1)=O |
| N-Methylcaprolactam | 3 | C1(CCCCCN1C)=O |
| 1,3,4,5-Tetrahydro-2H-1-benzazepin-2-one | 4 | C2(CCCC1=CC=CC=C1N2[H])=O |
| 2,3,4,5-Tetrahydro-benzo[c]azepin-1-one | 4 | C2CCC1=CC=CC=C1C(N2[H])=O |
| 1-Methyl-1,3,4,5-tetrahydro-2H-1-benzazepin-2-one | 4 | C1=CC2=C(C=C1)CCCC(N2C)=O |
| 5-Methyl-1H-benzo[e][1,4]diazepin-2(3H)-one | 4 | C2(CN=C(C1=CC=CC=C1N2[H])C)=O |

| | | |
|---|---|---|
| Benzamide | 4 | C1=CC(=CC=C1)C(N)=O |
| 1-Benzoylpiperidine | 4 | C1=CC(=CC=C1)C(N2CCCCC2)=O |
| (2R)-1-Benzoyl-2alpha-methylpiperidine | 4 | C1=CC(=CC=C1)C(=O)N2[C@@H](CCCC2)C |
| 2-(1Methyl-Imidazol-2-yl)pyrimidine | 7 | C1=NC(=NC=C1)C2=[N](C=C[N]2)C |
| 2-(5-Methylpyrazol-1-yl)pyrimidine | 7 | C1=NC(=NC=C1)[N]2N=CC=C2C |
| 3-(2-Pyridyl)furan | 7 | C1=CC(=NC=C1)C2=COC=C2 |
| 2-(2-Pyridyl)thiazole | 7 | C1=CC(=NC=C1)C2=NC=CS2 |
| Phenetole | 4 | C1(=CC=CC=C1)OCC |
| 1-Methyl-2-phenyl-1H-pyrrole | 7 | [N]1(C(=CC=C1)C2=CC=CC=C2)C |
| 2-Phenyl-1H-pyrrole | 7 | [N]1(C(=CC=C1)C2=CC=CC=C2)[H] |
| 1-Phenylpyrazole | 7 | [N]1(N=CC=C1)C2=CC=CC=C2 |
| Methyl2-(methylamino)benzoate | 4 | O(C(=O)C1=CC=CC=C1N(C)[H])C |
| 2-Aminobenzophenone | 7 | O=C(C1=CC=CC=C1N([H])[H])C2=CC=CC=C2 |
| (2-(methylamino)phenyl)phenyl-Methanone | 7 | O=C(C1=CC=CC=C1N(C)[H])C2=CC=CC=C2 |
| Dimethylaminobenzophenone | 7 | O=C(C1=CC=CC=C1N(C)C)C2=CC=CC=C2 |
| 2-Aminobiphenyl | 7 | N(C1=CC=CC=C1C2=CC=CC=C2)([H])[H] |
| 2-(Pyridin-2-yl)aniline | 7 | N1=CC=CC=C1C2=CC=CC=C2N([H])[H] |
| N-Methyl-2-(2-pyridyl)aniline | 7 | N(C1=CC=CC=C1C2=CC=CC=N2)(C)[H] |
| N,N-Dimethyl-2-pyridin-2-ylaniline | 7 | N(C1=CC=CC=C1C2=CC=CC=N2)(C)C |
| Ethylcyclopropane | 1 | C1C(C1)CC |
| 1-(Hydroxymethyl)cyclopropanol | 2 | C1C(C1)(CO)O |
| N-Cyclopropylacetamide | 3 | C1C(C1)NC(=O)C |
| N-Methylcyclopropanecarboxamide | 3 | C1C(C1)C(=O)NC |
| Ethylcyclopropanecarboxylate | 3 | C1C(C1)C(=O)OCC |
| Cyclopropylpropanoate | 3 | C1C(C1)OC(CC)=O |
| Methyl1-fluorocyclopropanecarboxyl | 3 | C1C(C1)(C(=O)OC)F |

| | | |
|---|---|---|
| ate | | |
| 1-Fluoro-N-methylcyclopropane-1-carboxamide | 3 | FC1(CC1)C(=O)N(C)[H] |
| 1-Cyclopropyl-3-methylurea | 3 | C1C(C1)NC(=O)N(C)[H] |
| Ethylcyclobutane | 1 | C1C(CC1)CC |
| 3-ethyl-3-methyl-oxetane | 2 | C1C(CO1)(CC)C |
| 3-(Hydroxymethyl)oxetan-3-ol | 2 | C1C(CO1)(CO)O |
| Thietane | 2 | C1CCS1 |
| 2,3-Diethylthietane | 2 | C1(C(CS1)CC)CC |
| Ethylcyclopentane | 1 | C1(CCCC1)CC |
| Tetrahydrothiophene | 2 | C1CCCS1 |
| Pyrrole | 5 | C1=CC=C[N]1 |
| N-Methyl-1H-imidazole-2-carboxamide | 5 | C1=C[N]C(=N1)C(=O)N(C)[H] |
| N,1-Dimethylimidazole-2-carboxamide | 5 | C1=C[N](C(=N1)C(=O)N(C)[H])C |
| Oxazole | 5 | C1=CN=CO1 |
| N-Methyl-1,3-oxazole-2-carboxamide | 5 | C1(=NC=CO1)C(=O)NC |
| Thiazole | 5 | C1=NC=CS1 |
| N-Methyl-1,3-thiazole-2-carboxamide | 5 | C1(=NC=CS1)C(=O)NC |
| Thiophene | 5 | C1=CC=CS1 |
| N-Methylthiophene-2-carboxamide | 5 | C1(=CC=CS1)C(=O)NC |
| 2-Acetylthiophene | 5 | C1(=CC=CS1)C(=O)C |
| N-methylfuran-2-carboxamide | 5 | C1(=CC=CO1)C(=O)N(C)[H] |
| N-Methyl-N-thiophen-2-ylacetamide | 5 | C1(=CC=CS1)N(C(C)=O)C |
| N-(Furan-2-yl)-N-methylacetamide | 5 | C1(=CC=CO1)N(C(C)=O)C |
| N-(Furan-2-yl)acetamide | 5 | C1(=CC=CO1)NC(C)=O |
| Pyrazole | 5 | C1=CC=N[N]1 |
| 3-Methoxy-1h-pyrazole | 5 | C1=CC(=N[N]1)OC |
| 5-Methoxy-1-methyl-1H-pyrazole | 5 | C1(=CC=N[N]1C)OC |
| N,N,1-Trimethyl-1H-imidazole-2-carboxamide | 5 | C1=CN=C([N]1C)C(=O)N(C)C |
| N,1-dimethylpyrrole-2-carboxamide | 5 | C1=CC=C([N]1C)C(=O)NC |

| | | |
|---|---|---|
| 1H-Pyrrol-3-ol | 5 | C1=CC(=C[N]1)O |
| 3-Methoxy-1h-pyrrole | 5 | C1=CC(=C[N]1)OC |
| 1H-Pyrrol-3-amine | 5 | C1=CC(=C[N]1)N |
| 3-Fluoro-1H-pyrrole | 5 | C1=CC(=C[N]1)F |
| 3-Chloro-1H-pyrrole | 5 | C1=CC(=C[N]1)Cl |
| Anisole | 5 | C1(=CC=CC=C1)OC |
| 2-Methoxypyridine | 5 | C1(=NC=CC=C1)OC |
| (Trifluoromethoxy)benzene | 4 | C1(=CC=CC=C1)OC(F)(F)F |
| (Trichloromethoxy)benzene | 4 | C1(=CC=CC=C1)OC(Cl)(Cl)Cl |
| 1-ethyl-2(1H)-Pyridinone | 5 | C1(C=CC=CN1CC)=O |
| 1-Cyclopropyl-1,2-dihydropyridine-2-one | 5 | C1(C=CC=CN1C2CC2)=O |
| 6-Ethyl-1H-pyridin-2-one | 5 | C1(C=CC=C(N1)CC)=O |
| 3-Ethylpyridin-2-ol | 5 | C1(C(=CC=CN1[H])CC)=O |
| 2-ethyl-1,3-oxazole | 5 | C1(=NC=CO1)CC |
| 2-Ethylthiazole | 5 | C1(=NC=CS1)CC |
| N,N-Dimethylaniline | 4 | C1(=CC=CC=C1)N(C)C |
| Cyclopropylbenzene | 4 | C1(=CC=CC=C1)C2CC2 |
| 2-Cyclopropylpyridine | 5 | C1(=NC=CC=C1)C2CC2 |
| 2-Cyclopropylpyrimidine | 5 | C1(=NC=CC=N1)C2CC2 |
| Cyclopentylbenzene | 4 | C1=CC(=CC=C1)C2CCCC2 |
| Benzylmethylether | 4 | C1=CC(=CC=C1)COC |
| N-Methylacetanilide | 4 | C1=CC(=CC=C1)N(C(C)=O)C |
| Acetanilide | 4 | C1=CC(=CC=C1)N(C(C)=O)[H] |
| N-Methylbenzenesulfonamide | 4 | C1=CC(=CC=C1)[S](=O)(=O)NC |
| Ethylphenylsulfone | 4 | C1=CC(=CC=C1)[S](=O)(=O)CC |
| N,N-dimethyl-1H-imidazole-2-sulfonamide | 5 | C1=CN=C([N]1)[S](=O)(=O)N(C)C |
| N,N-dimethyl-2-thiophenesulfonamide | 5 | C1=CC=C([S](=O)(=O)N(C)C)S1 |
| Phenyltrifluoromethylsulfone | 4 | C1(=CC=CC=C1)[S](C(F)(F)F)(=O)=O |
| trichloromethylphenylsulfone | 4 | C1(=CC=CC=C1)[S](C(Cl)(Cl)Cl)(=O)=O |
| Thiopyran | 5 | C1=CC=CCS1 |
| 2H-Pyran | 5 | C1=CC=CCO1 |
| 1,4-Benzoquinone | 5 | O=C1C=CC(=O)C=C1 |
| S-Ethylethanethioate | 3 | CC(=O)SCC |
| Thioaceticacid | 3 | CC(=O)S |
| Thioacetate | 3 | CC(=O)[S-] |
| 1H-Tetrazole | 5 | C1=N[N]N=N1 |
| 1,2,3-Triaza-4- | 5 | C1=N[N-]N=N1 |

| | | |
|---|---|---|
| azanidacyclopenta-2,5-diene | | |
| Ethoxyacetate | 3 | CC(=O)OOCC |
| Peraceticacid | 3 | CC(=O)OOH |
| Ethaneperoxoate | 3 | CC(=O)O[O-] |
| 1-Ethyl-2-pyrrolidinone | 3 | C1CCC(=O)N1CC |
| Pyrrolidinium | 2 | C1CCC[N+]1 |
| 1-Ethylpyrrolidinium | 2 | C1CCC[N+]1CC |
| Carbonate | 3 | C(=O)([O-])[O-] |
| Bicarbonate | 3 | C(=O)([OH])[O-] |
| 3-Hydroxytetrahydro-2h-pyran-2-one | 3 | C1(C(CCCO1)O)=O |
| 2-Piperidone | 3 | C1(CCCCN1)=O |
| alpha-Hydroxy-gamma-butyrolactone | 3 | C1(C(CCO1)O)=O |

**Table S2. Training set for out-of-plane bending**

The training set for out-of-plane bending has been shown with their name, the number of conformations chosen for each molecule, and the related group number as mentioned. (The equilibrium structure has been counted)

| Index | Name of molecule | Number of configurations |
|---|---|---|
| 1 | N-Methylacetamide | 13 |
| 2 | Acetic acid | 7 |
| 3 | Acetamide | 7 |
| 4 | Uracil | 25 |
| 5 | Pyridine | 19 |
| 6 | Benzene | 7 |
| 7 | Fluorobenzene | 25 |
| 8 | Chlorobenzene | 25 |
| 9 | Bromobenzene | 25 |
| 10 | Imidazole | 25 |
| 11 | Phenol | 25 |
| 12 | Indole | 55 |
| 13 | 1H-imidazol-3-ium | 19 |
| 14 | Guanidine | 19 |
| 15 | Acetate | 7 |
| 16 | Formimidamide | 13 |
| 17 | Formamide | 13 |
| 18 | Formicacid | 7 |
| 19 | 7-Methylindole | 55 |
| 20 | Methylguanidine | 25 |
| 21 | 4-Methylimidazole | 25 |
| 22 | 2-Aminopyridine | 31 |
| 23 | 2-Pyridone | 37 |
| 24 | Pyrazine | 7 |
| 25 | Thymine | 37 |
| 26 | Adenine | 43 |
| 27 | Formaldehyde | 7 |
| 28 | Methyl benzene | 25 |
| 29 | Acetone | 7 |
| 30 | Naphthalene | 19 |
| 31 | Guanine | 49 |
| 32 | Acetaldehyde | 7 |
| 33 | Cytosine | 37 |
| 34 | Phenolate | 25 |
| 35 | Aniline | 7 |
| 36 | Acetic Anhydride | 7 |
| 37 | Carbamic acid | 7 |
| 38 | 3-Methylindole | 55 |

