A LARGE-SCALE DATASET OF QUANTUM CHEMICAL PROPERTIES OF DRUG-LIKE MOLECULES FOR Δ-LEARNING MODELS

prepared by Frolov Dmitry

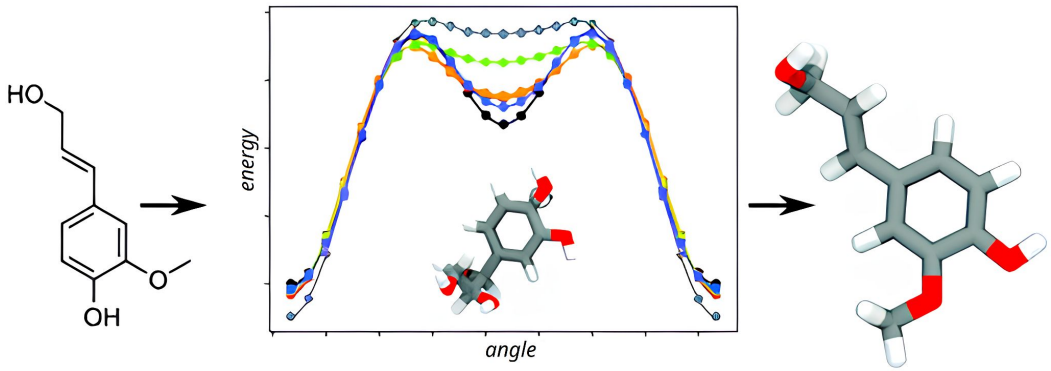
tutored by Sedov Igor

Motivation

Quantum chemistry methods can compute any properties of atomic systems and are widely used in molecular modeling and drug design.

Applications:

- Estimation of reaction directions and mechanisms [1];
- Force field parameterization [2];
- Descriptor calculation for QSAR studies [3].



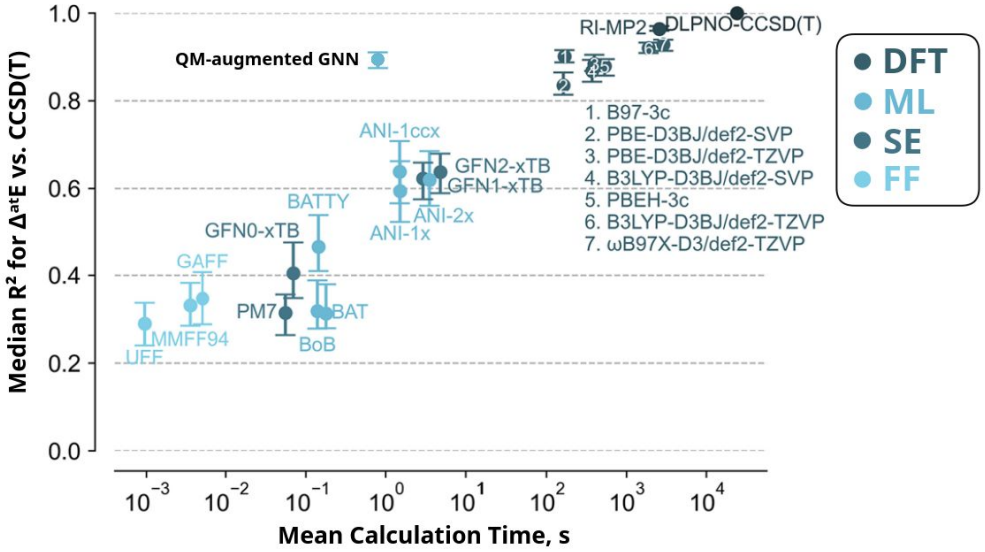
1. Hayashi H, et al. Quantum chemical calculations for reaction prediction in the development of synthetic methodologies. Chemical Science. 2023;14(42):11601-16.

2. Wang J, et al. Development and testing of a general amber force field. Journal of computational chemistry. 2004 Jul 15;25(9):1157-74.

3. Karelson M, et al. Quantum-chemical descriptors in QSAR/QSPR studies. Chemical reviews. 1996 May 9;96(3):1027-44.

Motivation

The computational cost of ***QM methods*** increases exponentially with the linear growth of accuracy [1]. The Δ-learning approach was proposed to refine ***semiempirical methods*** to the accuracy of ***DFT*** (***density functional theory***) methods [2].



1. Folmsbee D, Hutchison G. Assessing conformer energies using electronic structure and machine learning methods. International Journal of Quantum Chemistry. 2021 Jan 5;121(1):e26381.

