Machine Learning and Neural Network Models

The advent of machine learning (ML) or neural net (NN) technology offers the possibility of a new class of models that potentially afford both high reliability and quality as well as low (computational) cost. Implemented in the initial release of ***Spartan'24***, are a series of 5 neural nets designed to (i) improve conformer energy differences provided by MMFF molecular mechanics, (ii) accurately estimate ωB97X-D/6-31G* equilibrium geometries starting from MMFF geometries, and (iii) accurately estimate total energies from:

ωB97X-V/6-311+G(2df,2p)//ωB97X-D/6-31G*
ωB97M-V/6-311+G(2df,2p)//ωB97X-D/6-31G*
ωB97M(2)/6-311+G(2df,2p)//ωB97X-D/6-31G*

models starting from ωB97X-D/6-31G* equilibrium geometries and energies or estimated ωB97X-D/6-31G* equilibrium geometries and ωB97X-D/6-31G* energies. There are five neural nets in total, all of which are supported for closed shell uncharged molecules comprising the elements: H, C, N, O, F, S, Cl and Br (only).

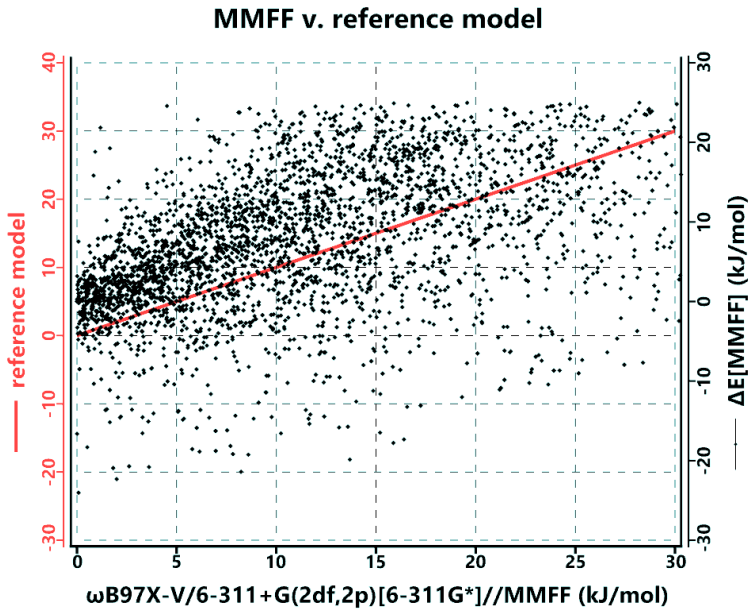
SpookyNet* has been used to generate all five neural nets. Data sets comprise ≈150k molecules with ≈1.8 million conformers obtained from ωB97X-V/6-311+G(2df,2p)[6-311G*]//MMFF calculations for the MMFF correction neural net, ≈120,000 molecules each with both equilibrium and 50 distorted geometries (times ≈6.1 million structures) obtained from ωB97X-D/6-31G* calculations for the estimated ωB97X-D/6-31G* geometry neural net, ≈295,000 molecules obtained from ωB97X-V/6-311+G(2df,2p)//ωB97X-D/6-31G* calculations for the estimated ωB97X-V/6-311+G(2df,2p) energy neural net, ≈289,000 molecules obtained from ωB97M-V/6-311+G(2df,2p)//ωB97X-D/6-31G* calculations for the estimated ωB97M-V/6-311+G(2df,2p) energy neural net, and the same ≈289,000 molecules obtained from ωB97M(2)/6-311+G(2df,2p)//ωB97X-D/6-31G* calculations for the estimated ωB97M(2)/6-311+G(2df,2p) energy neural net.