| | | |
|---|---|---|
| 39 | 4-Methylimidazole | 25 |
| 40 | 1-Propylguanidinium | 19 |
| 41 | Propionamide | 13 |
| 42 | Propionic acid | 7 |
| 43 | Formate | 7 |
| 44 | Propionyl chloride | 7 |
| 45 | Propionyl fluoride | 7 |
| 46 | Propionyl bromide | 7 |
| 47 | Diethyl carbonate | 7 |
| 48 | Ethyl hydrogen carbonate | 7 |
| 49 | Diethoxymethane | 7 |
| 50 | Benzonitrile | 19 |
| 51 | Propionate | 7 |
| 52 | Propionaldehyde | 7 |
| 53 | N-Ethylformamide | 13 |
| 54 | Dimethyl formamide | 13 |
| 55 | 1,3-Benzodioxole | 19 |
| 56 | Ethanamine | 7 |
| 57 | Ethanimine | 7 |
| 58 | 2-Propanimine | 7 |
| 59 | Ethyl Acetoacetate | 13 |
| 60 | Hydrogen oxalate | 13 |
| 61 | Oxalate | 7 |
| 62 | Malonic acid | 7 |
| 63 | Hydrogen malonate | 13 |
| 64 | Malonate | 7 |
| 65 | 2-Oxobutanoic Acid | 13 |
| 66 | 2-Oxobutanoate | 13 |
| 67 | Dipropionamide | 13 |
| 68 | Nitroethane | 7 |
| 69 | Nitrobenzene | 31 |
| 70 | Ethyl N-ethyl-N-methylcarbamate | 7 |
| 71 | Methylcarbamic acid | 13 |
| 72 | Carbamate | 13 |
| 73 | Benzenediazonium | 7 |
| 74 | Azobenzene | 7 |
| 75 | Benzene sulfonic acid | 19 |
| 76 | Benzene sulfonate | 25 |
| 77 | Ethyl acetate | 7 |
| 78 | Phenyl thiocyanate | 25 |
| 79 | Butadiene | 7 |
| 80 | Acrolein | 13 |
| 81 | Acetone oxime | 7 |

| | | |
|---|---|---|
| 82 | Benzyl alcohol | 7 |
| 83 | Ethylbenzene | 7 |
| 84 | Benzylamine | 7 |
| 85 | Benzaldehyde | 31 |
| 86 | Benzoic acid | 25 |
| 87 | Benzoate | 31 |
| 88 | AcrylicAcid | 13 |
| 89 | Acrylate | 13 |
| 90 | L-Alanine | 7 |
| 91 | Phenylazanium | 25 |
| 92 | Glyoxal | 7 |
| 93 | 2-Methyl-1-butene | 7 |
| 94 | (E)-but-2-ene | 7 |
| 95 | Glycolic acid | 7 |
| 96 | Glycolate | 7 |
| 97 | Malonaldehyde | 7 |
| 98 | Urea | 13 |
| 99 | Citric acid | 19 |
| 100 | Alpha-Ketoglutarate | 19 |
| 101 | Succinate | 7 |
| 102 | Malic acid | 13 |
| 103 | Cysteine | 7 |
| 104 | Ethyl propionate | 7 |
| 105 | Arginine | 25 |
| 106 | Glutamine | 19 |
| 107 | Isobutyric anhydride | 7 |
| 108 | Diisopropyl carbonate | 7 |
| 109 | N-Methyl-1-butanimine | 7 |
| 110 | 2-Methyl-1-nitropropane | 7 |
| 111 | Fumaric acid | 12 |
| 112 | Hexa-1,3,5-triene | 13 |
| 113 | 3-Pentanone | 7 |
| 114 | Butyraldehyde | 7 |
| 115 | Hexamethylbenzene | 7 |
| 116 | Salicylic acid | 43 |
| 117 | Cumene | 7 |
| 118 | p-Cresol | 25 |
| 119 | 3-Methylanilinium | 37 |
| 120 | 3-Methylbenzaldehyde | 43 |
| 121 | p-Toluidine | 30 |
| 122 | 3-Nitrotoluene | 13 |
| 123 | 4-Fluorophenol | 13 |
| 124 | 1,3,5-Trichlorobenzene | 13 |

| | | |
|---|---|---|
| 125 | 1,3-Dibromobenzene | 25 |
| 126 | 4-Aminopyridine | 19 |
| 127 | 1-Methylcytosine | 37 |
| 128 | 1-Methylthymine | 37 |
| 129 | 1-Methyluracil | 37 |
| 130 | 9-Methyladenine | 31 |
| 131 | 9-Methylguanine | 43 |
| 132 | Cytidine | 37 |
| 133 | Deoxycytidine | 37 |
| 134 | (2R)-2-aminopropanoic acid | 7 |
| 135 | 2-methylpropanoic acid | 7 |
| 136 | methyl propanoate | 7 |
| 137 | (2R)-2-hydroxypropanoic acid | 7 |
| 138 | (2S)-2-hydroxypropanoic acid | 7 |
| 139 | 3-methylbutan-2-one | 7 |
| 140 | (E)-pent-2-enal | 19 |
| 141 | (2S)-3-methylbut-3-en-2-ol | 7 |
| 142 | (E,2R)-hex-3-en-2-ol | 13 |
| 143 | (E,2S)-hex-3-en-2-ol | 13 |
| 144 | (Z)-1-ethoxybut-1-ene | 13 |
| 145 | N-ethylpropanamide | 13 |
| 146 | S-ethyl propanethioate | 7 |
| 147 | propan-2-yl acetate | 7 |
| 148 | (E)-pent-2-enoic acid | 19 |
| 149 | 2-methylpentan-3-one | 7 |
| 150 | (Z)-pent-2-enoic acid | 19 |
| 151 | propan-2-yl propanoate | 7 |
| 152 | N-propan-2-ylpropanamide | 13 |
| 153 | (2S)-N-ethyl-2-hydroxypropanamide | 13 |
| 154 | (2R)-N-ethyl-2-hydroxypropanamide | 13 |
| 155 | ethyl 2-methylpropanoate | 7 |
| 156 | (E)-4-oxohex-2-enal | 25 |
| 157 | 3-oxopentanoic acid | 13 |
| 158 | (E)-hex-3-enoic acid | 19 |
| 159 | (2S)-2-hydroxy-N-propan-2-ylpropanamide | 13 |
| 160 | (2R)-2-hydroxy-N-propan-2-ylpropanamide | 13 |
| 161 | 4-ethylphenol | 13 |
| 162 | (2S,3E,5E)-octa-3,5-dien-2-ol | 25 |
| 163 | (2R,3E,5E)-octa-3,5-dien-2-ol | 25 |
| 164 | (E)-4-oxohex-2-enoic acid | 25 |
| 165 | 2-(propanoylamino)acetic acid | 19 |

| | | |
|---|---|---|
| 166 | 5-ethylbenzene-1,3-diol | 25 |
| 167 | (2R)-2-(propanoylamino)propanoic acid | 19 |
| 168 | (2S)-2-(propanoylamino)propanoic acid | 19 |
| 169 | S-ethyl 3-oxopentanethioate | 13 |
| 170 | 2,5-diethyl-3,4-dimethylfuran | 13 |
| 171 | [(E)-3-methylpent-2-enyl] propanoate | 19 |
| 172 | D-Alaninate | 7 |
| 173 | D-Alaninium | 7 |
| 174 | D-Alanine zwitterion | 7 |
| 175 | 2-methylpropanoate | 7 |
| 176 | (2R)-2-hydroxypropanoate | 7 |
| 177 | (2S)-2-hydroxypropanoate | 7 |
| 178 | (E)-pent-2-enoate | 7 |
| 179 | (Z)-pent-2-enoate | 19 |
| 180 | 3-oxopentanoate | 13 |
| 181 | (E)-hex-3-enoate | 19 |
| 182 | 2-(propanoylamino)acetate | 19 |
| 183 | (2R)-2-(propanoylamino)propanoate | 19 |
| 184 | Benzotrifluoride | 7 |
| 185 | Benzotrichloride | 7 |
| 186 | 2,2,2-trifluoroethylbenzene | 7 |
| 187 | 2,2,2-trichloroethylbenzene | 7 |
| 188 | Tribromoethylbenzene | 7 |
| 189 | Iodobenzene | 7 |
| 190 | 1-Phenylpiperazinium | 6 |
| 191 | Tribromomethylbenzene | 7 |
| 192 | Biphenyl | 13 |
| 193 | Diphenylmethane | 7 |
| 194 | 2-Phenoxypyridine | 13 |
| 195 | 2-Anilinopyridine | 13 |
| 196 | Phenyl-2-pyridyl-methylamine | 13 |
| 197 | Diphenylamine | 7 |
| 198 | Caprolactam | 7 |
| 199 | N-Methylcaprolactam | 7 |
| 200 | 1,3,4,5-Tetrahydro-2H-1-benzazepin-2-one | 19 |
| 201 | 2,3,4,5-Tetrahydro-benzo[c]azepin-1-one | 19 |
| 202 | 1-Methyl-1,3,4,5-tetrahydro-2H-1-benzazepin-2-one | 19 |
| 203 | 5-Methyl-1H-benzo[e][1,4]diazepin-2(3H)-one | 25 |
| 204 | Benzamide | 13 |
| 205 | 1-Benzoylpiperidine | 13 |

| | | |
|---|---|---|
| 206 | (2R)-1-Benzoyl-2alpha-methylpiperidine | 13 |
| 207 | 2-(1Methyl-Imidazol-2-yl)pyrimidine | 37 |
| 208 | 2-(5-Methylpyrazol-1-yl)pyrimidine | 43 |
| 209 | 3-(2-Pyridyl)furan | 55 |
| 210 | 2-(2-Pyridyl)thiazole | 25 |
| 211 | Phenetole | 7 |
| 212 | 1-Methyl-2-phenyl-1H-pyrrole | 31 |
| 213 | 2-Phenyl-1H-pyrrole | 37 |
| 214 | 1-Phenylpyrazole | 31 |
| 215 | Methyl2-(methylamino)benzoate | 13 |
| 216 | 2-Aminobenzophenone | 37 |
| 217 | (2-(methylamino)phenyl)phenyl-Methanone | 31 |
| 218 | 2-Aminobiphenyl | 7 |
| 219 | 2-(Pyridin-2-yl)aniline | 43 |
| 220 | N-Methyl-2-(2-pyridyl)aniline | 37 |
| 221 | N,N-Dimethyl-2-pyridin-2-ylaniline | 7 |
| 222 | 1-Cyclopropyl-3-methylurea | 7 |
| 223 | Pyrrole | 13 |
| 224 | N-Methyl-1H-imidazole-2-carboxamide | 25 |
| 225 | N,1-Dimethylimidazole-2-carboxamide | 25 |
| 226 | Oxazole | 19 |
| 227 | N-Methyl-1,3-oxazole-2-carboxamide | 25 |
| 228 | Thiazole | 19 |
| 229 | N-Methyl-1,3-thiazole-2-carboxamide | 25 |
| 230 | Thiophene | 13 |
| 231 | N-Methylthiophene-2-carboxamide | 31 |
| 232 | 2-Acetylthiophene | 31 |
| 233 | N-methylfuran-2-carboxamide | 31 |
| 234 | N-Methyl-N-thiophen-2-ylacetamide | 25 |
| 235 | N-(Furan-2-yl)-N-methylacetamide | 31 |
| 236 | N-(Furan-2-yl)acetamide | 31 |
| 237 | Pyrazole | 19 |
| 238 | 3-Methoxy-1h-pyrazole | 19 |
| 239 | 5-Methoxy-1-methyl-1H-pyrazole | 19 |
| 240 | N,N,1-Trimethyl-1H-imidazole-2-carboxamide | 25 |
| 241 | N,1-dimethylpyrrole-2-carboxamide | 19 |
| 242 | 1H-Pyrrol-3-ol | 19 |
| 243 | 3-Methoxy-1h-pyrrole | 19 |
| 244 | 1H-Pyrrol-3-amine | 19 |
| 245 | 3-Fluoro-1H-pyrrole | 7 |
| 246 | 3-Chloro-1H-pyrrole | 7 |

| | | |
|---|---|---|
| 247 | Anisole | 7 |
| 248 | 2-Methoxypyridine | 7 |
| 249 | (Trifluoromethoxy)benzene | 7 |
| 250 | (Trichloromethoxy)benzene | 7 |
| 251 | 1-ethyl-2(1H)-Pyridinone | 31 |
| 252 | 1-Cyclopropyl-1,2-dihydropyridine-2-one | 31 |
| 253 | 6-Ethyl-1H-pyridin-2-one | 7 |
| 254 | 3-Ethylpyridin-2-ol | 13 |
| 255 | 2-ethyl-1,3-oxazole | 19 |
| 256 | 2-Ethylthiazole | 19 |
| 257 | N,N-Dimethylaniline | 7 |
| 258 | Cyclopropylbenzene | 7 |
| 259 | 2-Cyclopropylpyridine | 7 |
| 260 | 2-Cyclopropylpyrimidine | 19 |
| 261 | Cyclopentylbenzene | 7 |
| 262 | N-Methylacetanilide | 7 |
| 263 | Acetanilide | 7 |
| 264 | N-Methylbenzenesulfonamide | 7 |
| 265 | Ethylphenylsulfone | 7 |
| 266 | N,N-dimethyl-1H-imidazole-2-sulfonamide | 13 |
| 267 | N,N-dimethyl-2-thiophenesulfonamide | 25 |
| 268 | Phenyltrifluoromethylsulfone | 7 |
| 269 | trichloromethylphenylsulfone | 7 |
| 270 | Thiopyran | 25 |
| 271 | 2H-Pyran | 25 |
| 272 | 1,4-Benzoquinone | 13 |
| 273 | S-Ethylethanethioate | 7 |
| 274 | Thioaceticacid | 7 |
| 275 | Thioacetate | 7 |
| 276 | 1H-Tetrazole | 7 |
| 277 | 1,2,3-Triaza-4-azanidacyclopenta-2,5-diene | 7 |
| 278 | Ethoxyacetate | 7 |
| 279 | Peraceticacid | 7 |
| 280 | Ethaneperoxoate | 7 |
| 281 | 1-Ethyl-2-pyrrolidinone | 7 |
| 282 | Carbonate | 6 |
| 283 | Bicarbonate | 6 |
| 284 | 3-Hydroxytetrahydro-2h-pyran-2-one | 6 |
| 285 | 2-Piperidone | 6 |
| 286 | alpha-Hydroxy-gamma-butyrolactone | 6 |

**Table S3. The molecules in the validation set**

The validation set has been shown with their name, their classified group, and the SMILE strings (120 molecules). We have 7 groups: (1).alkanes and aliphatic rings; (2).molecules comprised of only single bonds; (3).molecules with double bonds or triple bonds; (4).benzene derivatives; (5).heterocyclics; (6).bio-fragments; (7).combinations of multiple aromatic cycles.