2. Ramakrishnan R, et al. Big data meets quantum chemistry approximations: the Δ-machine learning approach. Journal of chemical theory and computation. 2015 May 12;11(5):2087-96.

Existing Quantum Chemical Datasets

Dataset	Unique Structures	Max. Heavy Atoms	Atom Types	SE Method	DFT Method	Geometry	Wave Function	Thermo chemistry
QM9 [1]	134k	9	H, C, N, O, F	−	B3LYP/6-31G(2df,p)	DFT	−	+
QMugs [2]	666k	100	H, C, N, O, F, P, S, Cl, Br, I	GFN2-xTB	ωB97X-D/def2-SVP	SE	+	−
PubChemQC[a] [3]	86M	80	H, C, N, O, F, Na, Mg, Si, P, S, Cl, K, Ca, Br, I, etc.	PM6	B3LYP/6-31G*	SE	+	−
∇DFT [4]	1M	27	H, C, N, O, F, Cl, Br	−	ωB97X-D/def2-SVP	MM	+	−

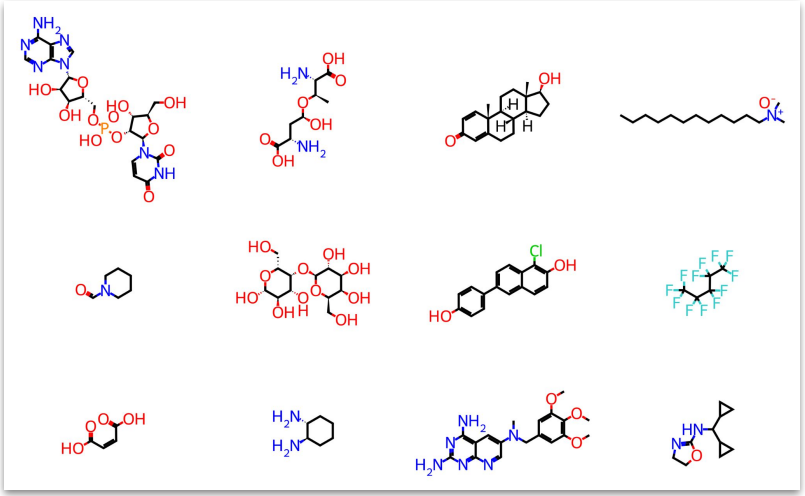
[a] B3LYP/6-31G*//PM6 variant

1. Ramakrishnan R, et al. Quantum chemistry structures and properties of 134 kilo molecules. Scientific data. 2014 Aug 5;1(1):1-7.
2. Isert C, et al. QMugs, quantum mechanical properties of drug-like molecules. Scientific Data. 2022 Jun 7;9(1):273.
3. Nakata M, et al. PubChemQC B3LYP/6-31G*//PM6 data set: The electronic structures of 86 million molecules using B3LYP/6-31G* calculations. Journal of Chemical Information and Modeling. 2023 Sep 7;63(18):5734-54.
4. Khrabrov K, et al. nabladft: Large-scale conformational energy and hamiltonian prediction benchmark and dataset. Physical Chemistry Chemical Physics. 2022;24(42):25853-63.