******Nat. Commun*. 12(1), **2021**, 1-14. (https://github.com/OUnke/SpookyNet).

Improving MMFF Molecular Mechanics

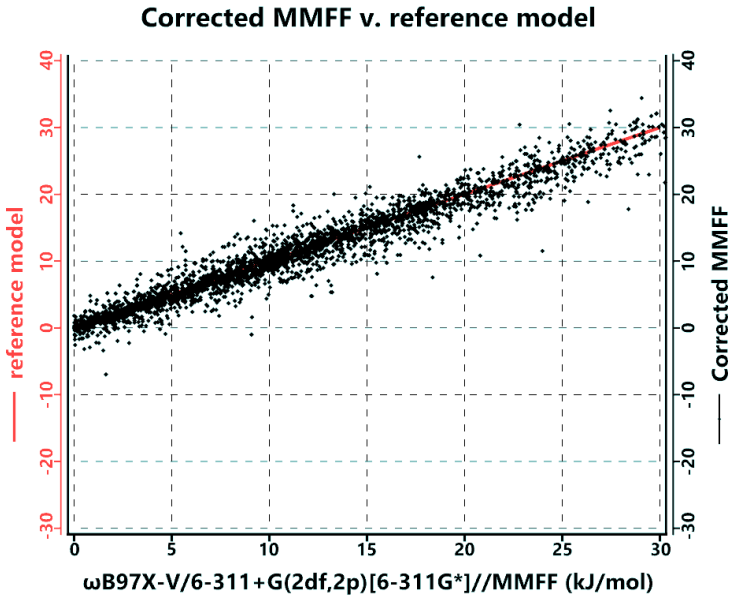
Molecular mechanics is perhaps the only class or methods that is routinely practical for calculations of conformer energy differences for flexible molecules with several degrees of conformational freedom. All accessible conformers, or at the very least a representative sample of these conformers, need to be identified and ranked based on energy. Assuming three-fold rotors, two and three degrees of freedom lead to only 9 and 27 conformers, respectively, whereas a molecule with 9 degrees of freedom (not at all unusual in natural products) leads to nearly 20,000 conformers. While only a relatively few conformers are likely to seriously affect calculated molecular properties, all must be examined to determine which few.

The MMFF94 approach has been shown to closely reproduce experimental conformational energy differences for small molecules with a single degree of conformational freedom. However, as suggested from the plot below, for a diverse series moderate size organic molecules with multiple degrees of conformational freedom, MMFF is not satisfactory. As experimental data are completely lacking, the reference energy is ωB97X-V/6-311+G(2df,2p)[6-311G*] density functional energy using MMFF equilibrium geometries, the same quantum chemical model used to train the neural net.



Comparison of Conformational Energy Differences Obtained from MMFF and ωB97X-V/6-311+G(2df,2p)[6-311G*] Calculations

MMFF structures corrected by neural net energies provides a far better accounting with only a modest (factor of two) increase in computer time. The overall RMS error reduces from ***10.2*** kJ/mol for MMFF to ***1.8*** kJ/mol with the neural net correction.



Comparison of Conformational Energy Differences Obtained from Neural Net and ωB97X-V/6-311+G(2df,2p)[6-311G*] Calculations

Estimating ωB97X-D/6-31G* Equilibrium Geometries

Two considerations generally dictate model selection, quality of the results and computation time. It is generally accepted that density functional models provide more consistent and better results than Hartree-Fock molecular orbital models but are also more time consuming, sometimes significantly so. The ωB97X-D/6-31G* density functional model is among the most commonly-used quantum chemical models for geometry calculation. It is both accurate and reliable, properly accounting for the geometries of molecules that are often poorly dealt with using Hartree-Fock models. While it is generally not sufficiently accurate for use in determining reaction energies, or energy differences among conformers, ωB97X-D/6-31G* structures provide a good foundation for energy calculations with more rigorous functionals and larger (more complete) basis sets.

The neural net fit to ωB97X-D/6-31G* incorporated into ***Spartan'24*** is many orders of magnitude faster than the underlying density functional model. Structures that

previously required hours of computer time may now be closely approximated in seconds (at most a few minutes). It must be emphasized each one of these neural network calculations is unique and specific to the underlying computational model it is trained to reproduce (a particular property (or properties) from a specific functional and with a specific basis set. A neural net constructed to reproduce equilibrium geometries from the popular B3LYP-D3/6-31G* model will be distinct from that directed at ωB97X-D/6-31G*. Also acknowledged: different training sets and different convergence criteria (as well as different starting geometries) will ultimately lead to different neural network definitions. The hope is that subtle differences in training sets will lead to nearly the same final (or very similar) equilibrium geometries.

We've applied three criteria to assess how well the geometry neural network performs, using a test set comprising 430 natural products with molecular weights ranging from ≈ 200 to 700 amu (mean of 447 amu), none of which were used in neural net training or validation.

The most obvious criterion is structure (itself): bond lengths, bond angles, and dihedral angles [excluding those involving hydrogen(s)]. The performance was substantive, with only 14 molecules showing bond lengths that deviate from ωB97X-D/6-31G* values by more than 0.02 Å. Similar results were observed in bond angles; only 4 and 7 molecules incorporate a bond angle that differs by more than 10^{o} and 5^{o}, respectively. Less than a quarter of the molecules show individual bond-angle deviations of more than 2^{o}.