| Name | Group | SMILE |
|---|---|---|
| 1,2-Ethanediol | 2 | (CO)O |
| 1-Amino-2-butanol | 2 | C(C(CN)O)C |
| 1,2-Butanediamine | 2 | C(C(CN)N)C |
| 1-Chloro-3-fluoropropane | 2 | C(CCCl)F |
| 1-Bromo-3-chloropropane | 2 | C(CCCl)Br |
| 1-Bromo-3-fluoropropane | 2 | C(CCF)Br |
| 1,3-Diiodopropane | 2 | C(CCI)I |
| Tetramethylammonium | 2 | C[N+](C)(C)C |
| 2-Hydroxyethyldimethylammonium | 2 | C[N+](CCO)C |
| 3,5-Dihydroxycyclohexanamine | 2 | C1C(CC(CC1O)O)N |
| 1,3,5-Trifluorocyclohexane | 2 | C1C(CC(CC1F)F)F |
| Cyclopentanemethanol | 2 | C1(CCCC1)CO |
| Cyclopentanemethylamine | 2 | C1(CCCC1)CN |
| 1,3-Diethylcyclobutane | 2 | C1C(CC1CC)CC |
| 2,4-Diethyloxetane | 2 | C1(CC(O1)CC)CC |
| 1,2,3-Trimethylcyclopropane | 2 | C1(C(C1C)C)C |
| Ethylisobutyldisulfide | 2 | C(SSCC)C(C)C |
| 1,2-Dimethoxyethane | 2 | COCCOC |
| 2-Methylpropane-1-thiolate | 2 | CC(C)CS- |
| Hexa-2,4-diene | 3 | C(C=CC)=CC |
| 3-Methyl-2-butenal | 3 | C(=CC=O)(C)C |
| 3-Ethylhexane-2,4-dione | 3 | C(C(C(=O)C)CC)(=O)CC |
| 5-Oxohexanal | 3 | C(CCCC(=O)[H])(=O)C |
| Succinamide | 3 | C(CC(=O)N)C(=O)N |
| Diethyloxalate | 3 | C(C(=O)OCC)(=O)OCC |
| Succinylfluoride | 3 | C(CCC(=O)F)(F)=O |
| Succinylchloride | 3 | C(CCC(=O)Cl)(Cl)=O |
| Butanedioyldibromide | 3 | C(CCC(=O)Br)(Br)=O |
| Isovaleronitrile | 3 | CC(CC#N)C |
| 4-Methyl-1-pentyne | 3 | CC(CC#C)C |
| N-Ethylbutane-2-imine | 3 | C(=NCC)(CC)C |
| Butan-2-imine | 3 | C(=N)(CC)C |

| | | |
|---|---|---|
| Monomethylsuccinate | 3 | C(CCC(=O)OC)(=O)O |
| Monomethylsuccinate-ion | 3 | C(CCC(=O)OC)(=O)[O-] |
| 2,3-Dimethyl-2-butenoicacid | 3 | C(=C(C)C)(C(=O)O)C |
| Thioisobutyricacid | 3 | C(C(S[H])=O)(C)C |
| 2-Methylpropanethioate | 3 | C(C([S-])=O)(C)C |
| 3-Methylbutanoylacetate | 3 | C(C(=O)OC(C)=O)C(C)C |
| 1-Ethyl-1-methylurea | 3 | C(=O)(N(C)CC)N |
| Cyclohexylammonium | 2 | C1(CCCCC1)[N+] |
| 1,4-Dimethylpiperazine | 3 | C1CN(CCN1C)C |
| Isobutylpropylsulfoxide | 3 | C([S](CC(C)C)=O)CC |
| 2-Methyl-1-propylsulfonylpropane | 3 | C([S](CC(C)C)(=O)=O)CC |
| Ethyl-2-methylpropane-1-sulfonate | 3 | [S](CC(C)C)(=O)(=O)OCC |
| Isopropylmethanesulfonate | 3 | [S](CC(C)C)(=O)(=O)[O-] |
| 2-Methyl-1-nitropropane | 3 | CC(C[N](=O)=O)C |
| Isopropylphosphate | 6 | CC(C)O[P](=O)([O-])[O-] |
| Urethane | 3 | C(OC(=O)N)C |
| Propan-2-yl-hydrogencarbonate | 3 | CC(C)OC(=O)O |
| D-gluconicacid | 6 | O([C@@H]([C@H](O[H])[C@@H](O[H])C(O[H])=O)[C@H](O[H])CO[H])[H] |
| Gluconolactone | 6 | O1[C@@H]([C@@H](O[H])[C@H](O[H])[C@@H](O[H])C1=O)CO[H] |
| N-Methylsalicylamide | 4 | C1(=CC=CC=C1C(=O)NC)O |
| 2-Chloroacetanilide | 4 | C1(=CC=CC=C1NC(C)=O)Cl |
| 2-Fluoroanisole | 4 | C1(=CC=CC=C1OC)F |
| 2-Methylanisole | 4 | C1(=CC=CC=C1OC)C |
| 3-Fluoro-2-methoxyphenol | 4 | C1(=C(C=CC=C1O)F)OC |
| N,2,6-Trimethylaniline | 4 | C1(=C(C=CC=C1C)C)NC |
| 3-methyl-2-(methylamino)phenol | 4 | C1(=C(C=CC=C1C)O)NC |
| 2-Hydroxyacetophenone | 4 | C1(=CC=CC=C1C(C)=O)O |
| 2-Methoxybenzamide | 4 | C1=CC=CC(=C1OC)C(=O)N |
| Phthalamide | 4 | C1=CC=CC(=C1C(=O)N)C(=O)N |
| 2-Aminobenzamide | 4 | C1(=CC=CC=C1C(=O)N)N |
| 2,4,6-Trinitrobenzoicacid | 4 | C1(=CC(=C(C(=C1)[N+](=O)[O-])C(=O)O)[N+](=O)[O-])[N+](=O)[O-] |
| 2,4,6-Trinitrobenzoate | 4 | C1(=CC(=C(C(=C1)[N+](=O)[O-])C(=O)[O-])[N+](=O)[O-])[N+](=O)[O-] |
| 3,5-Dimethylbenzenaminium | 4 | C1(=CC(=CC(=C1)C)C)[N+] |

| | | |
|---|---|---|
| 3,5-Dimethylbenzenesulfonicacid | 4 | C1(=CC(=CC(=C1)C)C)[S](O)(=O)=O |
| 3,5-Dimethylbenzenesulfonate | 4 | C1(=CC(=CC(=C1)C)C)[S]([O-])(=O)=O |
| 2-Methylbenzotrifluoride | 4 | C1=CC=CC(=C1C(F)(F)F)C |
| 2-Methylbenzotrichloride | 4 | C1=CC=CC(=C1C(Cl)(Cl)Cl)C |
| 3,5-Dichlorotoluene | 4 | C1=C(C=C(C=C1Cl)C)Cl |
| 3,5-Difluorotoluene | 4 | C1=C(C=C(C=C1F)C)F |
| 3,5-Dibromotoluene | 4 | C1=C(C=C(C=C1Br)C)Br |
| 5,6-Dimethyl-1,3-benzodioxole | 4 | C12=CC(=C(C=C1OCO2)C)C |
| o-Tolylaceticacid | 4 | C1(=CC=CC=C1CC(=O)O)C |
| 2-methylphenylacetate | 4 | C1(=CC=CC=C1CC(=O)[O-])C |
| 2-Methylbenzylalcohol | 4 | C1(=C(C=CC=C1)C)CO |
| 2-Methylbenzaldehyde | 4 | C1(=C(C=CC=C1)C)C=O |
| 2-Chloroacetophenone | 4 | C1(=C(C=CC=C1)Cl)C(=O)C |
| 2-Methoxy-3-methylpyridine | 5 | C1(=CC=CN=C1OC)C |
| 1-propan-2-ylpyridin-2-one | 5 | C1=CC=CN(C1=O)C(C)C |
| 3,4,5-trimethyl-1H-pyridin-2-one | 5 | C1(=C(C(=CNC1=O)C)C)C |
| 9-Methyl-2,3,4,5-tetrahydro-2-benzazepin-1-one | 4 | C1(C2=C(CCCN1[H])C=CC=C2C)=O |
| (2-methylphenyl)-piperidin-1-ylmethanone | 4 | C1CCCCN1C(C2=CC=CC=C2C)=O |
| 3-Methyl-2-(5-methyl-1H-imidazol-2-yl)pyridine | 7 | C1=NC(=C(C=C1)C)C2=NC=C([N]2)C |
| 3-methyl-1-(2-pyridyl)pyrazole | 7 | C1=NC(=CC=C1)[N]2C=CC(=N2)C |
| 2-Isopropyloxazole | 5 | C1=COC(=N1)C(C)C |
| 2-Isopropylthiazole | 5 | C1=CSC(=N1)C(C)C |
| 2-(4-Methylpyridin-2-yl)-1,3-thiazole | 7 | C1=NC(=CC(=C1)C)C2=NC=CS2 |
| 2-(2-Furyl)-4-methylpyridine | 7 | C1=NC(=CC(=C1)C)C2=CC=CO2 |
| 4-Methyl-2-pyridin-2-ylaniline | 7 | C1=NC(=CC=C1)C2=C(C=CC(=C2)C)N |
| 1-Methyl-4-(phenoxymethyl)benzene | 7 | C1=CC(=CC=C1COC2=CC=CC=C2)C |
| (3-Methylcyclopentyl)benzene | 4 | C1=CC=CC=C1C2CCC(C2)C |

| | | |
|---|---|---|
| 2-(3-Methylcyclopentyl)pyridine | 5 | C1=NC(=CC=C1)C2CCC(C2)C |
| 1-cyclopropyl-4-methylbenzene | 4 | C1=CC(=CC=C1C2CC2)C |
| 2-Cyclopropyl-5-methylpyridine | 5 | C1=NC(=CC=C1C)C2CC2 |
| 2-Cyclopropyl-5-methylpyrimidine | 5 | C1=NC(=NC=C1C)C2CC2 |
| 4-Benzyltoluene | 4 | C1=CC=C(C=C1)CC2=CC=C(C=C2)C |
| 5-Methyl-2-phenoxypyridine | 7 | C1=CC=C(C=C1)OC2=NC=C(C=C2)C |
| 5-Methyl-N-phenylpyridin-2-amine | 7 | C1=CC=C(C=C1)NC2=NC=C(C=C2)C |
| 4-methyl-N-phenylaniline | 7 | C1=CC=C(C=C1)N(C2=CC=C(C=C2)C) |
| N,5-Dimethyl-N-phenylpyridin-2-amine | 7 | C1=CC=C(C=C1)N(C2=NC=C(C=C2)C)C |
| 4-Benzyltoluene | 7 | C1=CC(=CC=C1C2=CC=CC=C2)C |
| N,4-Dimethylfuran-2-carboxamide | 5 | C1=C(OC=C1C)C(=O)NC |
| N,4-Dimethylthiophene-2-carboxamide | 5 | C1=C(C(=O)N(C)[H])SC=C1C |
| 4-Methylthiophene-2-carbaldehyde | 5 | C1=C(C=O)SC=C1C |
| 3-Methyl-5-acetamidofuran | 5 | C1=C(OC=C1C)NC(C)=O |
| N-(4-Methylthiophen-2-yl)acetamide | 5 | C1=C(N(C(C)=O)[H])SC=C1C |
| N,N-Dimethyl-1,3-oxazole-2-carboxamide | 5 | C1(=NC=CO1)C(=O)N(C)C |
| 1-(2-Methylphenyl)piperazin-4-ium | 5 | C1=CC(=C(C=C1)N2CC[N+]CC2)C |
| 2-Amino-4-methylbenzophenone | 7 | C1(=CC=C(C=C1)C)C(C2=CC=CC=C2N)=O |
| 1-Methylpyrazole | 5 | C1=C[N](N=C1)C |
| 4-Methyl-2H-pyran | 5 | C1=CC(=CCO1)C |
| 2-Methylperoxypropionicacid | 3 | C(C(=O)OO)(C)C |
| Peroxypropionate | 3 | C(C(=O)O[O-])C |
| 3-Methylpyrrolidin-1-ium | 2 | C1C(CC[N+]1)C |
| 1,3,4-Trimethylpyrrolidin-1-ium | 2 | C1C(C)C(C[N+]1C)C |
| 4-Triflyltoluene | 4 | C1(=CC=C(C=C1)C)[S](C(F)(F)F)(=O)=O |
| 4-Methyl-2H-thiopyran | 5 | C1C=C(C=CS1)C |

| | | |
|---|---|---|
| 2,6-Dimethylbenzoquinone | 5 | C1(C=C(C(C(=C1)C)=O)C)=O |
| 1-Iodopropane | 2 | CCCI |
| 1,1,1-Triiodopropane | 2 | CCC(I)(I)I |
| 1-Iodo-3,5-dimethylbenzene | 4 | C1(=CC(=CC(=C1)C)C)I |
| 1-(iodomethyl)-3,5-dimethylbenzene | 4 | C1(=CC(=CC(=C1)C)C)CI |

