Contribution ID: **9** Type: **Oral talk**

Revealing inaccuracies of the GFN1-xTB semi-empirical method with machine-learning toolset

Wednesday, 17 January 2024 10:45 (20 minutes)

The change in conformation during biological functioning is a characteristic feature of biomolecules. The presence of various conformers complicates their physics-based modeling, as the properties of each conformer need to be determined separately. Moreover, the relative energies of the conformers become an additional quantity of interest because they determine the probability of each conformation in a real system.

Although ab initio quantum-mechanical methods are the most accurate for determining the physical properties of molecules, their computational costs become prohibitive when applied to biomolecules with a large number of atoms and/or a large number of conformations. Semi-empirical quantum mechanical methods, particularly those based on the tight-binding approximation in density functional theory (DFTB), significantly reduce computational requirements by utilizing parameterized models for the elements of the Fock or Kohn-Sham matrices. However, due to such simplifications, the resulting method can show low accuracy in some cases, despite performing accurately in other cases.

In this work, we demonstrate that machine learning (ML) methods such as graph neural networks and ensemble models can be employed to predict errors in determining the relative energies of conformers by the semi-empirical method GFN1-xTB from the DFTB family, based solely on the molecular structural formula as input.

The developed ML ensemble model achieves an accuracy of about 80% in determining whether a given molecule is 'challenging' for GFN1-xTB. This approach can be used to find the molecules that most vividly highlight the shortcomings of the physical model underlying the GFN1-xTB method. We demonstrate that for the 190 molecules selected using the developed model, the average error in relative energies of their conformers obtained by the GFN1-xTB method is 4.2 kcal/mol, in contrast to 1.8 kcal/mol when the same number of test molecules are selected randomly. This indicates that the ML toolset indeed allows for the identification of challenging molecules and can therefore be useful in improving the approximations utilized by semi-empirical methods.

Primary authors: TERETS, Andrii (1) Faculty of Physics, Taras Shevchenko National University of Kyiv. 2) Chuiko Institute of Surface Chemistry, NAS of Ukraine,); Dr NIKOLAIENKO, Tymofii (Faculty of Physics, Taras Shevchenko National University of Kyiv)

Presenter: TERETS, Andrii (1) Faculty of Physics, Taras Shevchenko National University of Kyiv. 2) Chuiko Institute of Surface Chemistry, NAS of Ukraine,)

Session Classification: Morning Session 2

Track Classification: Physics of Biological Macromolecules