Dihedral angles are perhaps the most important structural parameter in that large deviations may result in molecular shapes that differ greatly from those obtained using ωB97X-D/6-31G* (the training model). One concern is that further calculation of NMR chemical shifts, which are sensitive to medium to long-range interactions, will be impacted. Only two molecules show very large (>50^{o}) deviations; 34 (8%) show deviations >20; 30% of molecules show deviations of >10^{o}.

The second criterion involves differences in energies resulting from use of neural net equilibrium geometries in lieu of ωB97X-D/6-31G* geometries. The RMS error for the neural net is 8 kJ/mol and the energies of only 11 molecules deviate from ωB97X-D/6-31G* by more than 20 kJ/mol, the largest of which is 42 kJ/mol. 54 molecules deviate by more than 10 kJ/mol.

Perhaps the most important metric is the effect of equilibrium geometry on proton and especially ^{13}C NMR chemical shifts. This is because geometry calculation is key

in any effort to realize an accurate rapid procedure to calculate chemical shifts of conformationally flexible organic molecules, natural products among them, and to distinguish among stereoisomers that closely reproduces the results of the accurate (but time consuming) multi-step protocols now in place*.

Only 4 and 29 molecules exhibit individual deviations greater than 1 and 0.5 ppm, respectively, more than a third of the molecules have one or more protons with chemical shifts that deviate from ωB97X-D/6-31G* values by more than 0.2 ppm.

Statistics for ^{13}C chemical shifts mimic those for proton shifts. Only one molecule exhibits an individual deviation >10 ppm and only 11 exhibit individual deviations >5 ppm. Roughly a third of the molecules based on neural net geometries show at least one individual deviation >2 ppm. The overall RMS (all carbons) and the RMS of the maximum individual ^{13}C shift errors (per molecule) from the neural net geometries are 0.8 ppm and 2.3 ppm, respectively.

Improved Reaction and Conformer Energy Calculations

Reaction energies and conformer energies differences are among the most important quantities obtained from quantum chemical calculations. Organic chemists, natural products chemists in particular, will benefit from calculations able to identify the dominant stereoisomer arising from a chemical reaction, and from determination of conformer energy differences, providing Boltzmann weights and with these, accurate prediction of NMR shifts of flexible molecules.

The ”gold standard” CCSD(T)/CBS calculations scale formally as the seventh power of the number of basis functions, and can take many hours (even days) to complete. Reducing computational cost while achieving CCSD(T) accuracy has been the primary motivation for development of many density functional models. In previous versions of ***Spartan***, the primary focus has been on the Head-Gordon functionals: ωB97X-V, ωB97M-V and ωB97M(2), all with the 6-311+G(2df,2p) basis set and all starting from ωB97X-D/6-31G* geometries. The CCSD(T) procedure is superior to these three functionals, and while the 6-311+G(2df,2p) basis set is moderately large and diverse, the recipe developed to yield a “complete basis set” (CBS) ensures that further additions will at most have modest effect on the energy. Finally, the geometry is different from that used in CCSD(T) calculations (MP2/aug-cc-pVDZ basis set vs. ωB97X-D/6-31G*).

J. Nat. Prod.* **2019, 82 8, 2299-2306. (https://pubs.acs.org/doi/10.1021/acs.jnatprod.9b00603)

The two models ωB97X-V/6-311+G(2df,2p)//ωB97X-D/6-31G* and (to a lesser extent) ωB97M-V/6-311+G(2df,2p)//ωB97X-D/6-31G* are practical for routine calculations on rigid molecules (or those with only a few degrees of conformational freedom) up to molecular weights <400 amu, although they may require tens of minutes to several hours to complete. They are clearly impractical, at least routinely so, for significantly larger systems, or for molecules with multiple degrees of conformational freedom. The ωB97M(2)/6-311+G(2df,2p)//ωB97X-D/6-31G* is more computationally costly and practical ranges are even smaller.

Spartan'24 provides neural networks trained to these three density functional models. These require an ωB97X-D/6-31G* geometry (or an estimated geometry from the Est. ωB97X-D/6-31G* neural network along with a single point ωB97X-D/6-31G* energy). Because of these requirements, application of these neural nets is not "instantaneous" (unlike the two previous described machine learning models, these three are not devoid of quantum chemical calculations as input). This said, the overall performance is easily an *order of magnitude faster* utilizing these neural nets to provide energies (compared to the pure quantum chemical calculations). As quality conformer energy differences are needed to establish accurate **Conformer Distributions (**as well as Boltzmann weighted NMR shifts), these energy-focused neural networks provide significant computational savings in tasks including conformational searching, in particular the multi-step **NMR Spectrum** task.