**Table S4. Validation set for out-of-plane bending**

The validation set for out-of-plane bending has been shown with their name, the number of conformations chosen for each molecule, and the related group number as mentioned. (The equilibrium structure has been counted)

| Index | Name of molecule | Number of configurations |
|---|---|---|
| 1 | Hexa-2,4-diene | 11 |
| 2 | 3-Methyl-2-butenal | 16 |
| 3 | 3-Ethylhexane-2,4-dione | 6 |
| 4 | 5-Oxohexanal | 11 |
| 5 | Succinamide | 6 |
| 6 | Diethyloxalate | 6 |
| 7 | Succinylfluoride | 6 |
| 8 | Succinylchloride | 6 |
| 9 | Butanedioyldibromide | 6 |
| 10 | N-Ethylbutane-2-imine | 6 |
| 11 | Butan-2-imine | 6 |
| 12 | Monomethylsuccinate | 6 |
| 13 | Monomethylsuccinate-ion | 6 |
| 14 | 2,3-Dimethyl-2-butenoicacid | 11 |
| 15 | Thioisobutyricacid | 6 |
| 16 | 2-Methylpropanethioate | 6 |
| 17 | 3-Methylbutanoylacetate | 11 |
| 18 | 1-Ethyl-1-methylurea | 6 |
| 19 | 2-Methyl-1-nitropropane | 6 |
| 20 | Urethane | 6 |
| 21 | Propan-2-yl-hydrogencarbonate | 6 |
| 22 | Gluconolactone | 6 |
| 23 | N-Methylsalicylamide | 6 |
| 24 | 2-Chloroacetanilide | 6 |
| 25 | 2-Fluoroanisole | 11 |
| 26 | 2-Methylanisole | 11 |
| 27 | 3-Fluoro-2-methoxyphenol | 16 |
| 28 | N,2,6-Trimethylaniline | 21 |
| 29 | 3-methyl-2-(methylamino)phenol | 16 |
| 30 | 2-Hydroxyacetophenone | 16 |
| 31 | 2-Methoxybenzamide | 21 |
| 32 | Phthalamide | 11 |
| 33 | 2-Aminobenzamide | 21 |
| 34 | 2,4,6-Trinitrobenzoicacid | 11 |
| 35 | 2,4,6-Trinitrobenzoate | 6 |
| 36 | 3,5-Dimethylbenzenaminium | 6 |
| 37 | 3,5-Dimethylbenzenesulfonicacid | 11 |
| 38 | 3,5-Dimethylbenzenesulfonate | 11 |

| | | |
|---|---|---|
| 39 | 2-Methylbenzotrifluoride | 16 |
| 40 | 2-Methylbenzotrichloride | 16 |
| 41 | 3,5-Dichlorotoluene | 11 |
| 42 | 3,5-Difluorotoluene | 11 |
| 43 | 3,5-Dibromotoluene | 11 |
| 44 | 5,6-Dimethyl-1,3-benzodioxole | 16 |
| 45 | o-Tolylaceticacid | 16 |
| 46 | 2-methylphenylacetate | 16 |
| 47 | 2-Methylbenzylalcohol | 6 |
| 48 | 2-Methylbenzaldehyde | 11 |
| 49 | 2-Chloroacetophenone | 11 |
| 50 | 2-Methoxy-3-methylpyridine | 11 |
| 51 | 1-propan-2-ylpyridin-2-one | 16 |
| 52 | 3,4,5-trimethyl-1H-pyridin-2-one | 16 |
| 53 | 9-Methyl-2,3,4,5-tetrahydro-2-benzazepin-1-one | 11 |
| 54 | (2-methylphenyl)-piperidin-1-ylmethanone | 6 |
| 55 | 3-Methyl-2-(5-methyl-1H-imidazol-2-yl)pyridine | 16 |
| 56 | 3-methyl-1-(2-pyridyl)pyrazole | 16 |
| 57 | 2-Isopropyloxazole | 11 |
| 58 | 2-Isopropylthiazole | 11 |
| 59 | 2-(4-Methylpyridin-2-yl)-1,3-thiazole | 21 |
| 60 | 2-(2-Furyl)-4-methylpyridine | 11 |
| 61 | 4-Methyl-2-pyridin-2-ylaniline | 6 |
| 62 | (3-Methylcyclopentyl)benzene | 6 |
| 63 | 2-(3-Methylcyclopentyl)pyridine | 6 |
| 64 | 1-cyclopropyl-4-methylbenzene | 6 |
| 65 | 2-Cyclopropyl-5-methylpyridine | 6 |
| 66 | 2-Cyclopropyl-5-methylpyrimidine | 16 |
| 67 | 5-Methyl-2-phenoxypyridine | 6 |
| 68 | 5-Methyl-N-phenylpyridin-2-amine | 6 |
| 69 | 4-methyl-N-phenylaniline | 6 |
| 70 | N,5-Dimethyl-N-phenylpyridin-2-amine | 6 |
| 71 | 4-Benzyltoluene | 6 |
| 72 | N,4-Dimethylfuran-2-carboxamide | 16 |
| 73 | N,4-Dimethylthiophene-2-carboxamide | 16 |
| 74 | 4-Methylthiophene-2-carbaldehyde | 26 |
| 75 | 3-Methyl-5-acetamidofuran | 16 |
| 76 | N-(4-Methylthiophen-2-yl)acetamide | 16 |

| | | |
|---|---|---|
| 77 | N,N-Dimethyl-1,3-oxazole-2-carboxamide | 16 |
| 78 | 1-(2-Methylphenyl)piperazin-4-ium | 6 |
| 79 | 1-Methylpyrazole | 16 |
| 80 | 4-Methyl-2H-pyran | 11 |
| 81 | 2-Methylperoxypropionicacid | 6 |
| 82 | Peroxypropionate | 6 |
| 83 | 4-Triflyltoluene | 6 |
| 84 | 4-Methyl-2H-thiopyran | 11 |
| 85 | 2,6-Dimethylbenzoquinone | 11 |
| 86 | 1-Iodo-3,5-dimethylbenzene | 6 |

**Table S5. The overall atomic types based on SMARTS strings**

| SMARTS strings | General index | Type index | Short name | Column6 |
|---|---|---|---|---|
| [H]* | 1 | 1 | H* | H*, wild matching Hydrogen |
| [C]* | 2 | 2 | C* | C*, wild matching Carbon |
| [c]* | 3 | 3 | c* | c*, wild matching Carbon (aromatic) |
| [N]* | 4 | 4 | N* | N*, wild matching Nitrogen |
| [n]* | 5 | 5 | n* | N*, wild matching Nitrogen (aromatic) |
| [O]* | 6 | 6 | O* | O*, wild matching Oxygen |
| [o]* | 7 | 7 | o* | o*, wild matching oxygen (aromatic) |
| [P]* | 8 | 8 | P* | P*, wild matching Phosphor |
| [S]* | 9 | 9 | S* | S*, wild matching Sulphur |
| [s]* | 10 | 10 | s* | s*, wild matching sulfur (aromatic) |
| [F]* | 11 | 11 | F | F*, wild matching Fluorine |
| [Cl]* | 12 | 12 | Cl | Cl*, wild matching Chlorine |
| [Br]* | 13 | 13 | Br | Br*, wild matching Bromine |
| [I]* | 14 | 14 | I | I*, wild matching Iodine |
| [OH2] | 15 | 15 | OW | OW: O on Water |
| [H][OH2] | 16 | 16 | HW | HW: H on Water |
| [H][OH1] | 17 | 17 | HO | HO, H on Oxygen |
| [H][C] | 18 | 18 | HC | HC, H on carbon |
| [H][CH] | 19 | 19 | HCH | HCH, H on sp3 carbon, -CH- |
| [H][CH2] | 20 | 20 | HCH2 | HCH2, H on sp3 carbon, -CH2- |
| [H][CH3] | 21 | 21 | HCH3 | HCH3, H on sp3 carbon, -CH3 |
| [H][C]=[C,N,O] | 22 | 22 | HC2 | HC2, H on sp2 carbon |
| [H][c] | 23 | 23 | Hc | HC, H on aromatic |

| | | | | |
|---|---|---|---|---|
| | | | | carbon |
| [H][N] | 24 | 24 | HN | HN, H on Nitrogen |
| [H][S] | 25 | 25 | HS | HS, H on Sulfur |
| [H][P] | 26 | 26 | HP | HP, H on Phosphor |
| [H][NX4] | 27 | 27 | H+ | H+, H on NH4+, R-NH3+, R2-NH2+, R3-NH+ |
| [H][N]=[C](N)[N] | 28 | 27 | H+ | H+, H on Guandinium, Imidazolium |
| [H][N]-[C](=N)[N] | 29 | 27 | H+ | H+, H on Guandinium, Imidazolium |
| [H][n] | 30 | 28 | Hn | Hn, H on aromatic N |
| [c] | 31 | 29 | car | car, aromatic carbon |
| [CH]1C=CC=CC1=O | 32 | 29 | car | car, |
| [C]1=CC=CC(=O)[CH]1 | 33 | 29 | car | car, |
| [C]1=C[CH]C(=O)C=C1 | 34 | 29 | car | car, |
| [C]1=CC(=O)[CH]C=C1 | 35 | 29 | car | car, |
| [C]1=CC=C[CH]C1(=O) | 36 | 29 | car | car, |
| [C]1(=O)[CH]C=CC=C1 | 37 | 29 | car | car, |
| [cr5H0] | 38 | 30 | c0r5 | c0r5, aromatic carbon in 5-membered cycle without H |
| [c]([c])([c])[c] | 39 | 31 | car0 | car, aromatic carbon without H |
| [CX4H0] | 40 | 32 | C30 | C30, sp3 carbon with 0 Hydrogen |
| [CX4H1] | 41 | 33 | C31 | C31, sp3 carbon with 1 Hydrogen |
| [CX4H2] | 42 | 34 | C32 | C32, sp3 carbon with 2 Hydrogens |
| [CX4H3] | 43 | 35 | C33 | C33, sp3 carbon with 3 Hydrogens |
| [CX4H4] | 44 | 36 | C34 | C34, sp3 carbon with 4 Hydrogens, methane |
| [CX4]([O])[O] | 45 | 37 | C3O2 | C3O2, sp3 carbon with 2 oxygens |
| [CX4r] | 46 | 38 | C3R | C3R, sp3 carbon on |

| | | | | |
|---|---|---|---|---|
| | | | | the cycle |
| [CX4r5] | 47 | 39 | C3R5 | C3R5, sp3 carbon on 5-membered cycle |
| [CX3R5] | 48 | 40 | C2R5 | C2R5, sp2 carbon on cycle |
| [CX4r5]([Or5]) | 49 | 41 | C3O1R5 | C3O1R5, sp3 carbon on 5-membered cycle bonded to 1 O on cycle |
| [CX4r5]([Or5])[Or5] | 50 | 42 | C3O2R5 | C3O2R5, sp3 carbon on 5-membered cycle bonded to 2 O on cycle |
| [CX4r6] | 51 | 43 | C3R6 | C3R6, sp3 carbon on 6-membered cycle |
| [CX4r3] | 52 | 44 | C3R3 | C3R3, sp3 carbon on 3-membered cycle |
| [CX4r4] | 53 | 45 | C3R4 | C3R4, sp3 carbon on 4-membered cycle |
| [C]1(CCO1) | 54 | 46 | C3R4O1 | C3R4O1, sp3 C in oxetane ring closer to O |
| [CX4r7] | 55 | 47 | C3R7 | C3R7, sp3 carbon in 7-membered ring |
| [CX3r7](=[O]) | 56 | 48 | C=OR7 | C=OR7, C=O in 7-membered ring (after CNN) |
| [CX4r6]([Or6]) | 57 | 49 | C3O1R | C3O1R, sp3 carbon on 6-membered cycle bonded to 1 O on cycle |
| [CX4r5]([NH1]) | 58 | 50 | C3N1R5 | C3N1R5, sp3 carbon on 5-membered cycle bonded to 1 NH |
| [CX4r5]([NH1])([NH1]) | 59 | 51 | C3N2R5 | C3N2R5, sp3 carbon on 5-membered cycle bonded to 2 NH |
| [CX4]([C](=O))([C](=O)) | 60 | 52 | C-Bket | C-Bket, sp3 carbon with 2 C=O |
| [$([#6](-[#7])(-[#6](=[#8])(-[#8])))] | 61 | 53 | CA | CA, sp3 carbon on alanine C chain |
| [C]([cr6])[cr6] | 62 | 54 | C3c6c6 | C3c6c6, sp3 C connected with 2 benzene rings |

| | | | | |
|---|---|---|---|---|
| [CX4]([F])[F] | 63 | 55 | CF2 | CF2, sp3 carbon with 2 F |
| [CX4]([F])([F])[F] | 64 | 56 | CF3 | CF3, sp3 carbon with 3 F |
| [CX4]([F])([F])([F])[F] | 65 | 57 | CF4 | CF4, sp3 carbon with 4 F |
| [CX4][Cl] | 66 | 58 | CCl1 | CCl1, sp3 carbon with 1 Cl |
| [CX4]([Cl])[Cl] | 67 | 59 | CCl2 | CCl2, sp3 carbon with 2 Cl |
| [CX4]([Cl])([Cl])[Cl] | 68 | 60 | CCl3 | CCl3, sp3 carbon with 3 Cl |
| [CX4]([Cl])([Cl])([Cl])[Cl] | 69 | 61 | CCl4 | CCl4, sp3 carbon with 4 Cl |
| [CX4][Br] | 70 | 62 | CBr1 | CBr1, sp3 carbon with 1 Br |
| [CX4]([Br])[Br] | 71 | 63 | CBr2 | CBr2, sp3 carbon with 2 Br |
| [CX4]([Br])([Br])[Br] | 72 | 64 | CBr3 | CBr3, sp3 carbon with 3 Br |
| [CX4]([Br])([Br])([Br])[Br] | 73 | 65 | CBr4 | CBr4, sp3 carbon with 4 Br |
| [$([CX3]=[CX3])] | 74 | 66 | C2C | C2C, sp2 carbon, CH2=CH2 |
| [$([#6](=[#6])-[#6]=[#6])] | 75 | 67 | C2C2C | C2C2C, sp2 carbon, interior C of CH2=CH-CH=CH2 |
| [$([#6](=[#6])-[#6]=[#8])] | 76 | 68 | C2C2O | C2C2O, sp2 carbon, interior C of CH2=CH-C=O |
| [CX3H1](=O) | 77 | 69 | CHO | C=O, sp2 carbon, aldehyde |
| [CX3H2](=O) | 78 | 70 | CH2O | C=O, sp2 carbon, formaldehyde |
| [CX3H1](=O)([CX3]([H])(=O)) | 79 | 71 | CHO-2 | C=O, sp2 carbon, glyoxal |
| [$([#6](=[#8])-[#6]=[#6])] | 80 | 72 | CHOC2C | CHOC2C, sp2 carbon, CHO of CHO-CH=CH2 |
| [$([#6](=[#8])([H])([ccccc]))] | 81 | 73 | CHOben | CHOben, sp2 carbon, CHO of CHO-Ph |
| [CX3H0](=O) | 82 | 74 | C-Ket | C=O, sp2 carbon, ketone |
| [CX3H0](=O)([C](=O) | 83 | 75 | C-PRV | C=O, sp2 carbon, |

| | | | | |
|---|---|---|---|---|
| ([O])) | | | | pyruvate |
| [CX3](=O)([OH0]) | 84 | 76 | CO=O- | C=O, sp2 carbon, carboxylate |
| [CX3](=O)([OH0])(-[#6]=[#6]) | 85 | 77 | COOC2C- | C=O, sp2 carbon, carboxylate bonded to C=C |
| [CX3](=O)([OH0])[ccccc] | 86 | 78 | COOben- | C=O, sp2 carbon, carboxylate on benzene |
| [$([#6](-[#6](-[#8])=[#8])(=[#8])(-[OH0]))] | 87 | 79 | C-Oxa- | C=O, sp2 carbon, oxalate |
| [CX3](=O)([OH1]) | 88 | 80 | CO=O | C=O, sp2 carbon, carboxylic acid |
| [$([CX3](=O)([O][C]))] | 89 | 80 | CO=O | C=O, sp2 carbon, carboxylic ester |
| [CX3]1(=[O])(OCCCC1) | 90 | 81 | COOR6 | COOR6, |
| [CX3]1(=[O])(OCCC1) | 91 | 82 | COOR5 | COOR5, |
| [CX3](=O)([SH1]) | 92 | 83 | CS=O | C=O, sp2 carbon, thioate acid |
| [$([CX3](=O)([S][C]))] | 93 | 83 | CS=O | C=O, sp2 carbon, thioate ester |
| [CX3](=O)([OH1,OC1])(-[#6]=[#6]) | 94 | 84 | COOC2C | C=O, sp2 carbon, carboxylic acid bonded to C=C |
| [$([CX3](=O)([O][C]))(-[#6]=[#6]))] | 95 | 84 | COOC2C | C=O, sp2 carbon, carboxylic ester bonded to C=C |
| [CX3](=O)([OH1])[ccccc] | 96 | 85 | COOben | C=O, sp2 carbon, carboxylic acid on benzene |
| [$([CX3](=O)([O][C]))[cccccc])] | 97 | 85 | COOben | C=O, sp2 carbon, carboxylic ester on benzene |
| [$([#6](-[#6](-[#8])=[#8])(=[#8])(-[OH]))] | 98 | 86 | C-Oxa | C=O, sp2 carbon, oxalic acid |
| [CX3](=[OX1])[OX2][CX3](=[OX1]) | 99 | 87 | CO=OC | C=O, sp2 carbon, anhydride |
| [CX3](=O)([OH0])[OH0] | 100 | 88 | CO3-- | C=O, sp2 carbon, carbonate |
| [CX3](=O)([OH1])[OH | 101 | 89 | CO3- | C=O, sp2 carbon, |

| | | | | |
|---|---|---|---|---|
| 0] | | | | bicarbonate |
| [CX3](=O)(O[C])(O[C]) | 102 | 90 | CO3 | C=O, sp2 carbon, carbonic ester |
| [CX3](=O)(O[C])(O[H]) | 103 | 91 | CO3H | C=O, sp2 carbon, bicarbonate ester |
| [CX3](=O)([NX3]) | 104 | 92 | CON | CON, sp2 carbon, peptide bond (amide) |
| [Cr5](=O)[Nr5] | 105 | 93 | CONR5 | CONR5, CON in 5-membered ring |
| [CX3]1(=[O])(NCCCC1) | 106 | 94 | CONR6 | CONR6, lactam |
| [C](=O)([NX3])([ccccc]) | 107 | 95 | CONben | CONben, CON connected to benzene ring |
| [CX3](=O)([O])([NX3]) | 108 | 96 | C-carb | C-carb, sp2 carbon, carbamic acid |
| [CX3](=O)([F]) | 109 | 97 | COF | COF, sp2 carbon, acyl fluoride |
| [CX3](=O)([Cl]) | 110 | 98 | COCl | COCl, sp2 carbon, acyl chloride |
| [CX3](=O)([Br]) | 111 | 99 | COBr | COBr, sp2 carbon, acyl bromide |
| [C](=O)([O]([OH0])) | 112 | 100 | COOO- | COOO-, C=O in peracetate ion |
| [C](=O)([O]([OH1,OC1])) | 113 | 101 | COOOH | COOOH, C=O in peracetic acid or ester |
| [C]=[N] | 114 | 102 | C=N | C=N, sp2 carbon, formimidamide |
| [CX3R]=[N] | 115 | 103 | C=NR | C=N, sp2 carbon, connected to Nitrogen in a ring |
| [CH0]([c])([N])[NH2] | 116 | 104 | C-A6 | C-A6, 6'-C in Adenine |
| [C](=O)([cr6])[cr6] | 117 | 105 | C=Oc6c6 | C=Oc6c6, C=O connected with 2 benzene rings |
| [CH1R](=[N])([N]([c])) | 118 | 106 | C-A2 | C-A2, 2'-C in Adenine |
| [C](=[NH2])(N)[N] | 119 | 107 | C+ | C+, positively charged carbon, C+(NH2)3 and C+(NH2)2(NHCH3) |
| [$([CX2]#[CX2])] | 120 | 108 | C1C | C1C, sp carbon, triple bond |
| [$([CX2](#[CX2])-[CX2]#[CX2])] | 121 | 109 | C1C1C | C1C1C, sp carbon, triple bond bonded |

| | | | | |
|---|---|---|---|---|
| | | | | with another triple bond |
| [CX2]#[N] | 122 | 110 | C1N | C1N, sp carbon, nitrile |
| [$([CX3N1]=[CX3N1])] | 123 | 111 | CCN | CCN, sp2 carbon, C=C in imidazole with positive charge |
| [CX3N2H1] | 124 | 112 | CNN | CNN, sp2 carbon, C between Ns in imidazole with positive charge |
| [CX3]=[S] | 125 | 113 | C=S | C=S, Carbon of C=S bond |
| [CX2](#[NX1])[SX2] | 126 | 114 | C-SCN | |
| [C]1([O,S]C=CC=C1) | 127 | 115 | C3-pyran | C3-pyran, sp3 C in pyran ring |
| [$([C]1(=CC=C[O,S]C1)),$([C]1(=CC[O,S]C=C1)),$([C]1(=CC=CC[O,S]1)),$([C]1(=C[O,S]CC=C1))] | 128 | 116 | C2-pyran | C2-pyran, sp2 C in pyran ring |
| [c]([c])[n] | 129 | 117 | ccn | ccn, aromatic carbon bonded to aromatic nitrogen |
| [cr5]([cr5])[nH0r5] | 130 | 118 | ccnr5 | ccnr5, aromatic carbon bonded to aromatic nitrogen in 5-membered cycle |
| [c]([c]([c](=O)))([c]) | 131 | 119 | c-PryO | c-PryO, side sp2-C in 2-Prydone close to aromatic C=O |
| [cH1]([c]([nH1]))([c]) | 132 | 120 | c-PryN | c-PryN, side sp2-C in 2-Prydone close to aromatic NH |
| [c]([c])([c])[n] | 133 | 121 | cccn | cccn, aromatic carbon bonded to aromatic nitrogen without H |
| [cH0]([c])[nH1] | 134 | 122 | c0cn1 | c0cn1, aromatic carbon bonded to aromatic NH and no H |
| [cH1]([c])[nH1] | 135 | 123 | c1cn1 | c1cn1, aromatic carbon bonded to aromatic NH and 1 H |
| [cH1r5]([cr5])[nH1r5] | 136 | 124 | c1cn1r5 | c1cn1r5, aromatic carbon bonded to |

| | | | | |
|---|---|---|---|---|
| | | | | aromatic NH and 1 H in 5-membered cycle |
| [c]([c])([C]([NH2]))[n] | 137 | 125 | c-A5 | c-A5, 5'-C in Adenine |
| [c]([n])[NH2] | 138 | 126 | cnN2 | cnN2, aromatic carbon bonded to 1 aromatic nitrogens and aliphatic NH2 |
| [c]([c])([nH0])[NH2] | 139 | 127 | ccnN2 | ccnN2, aromatic carbon bonded to aromatic nitrogen and aliphatic NH2 without H |
| [c]([c])([nH1])[NH2] | 140 | 128 | ccn1N2 | ccn1N2, aromatic carbon bonded to aromatic NH and aliphatic NH2 without H |
| [c]([c]([n]))([n])[NH0] | 141 | 129 | c-A4 | c-A4, 4'-C in Adenine |
| [c]([n])[n] | 142 | 130 | cnn | cnn, aromatic carbon bonded to 2 aromatic nitrogens |
| [c]([n])[nH1] | 143 | 131 | cnn1 | cnn1, aromatic carbon bonded to 2 aromatic nitrogens and H |
| [cr5]([nr5])[nH1r5] | 144 | 132 | cnn1r5 | cnn1r5, aromatic carbon bonded to 2 aromatic nitrogens and H in 5-membered cycle |
| [c]([n])([n])[c]([n]) | 145 | 133 | c-G4 | c-G4, 4'-C in Guanine and C in similar structure |
| [c]([n])([n])[N] | 146 | 134 | cnnN | cnnN, aromatic carbon bonded to 2 aromatic nitrogens and aliphatic N |
| [c]([c])[c](=O) | 147 | 135 | ccco | ccco, aromatic carbon bonded to aromatic C=O |
| [c]([c](=O))([c]([n])([n]))[n] | 148 | 136 | c-G5 | c-G5, 5'-C in Guanine and C in similar structure |
| [c](=O)([c])[nH1] | 149 | 137 | c=On1 | c=On1, aromatic C=O bonded to aromatic |

| | | | | |
|---|---|---|---|---|
| | | | | NH and carbon |
| [c](=O)([n])[nH1] | 150 | 138 | c=Onn1 | c=Onn1, aromatic C=O bonded to aromatic N and NH |
| [c](=O)([nH1])[nH1] | 151 | 139 | c=On1n1 | c=On1n1, aromatic C=O bonded to 2 aromatic NH |
| [c](=O)([c]([n]))[nH1] | 152 | 140 | c-G6 | c-G6, 6'-C in Guanine and C in similar structure |
| [c]1(ccco1) | 153 | 141 | c-Fur-1 | c-Fur-1, c in Furan ring closer to O |
| [c]1(cocc1) | 154 | 142 | c-Fur-2 | c-Fur-2, c in Furan ring not near O |
| [c]1(cccs1) | 155 | 143 | c-Thp-1 | c-Thp-1, c in Thiophene ring closer to S |
| [c]1(cscc1) | 156 | 144 | c-Thp-2 | c-Thp-2, c in Thiophene ring not near S |
| [c]1(ccnn1) | 157 | 145 | c-Prz-1 | c-Prz-1, c in Pyrazole ring closer to n |
| [c]1(cnnc1) | 158 | 146 | c-Prz-2 | c-Prz-2, c in Pyrazole ring which is not near n |
| [c]1(ncco1) | 159 | 147 | c-Oxz-1 | c-Oxz-1, c in Oxazole ring between n and o |
| [c]1(ncoc1) | 160 | 148 | c-Oxz-n | c-Oxz-n, c in Oxazole ring closer to n |
| [c]1(ocnc1) | 161 | 149 | c-Oxz-o | c-Oxz-o, c in Oxazole ring closer to o |
| [c]1(nccs1) | 162 | 150 | c-Thz-1 | c-Thz-1, c in Thiazole ring between n and s |
| [c]1(ncsc1) | 163 | 151 | c-Thz-n | c-Thz-n, c in Thiazole ring closer to n |
| [c]1(scnc1) | 164 | 152 | c-Thz-o | c-Thz-o, c in Thiazole ring closer to s |
| [c]1(ccccc1)([c]2(cccc c2)) | 165 | 153 | c6-6 | c6-6, The connecting C between 2 benzene rings |
| [c]1(ccccc1)([c]2(nccc 2)) | 166 | 154 | c6-cn5 | c6-cn5, connecting c between 6-ring and 5-ring with n |
| [c]1(nccc1)([c]2(ccccc | 167 | 155 | cn5-c6 | cn5-c6, connecting c |

| | | | | |
|---|---|---|---|---|
| 2)) | | | | between 5-ring with n and 6-ring |
| [c]1(ncccc1)([c]2(nccn2)) | 168 | 156 | cn6-cnn5 | cn6-cnn5, connecting c between 6-ring with n and 5-ring with 2 n |
| [c]1(nccn1)([c]2(nccc2)) | 169 | 157 | cnn5-cn6 | cnn5-cn6, connecting c between 5-ring with 2 n and 6-ring with n |
| [c]1(ncccn1)([c]2(nccn2)) | 170 | 158 | cnn6-cnn5 | cnn6-cnn5, connecting c between 6-ring with 2 n and 5-ring with 2 n |
| [c]1(nccn1)([c]2(ncccn2)) | 171 | 159 | cnn5-cnn6 | cnn5-cnn6, connecting c between 5-ring with 2 n and 6-ring with 2 n |
| [c]1(ncccc1)([n]2(nccc2)) | 172 | 160 | cn6-nnc5 | cn6-nnc5, connecting c to n between 6-ring with n and 5-ring with 2 n |
| [c]1(ncccn1)([n]2(nccc2)) | 173 | 161 | cnn6-nnc5 | cnn6-nnc5, connecting c to n between 6-ring with 2 n and 5-ring with 2 n |
| [c]1(ncccc1)([c]2(nccs2)) | 174 | 162 | cn6-cns5 | cn6-cns5, connecting c between 6-ring with n and 5-ring with n & s |
| [c]1(nccs1)([c]2(ncccc2)) | 175 | 163 | cns5-cn6 | cns5-cn6, connecting c between 5-ring with n & s and 6-ring with n |
| [c]1(ncccc1)([c]2(ccco2)) | 176 | 164 | cn6-co5 | cn6-co5, connecting c between 6-ring with n and 5-ring with o |
| [c]1(ccco1)([c]2(ncccc2)) | 177 | 165 | co5-cn6 | co5-cn6, connecting c between 5-ring with o and 6-ring with n |
| [c]1(ccccc1)([c]2(cccn2)) | 178 | 166 | c6-cn6 | c6-cn5, connecting c between 6-ring and 6-ring with n |
| [c]1(ccccn1)([c]2(ccccc2)) | 179 | 167 | cn6-c6 | cn6-c6, connecting c between 6-ring with n and 6-ring |
| [c]1(ccco1)([c]2(ncccc2)) | 180 | 165 | co5-cn6 | co5-cn6, connecting c between 6-ring with n |

| | | | | |
|---|---|---|---|---|
| | | | | and 6-ring with o |
| [c]1(nnnn1) | 181 | 168 | cnnnnr | cnnnnr, |
| [n] | 182 | 169 | nar | nar, Aromatic Nitrogen in a ring |
| [NX4] | 183 | 170 | N4 | N4, Nitrogen in Ammonium |
| [NX3] | 184 | 171 | N3 | N3, sp3 Nitrogen |
| [NX3r5] | 185 | 172 | N3R5 | N3R5, sp3 Nitrogen on 5-membered cycle |
| [NX4r5] | 186 | 173 | N4R5 | N4R5, sp3 N with charge in 5-membered ring |
| [NX3r6] | 187 | 174 | N3R6 | N3R6, sp3 N in 6-membered ring |
| [NX4r6] | 188 | 175 | N4R6 | N4R6, sp3 N in 6-membered ring with charge |
| [Nr7] | 189 | 176 | N3R7 | N3R7, sp3 N in 7-membered ring |
| [NX3H3] | 190 | 177 | N3H3 | N3H3, ammonia N |
| [NX3H2]([C](=[N])) | 191 | 178 | N3C=N | N3C=N, |
| [NX3;!r]([CX3](=O)) | 192 | 179 | NC=O | NC=O, sp3 Nitrogen in amide |
| [NX3H2]([C](=O)([O])) | 193 | 180 | N-carbH | N-carbH, sp3 Nitrogen in carbamic acid with H |
| [NX3H1]([C](=O)([O])) | 194 | 180 | N-carbH | N-carbH, sp3 Nitrogen in carbamic acid with H |
| [NX3H0]([C](=O)([O])) | 195 | 181 | N-carb | N-carb, sp3 Nitrogen in carbamic acid with no H |
| [NX3]([CX3](=O))([CX3](=O)) | 196 | 182 | N-Imi | N-Imi, sp3 Nitrogen in imide |
| [NX3](=[O])[C] | 197 | 183 | NO2 | NO2, nitryl |
| [NX3](=[O])[cccccc] | 198 | 184 | NO2ben | NO2ben, nitryl on benzene |
| [NX3](-[O])[O] | 199 | 185 | NOO | NOO, nitrite |
| [NX3H0]([NX3H0c0])[cccccc] | 200 | 186 | N2N | N2N, N in azobenzene |
| [NX1] | 201 | 187 | NT | , Triple bond Nitrogen |
| [NX1](#[CX2]([SX2])) | 202 | 188 | N-SCN | N-SCN |
| [NX3H0]([N])[cccccc] | 203 | 189 | N1Ni | N, Triple bond |

| | | | | |
|---|---|---|---|---|
| | | | | Nitrogen in diazonium (bond to other atoms) |
| [NX3H0]([N])[C] | 204 | 189 | N1Ni | N, Triple bond Nitrogen in diazonium (bond to other atoms) |
| [NX3]([NH0][cccccc]) | 205 | 190 | N1No | N, Triple bond Nitrogen in diazonium (outer side) |
| [NH0]([NH0][cccccc])[cccccc] | 206 | 191 | N-Azo | N-Azo, N in Azobenzene |
| [NX2]=[C] | 207 | 192 | N=C | N=C, Double bond N=C |
| [NH2]=[C] | 208 | 193 | N=C+ | N=C+, Double bond N+ |
| [NH2]-[C](=[NH2]) | 209 | 193 | N=C+ | N=C+, Double bond N+ |
| [NH1]-[C](=[NH2]) | 210 | 193 | N=C+ | N=C+, Double bond N+ |
| [NH1]([C])([C](=[NH2])) | 211 | 194 | N=CC+ | N=C+, Double bond N+ with 2 C |
| [NX2]([C]([NH2]))(=[CH1]) | 212 | 195 | N-A1 | N-A1, 1'-N in Adenine |
| [NH0]([c]([c]))([CH]) | 213 | 196 | N-A3 | N-A3, 3'-N in Adenine |
| [NH2]([cccccc]) | 214 | 197 | NH2ben | NH2, Aniline N |
| [NH3]([cccccc]) | 215 | 198 | NH3ben | NH3, Phenylazanium N |
| [NX3]([cr6])[cr6] | 216 | 199 | N3c6c6 | N3c6c6, sp3 N connected with 2 benzene rings |
| [nX2H0] | 217 | 200 | n0 | n0, sp2 aromatic nitrogen with no H |
| [nX2H0]([c]([NH2]))([cH0]) | 218 | 201 | n-G3 | n-G3, 3'-G in Guanine |
| [nX3H1] | 219 | 202 | n1 | n1, sp3 aromatic nitrogen with H |
| [nX3]([c](=O)) | 220 | 203 | nc=O | nc=O, sp3 aromatic nitrogen bonded to aromatic C=O |
| [nX3]([c](=O))([c](=O)) | 221 | 204 | n2c=O | n2c=O, sp3 aromatic nitrogen bonded to 2 aromatic C=O |
| [n]1(nccc1)([c]2(nccc2)) | 222 | 205 | nnc5-cn6 | nnc5-cn6, connecting n to c between 5-ring with 2 n and 6-ring |

| | | | | |
|---|---|---|---|---|
| | | | | with n |
| [n]1(nccc1)([c]2(nccc n2)) | 223 | 206 | nnc5-cnn6 | nnc5-cnn6, connecting n to c between 5-ring with 2 n and 6-ring with 2 n |
| [n]1(cccn1) | 224 | 207 | n-Prz | n-Prz, n in Pyrazole |
| [n]1(cocc1) | 225 | 208 | n-Oxz | n-Oxz, n in Oxazole |
| [n]1(cscc1) | 226 | 209 | n-Thz | n-Thz, n in Thiazole |
| [n]1(ncnn1) | 227 | 210 | nncnnr | nncnnr, |
| [n]1(nnnc1) | 228 | 211 | nnnncr | nnnncr, |
| [OX2H1] | 229 | 212 | OH | OH, Oxygen of Hydroxyl group |
| [OX2H1]([C](=O)) | 230 | 213 | OCO | OCO, Oxygen of Carboxylic acid (protonated) |
| [OX2H1]([P]) | 231 | 214 | O-P | O-P, Oxygen of phosphate, with H |
| [OX2H0]([P]) | 232 | 215 | O-P- | O-P, Oxygen of phosphate, ionic state |
| [O]([N])(-[C]) | 233 | 216 | O-N | O-N, Oxygen of nitrite |
| [OX2H1]([SX4]) | 234 | 217 | O-S | O-S, Oxygen of Sulfoxide |
| [OX2]([N])(-[!C]) | 235 | 218 | O=N | O=N, Oxygen of nitrite |
| [OX1](=N) | 236 | 218 | O=N | O=N, Oxygen of nitryl |
| [OX1]=[CX3][OH1] | 237 | 219 | O=C | O=C, Oxygen of carbonyl group, Acetic Acid |
| [OX1]=[CX3] | 238 | 219 | O=C | O=C, Oxygen of carbonyl group |
| [$([OX1]=[CX3][H])] | 239 | 219 | O=C | O=C, Oxygen of carbonyl group, O=C-H |
| [OX1]=[CX3][N] | 240 | 219 | O=C | O=C, Oxygen of carbonyl group, Amide, Formamide |
| [OX1](=[P]) | 241 | 220 | O=P | O=P, Oxygen of phosphate |
| [OX2H0]([P])[C] | 242 | 221 | OP-est | OP-est, Oxygen in phosphate (bonded to C) |
| [OX2]([P])[P] | 243 | 222 | O2P | O2P, Oxygen to connect 2 P |
| [OX2H0]([C])[C] | 244 | 223 | OR2 | OR2, Oxygen of ether |

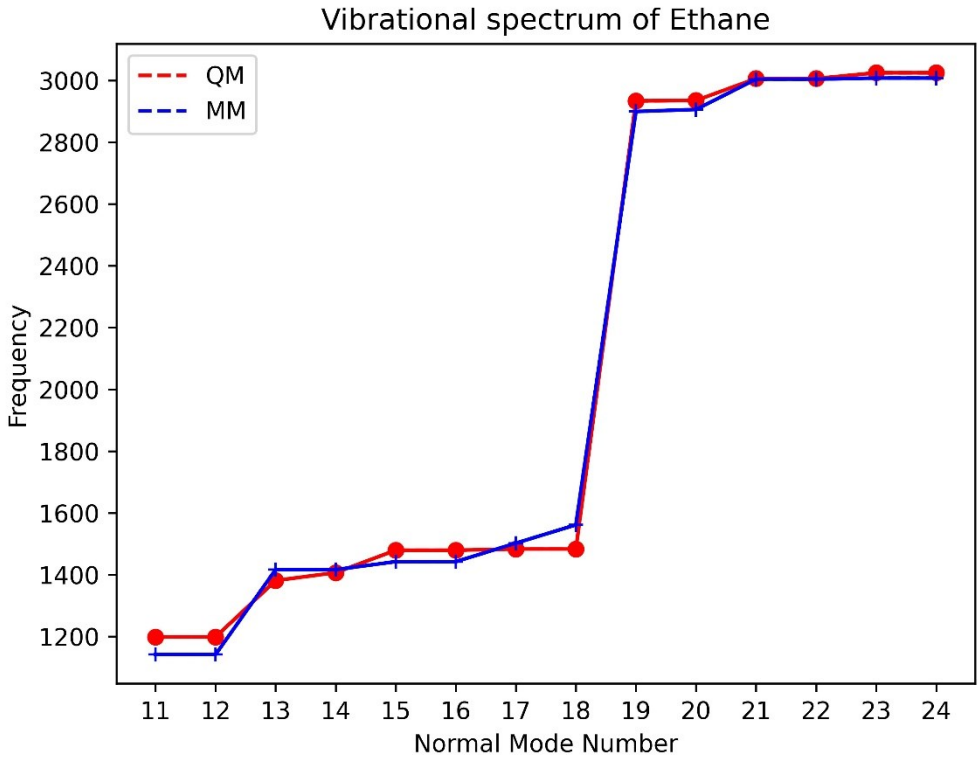
| | | | | |
|---|---|---|---|---|
| [OX2r6]([C])[C] | 245 | 224 | OR6 | OR6, Oxygen in 6-membered cycle |
| [$([OH0]=[C]([C,N])[OH0]),$([OH0][C]([C,N])=[OH0])] | 246 | 225 | O- | O-, Oxygen of carboxinate |
| [$([OH0]([CX3H1](=[O]))),$([OH0](=[CX3H1]([OH0]))),$([OH0]([CX3H0](=O)([OH0,OH1]))) | 247 | 225 | O- | O-, Oxygen of carboxinate |
| [$([OH0]=[C]([cccccc])[OH0]),$([OH0][C]([ccccc])=[OH0])] | 248 | 225 | O- | O-, Oxygen of carboxinate on benzene |
| [O]=[c] | 249 | 226 | Oar | Oar, sp2 Oxygen connected to aromatic ring |
| [OX2H0]([C](=O))[C] | 250 | 227 | O-est | O-est, Oxygen of ester |
| [$([OH0]=[C]([O]([C])))] | 251 | 219 | O=C | O=C, Oxygen of carbonyl group (correct the O=C in ester) |
| [OH0]c1ccccc1 | 252 | 228 | O-Phe- | O on Phenol minus |
| [OH1]([cccccc]) | 253 | 229 | O-Phe | O on Phenol |
| [O;!r]([cccccc])([C]) | 254 | 230 | O-Phe-est | O on Phenol |
| [OX2H0]([SX4]) | 255 | 231 | O-S- | O-S-: Oxygen of sulfonate |
| [O]=[SX3] | 256 | 232 | O=S | O=S, Oxygen of DMSO |
| [O]=[SX4] | 257 | 233 | O=S=O | O=S=O, Oxygen of sulfone |
| [$([O]([S]([O])([O][C])))] | 258 | 234 | OSO2 | OSO2, oxygen of EMsulfonate |
| [OH0]([S]([cccccc])) | 259 | 235 | O-Sben- | O-Sben-, Oxygen of benzene sulfonate |
| [OH1]([S]([cccccc])) | 260 | 236 | O-Sben | O-Sben, Oxygen of benzene sulfonic acid |
| [OX2H0]([SX4]([OH1])) | 261 | 237 | O=SO2 | O=SO2: Oxygen of Sulfonic acid |
| [OX2H0]([SX4]([O]([C]))([OH0]) | 262 | 237 | O=SO2 | O=SO2: Oxygen of Sulfonic acid |
| [O]([S]([O])([O]))([C]) | 263 | 238 | OS-est | OS-est, oxygen of EMsulfonate |
| [OX2r5]([CX4r5]([OH | 264 | 239 | Oxo | Oxo, example: |

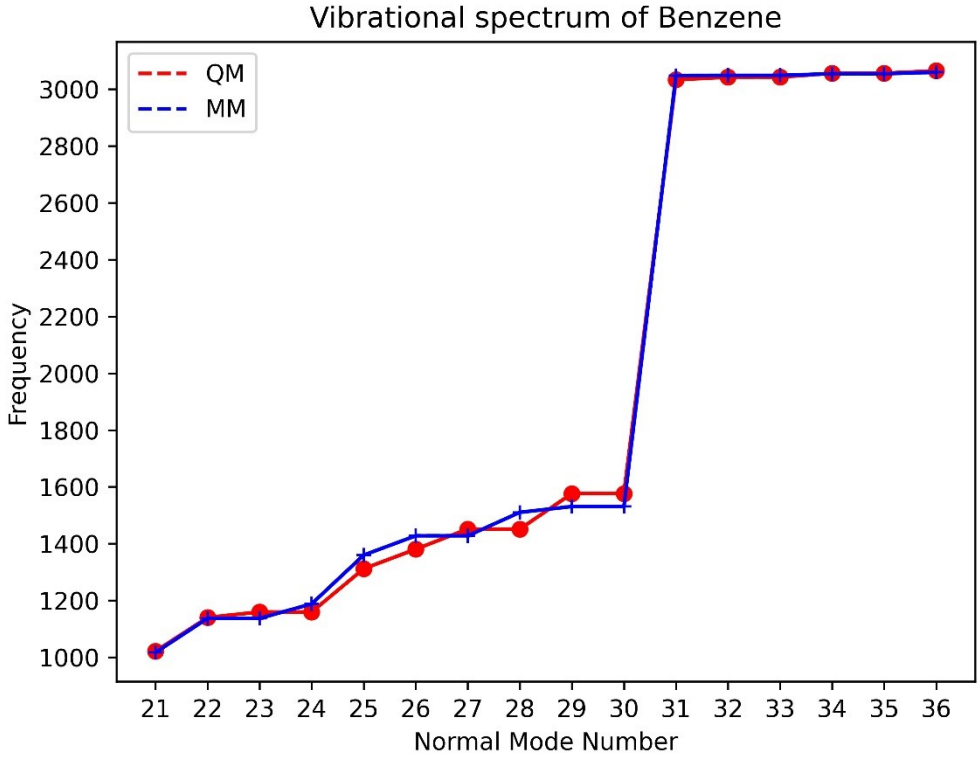
| | | | | |
|---|---|---|---|---|
| 0r5])) | | | | benzodioxole |
| [OX2r5]([CX4r5]([OH 0r5]))[cccccc] | 265 | 239 | Oxo | Oxo, example: benzodioxole |
| [O][C](=[O])([NH2]) | 266 | 225 | O- | O-, example: carbamate |
| [o]1cccc1 | 267 | 240 | O-fur | O-fur, O on furan ring |
| [O](=[C]([O]([OH0]))) | 268 | 241 | O=COO- | O=COO-, O=C in peracetate ion |
| [O](=[C]([O]([OH1]))) | 269 | 242 | O=COOH | O=COOH, O=C in peracetic acid or ester |
| [O]([C](=[O]))([OH0]) | 270 | 243 | OCO-1- | OCO-1-, O in the middle of (C=O)-O-O- |
| [OH0]([O]([C](=[O]))) | 271 | 244 | OCO-2- | OCO-2-, O in the side of (C=O)-O-O- |
| [O]([C](=[O]))([OH1,OC1]) | 272 | 245 | OCO-1 | OCO-1, O in the middle of peracetic acid or ester |
| [OH1]([O]([C](=[O]))) | 273 | 246 | OCO-2 | OCO-2, O in the side of peracetic acid |
| [OC1]([O]([C](=[O])))([C]) | 274 | 247 | OCO-est | OCO-est, O in the side of peracetic ester |
| [O]1(CC=CC=C1) | 275 | 248 | O-pyran | O-pyran, O in pyran ring |
| [o]1(cncc1) | 276 | 249 | o-Oxz | o-Oxz, o in Oxazole |
| [SX2H2] | 277 | 250 | SH2 | SH2, Sulphur in H2S |
| [SX2H0] | 278 | 251 | S3- | S3-, Sulphur in thiolate |
| [SX2H0]([C])[C] | 279 | 252 | SC2 | SC2, Sulphur in SC2 |
| [SX2H1] | 280 | 253 | S3 | S3, Sulphur in thiol |
| [SX2H1]([C](=O)) | 281 | 254 | SCO | SCO, Sulfur of thioate acid (protonated) |
| [SX2H0]([C](=O))[C] | 282 | 255 | S-est | S-est, Sulfur of thioate ester |
| [SX2]([S])[C] | 283 | 256 | SS | SS, Sulphur in disulfide |
| [SX3]=[O] | 284 | 257 | S=O | S=O, Sulphur in DMSO |
| [SX4](=[O])(=[O]) | 285 | 258 | SO2 | SO2, Sulphur in sulfone |
| [SX4]([OH0])([OH0])[OH0] | 286 | 259 | SO3- | SO3-, Sulphur in sulfonate |
| [SX4]([OH1])([OH0])([OH0]) | 287 | 260 | SO3 | SO3, Sulphur in sulfonic acid |

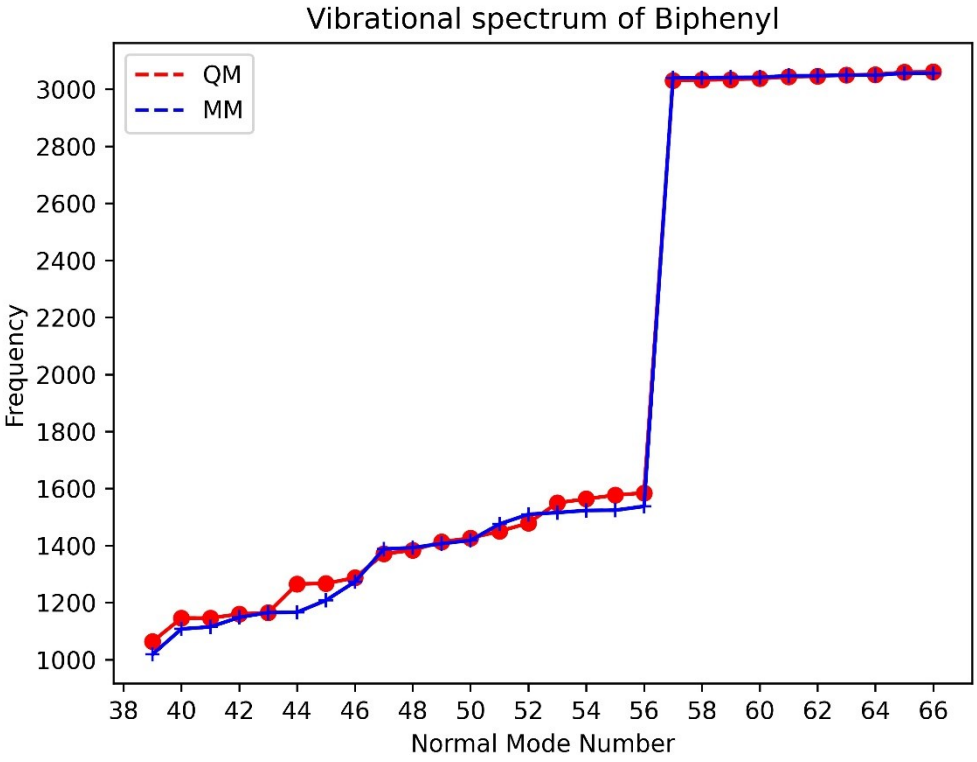
| | | | | |
|---|---|---|---|---|
| [SX4]([O]([C]))([OH0])([OH0]) | 288 | 260 | SO3 | SO3, Sulphur in sulfonic ester |
| [SX4]([OH0])([OH0])([OH0])[cccccc] | 289 | 261 | SO3ben- | SO3ben-, Sulphur in sulfonate on benzene |
| [SX4]([OH1])([OH0])([OH0])[cccccc] | 290 | 262 | SO3ben | SO3ben, Sulphur in sulfonic acid on benzene |
| [SX4]([O]([C]))([OH0])([OH0])[cccccc] | 291 | 262 | SO3ben | SO3ben, Sulphur in sulfonic acid on benzene |
| [SX2]([CX2](#[NX1])) | 292 | 263 | S-SCN | S-SCN, |
| [SX2]([CX2](#[NX1]))[cccccc] | 293 | 264 | S-SCNben | S-SCNben, |
| [S](=O)(=O)([c]) | 294 | 265 | SO2ben | SO2ben, S in sulfone bonded to the benzene ring |
| [S](=O)(=O)([c])([NX3]) | 295 | 266 | SO2Nben | SO2Nben, S in sulfone bonded to the benzene ring and sp3 N |
| [S]1(CC=CC=C1) | 296 | 267 | S-pyran | S-pyran, S in pyran ring |
| [s]1(cccc1) | 297 | 268 | s-Thp | s-Thp, s in Thiophene |
| [s]1(cncc1) | 298 | 269 | s-Oxz | s-Oxz, s in Thiazole |
| [PH3] | 299 | 270 | PH3 | PH3, Phosphor in PH3 |
| [PX4](=[O])([OH0])([OH0])[OH0] | 300 | 271 | PO4 | PO4, Phosphor in Phosphate |
| [PX4](=[O])([OH1])([OH0])[OH0] | 301 | 272 | PO4H1 | PO4H1, Phosphor in Phosphate |
| [PX4](=[O])([OH1])([OH1])[OH0] | 302 | 273 | PO4H2 | PO4H2, Phosphor in Phosphate |
| [PX4](=[O])([OH1])([OH1])[OH1] | 303 | 274 | PO4H3 | PO4H3, Phosphor in Phosphate |
| [PX4](=[O])([O][C])([O])[O] | 304 | 275 | PO4C1 | PO4C1, Phosphor in Phosphate |
| [PX4](=[O])([O][C])([O][C])[O] | 305 | 276 | PO4C2 | PO4C2, Phosphor in Phosphate |
| [PX4](=[O])([O][C])([O][C])([O][C]) | 306 | 277 | PO4C3 | PO4C3, Phosphor in Phosphate |
| [PX4](=[O])([O]([P]))([OH0])[OH0] | 307 | 278 | P-pyro | P-pyro, Phosphor in Pyrophosphate |
| [PX4](=[O])([O]([P]))([O]([P]))[OH0] | 308 | 279 | P-tri | P-tri, Phosphor in the middle of |

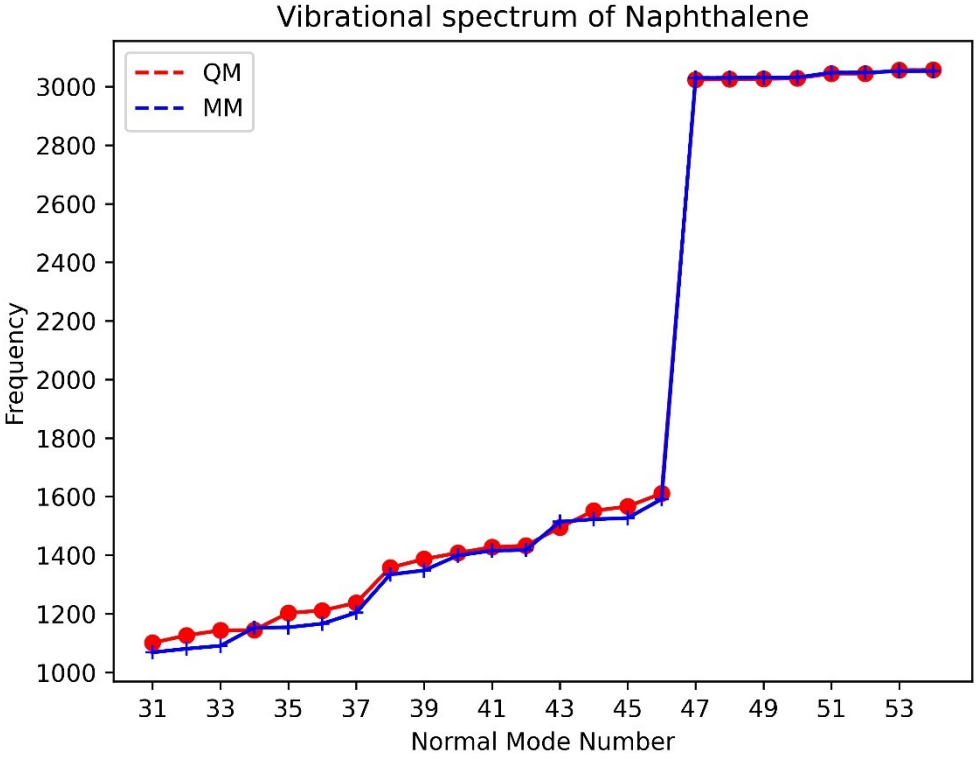
| | | | | |
|---|---|---|---|---|
| | | | | Triphosphate |

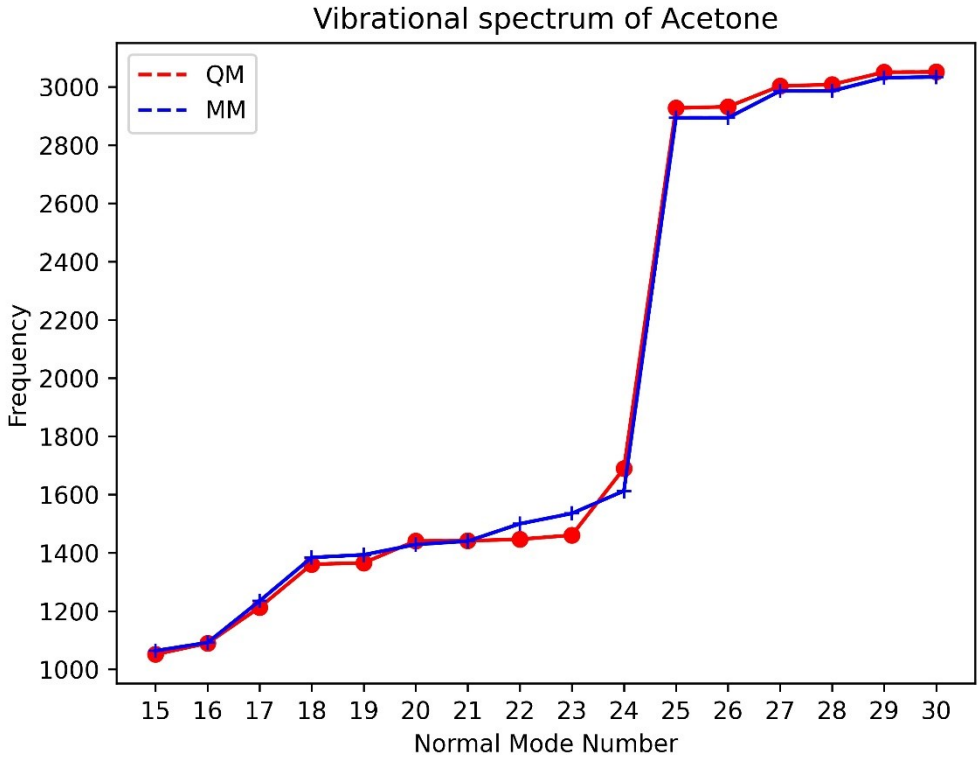
**Figure S1.** The normal mode frequencies of 10 common molecules by QM and MM. (Units: $cm^{-1}$)

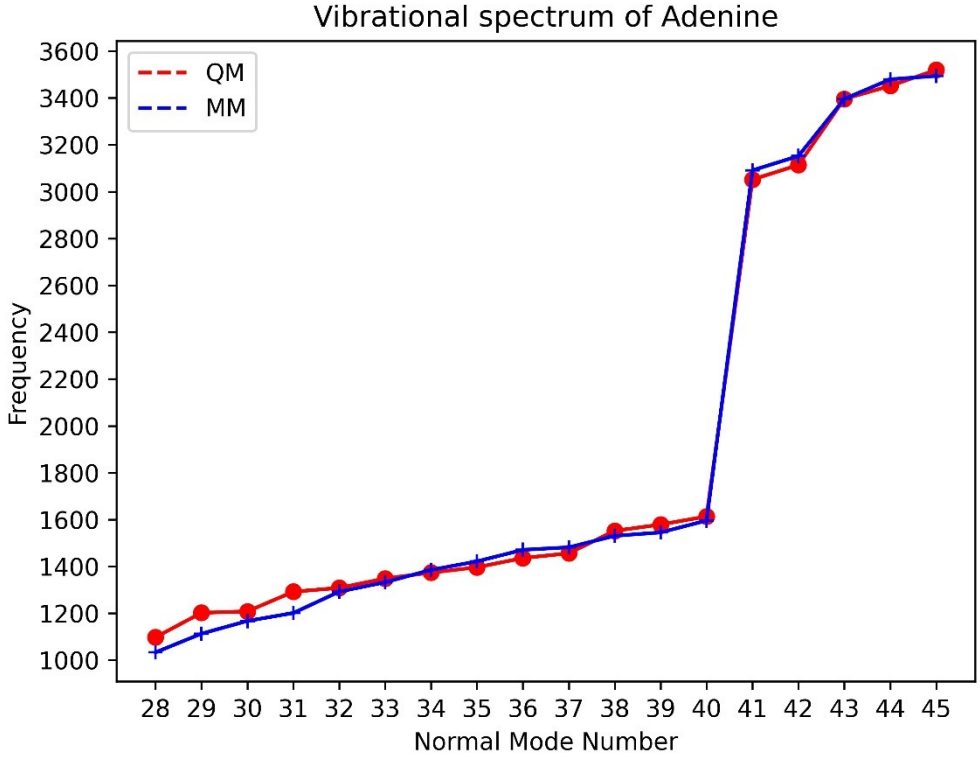


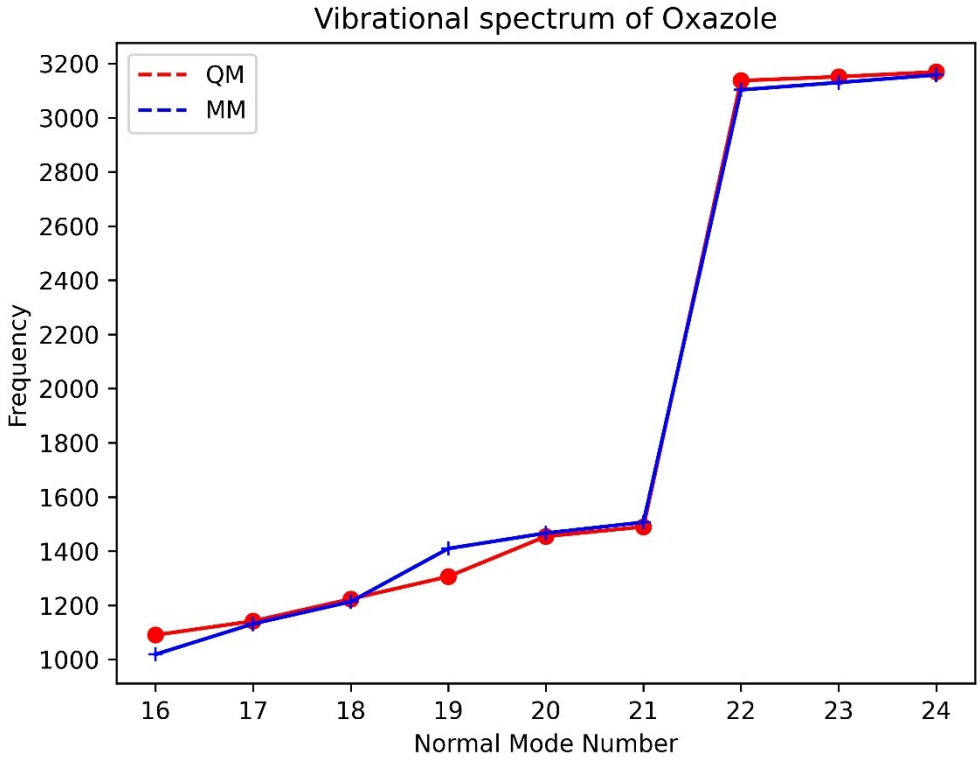



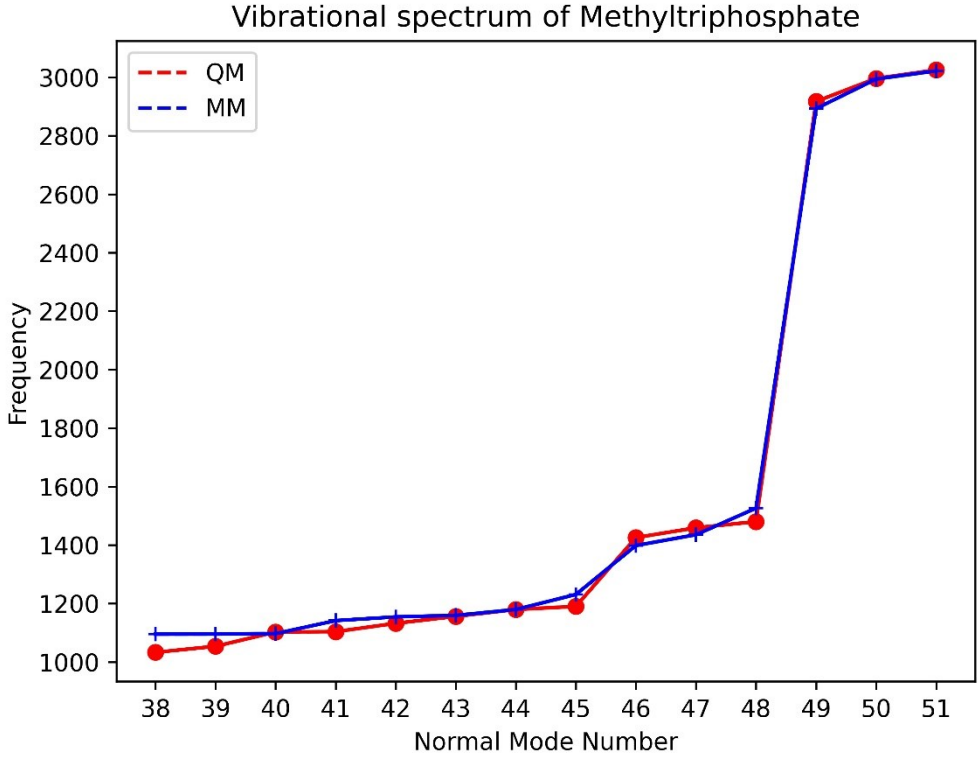

Vibrational spectrum of Biphenyl
QM
MM
Frequency
3000
2800
2600
2400
2200
2000
1800
1600
1400
1200
1000
38
40
42
44
46
48
50
52
54
56
58
60
62
64
66
Normal Mode Number

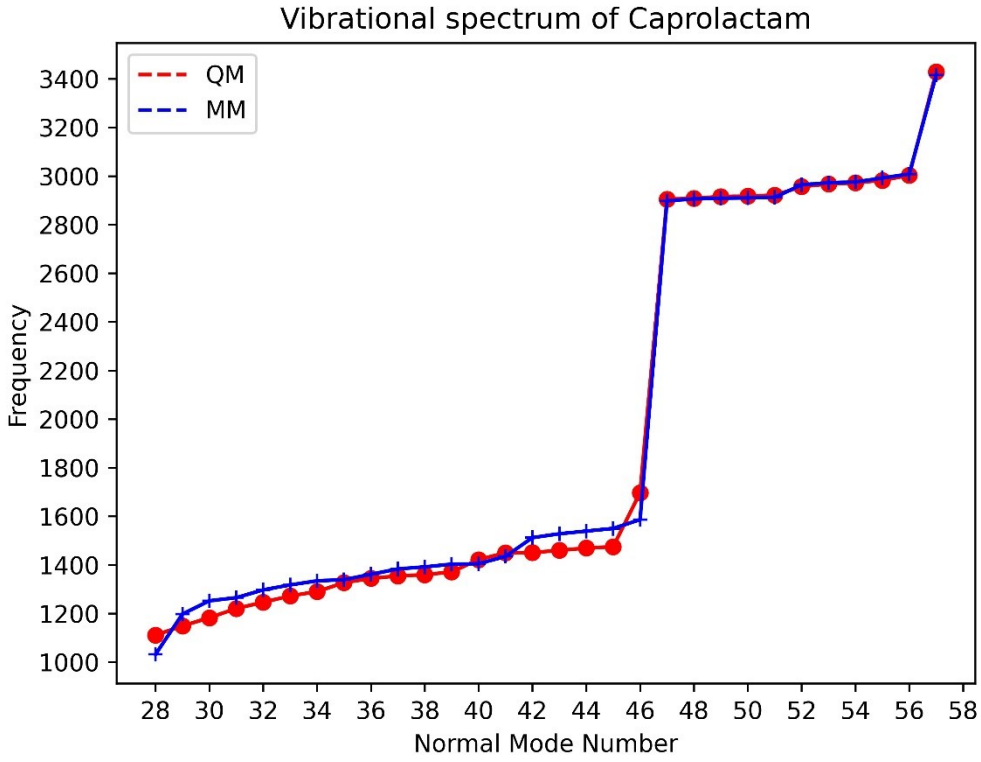

Vibrational spectrum of Naphthalene
QM
MM
Frequency
3000
2800
2600
2400
2200
2000
1800
1600
1400
1200
1000
31
33
35
37
39
41
43
45
47
49
51
53
Normal Mode Number

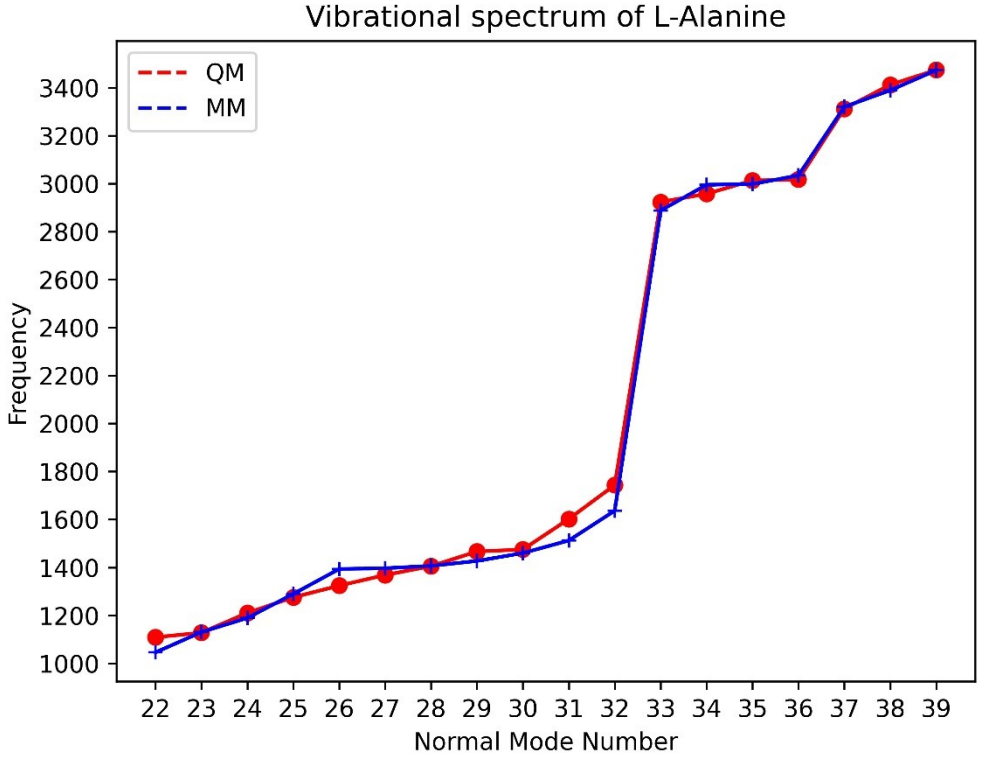

Vibrational spectrum of Acetone
QM
MM
Frequency
3000
2800
2600
2400
2200
2000
1800
1600
1400
1200
1000
15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30
Normal Mode Number


Vibrational spectrum of Adenine
QM
MM
Frequency
3600
3400
3200
3000
2800
2600
2400
2200
2000
1800
1600
1400
1200
1000
28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45
Normal Mode Number


Vibrational spectrum of Oxazole
QM
MM
Frequency
3200
3000
2800
2600
2400
2200
2000
1800
1600
1400
1200
1000
16
17
18
19
20
21
22
23
24
Normal Mode Number


Vibrational spectrum of Methyltriphosphate
QM
MM
Frequency
3000
2800
2600
2400
2200
2000
1800
1600
1400
1200
1000
38
39
40
41
42
43
44
45
46
47
48
49
50
51
Normal Mode Number


Vibrational spectrum of Caprolactam
QM
MM
Frequency
Normal Mode Number
3400
3200
3000
2800
2600
2400
2200
2000
1800
1600
1400
1200
1000
28 30 32 34 36 38 40 42 44 46 48 50 52 54 56 58


Vibrational spectrum of L-Alanine
QM
MM
Frequency
Normal Mode Number
3400
3200
3000
2800
2600
2400
2200
2000
1800
1600
1400
1200
1000
22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39