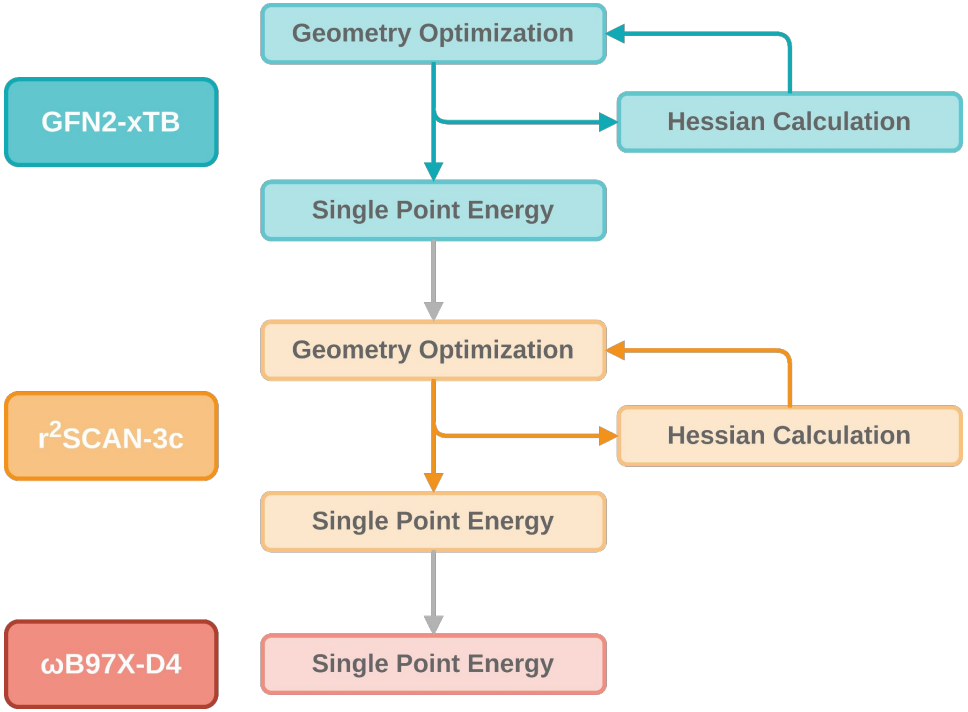
DrugBank Database

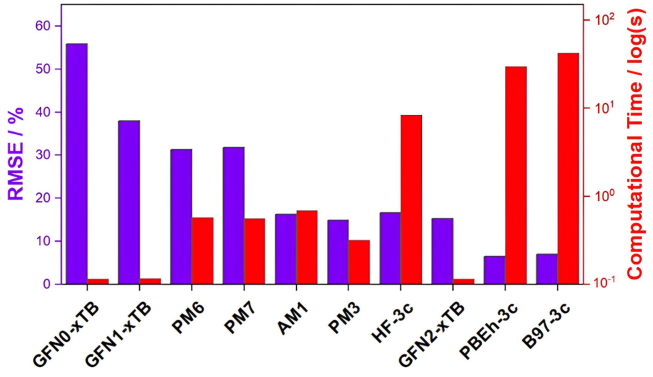
DrugBank is a database containing FDA-approved, experimental drugs, and other compounds of pharmaceutical interest. At the time of download, it contained 11,928 structures.

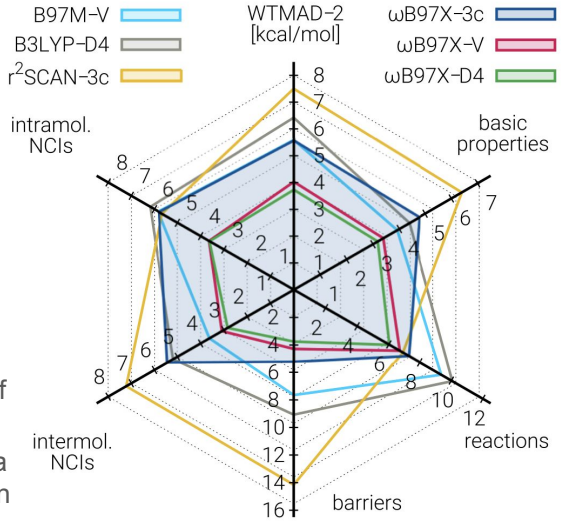
Instances were carefully prepared. 9,993 molecules (84%) remained after standardization, deduplication, and removal of irrelevant structures.



Choice of QM Methods







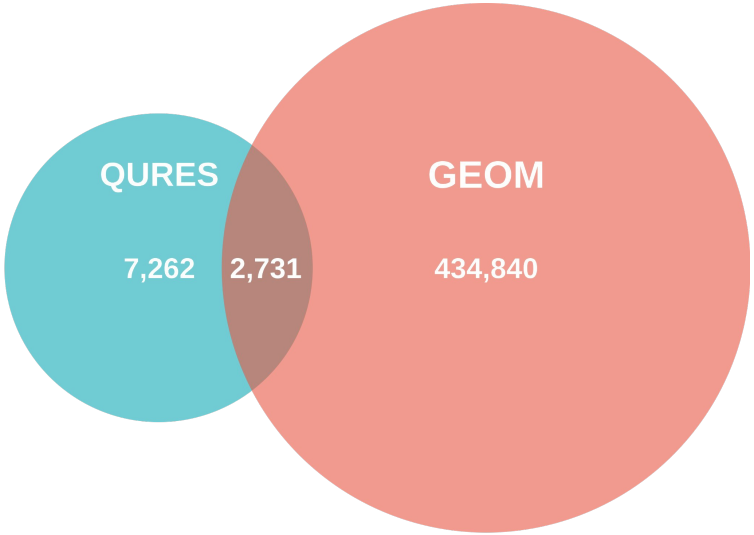
1. Soyemi A, Szilvási T. Benchmarking semiempirical qm methods for calculating the dipole moment of organic molecules. The Journal of Physical Chemistry A. 2022 Mar 15;126(11):1905-21.

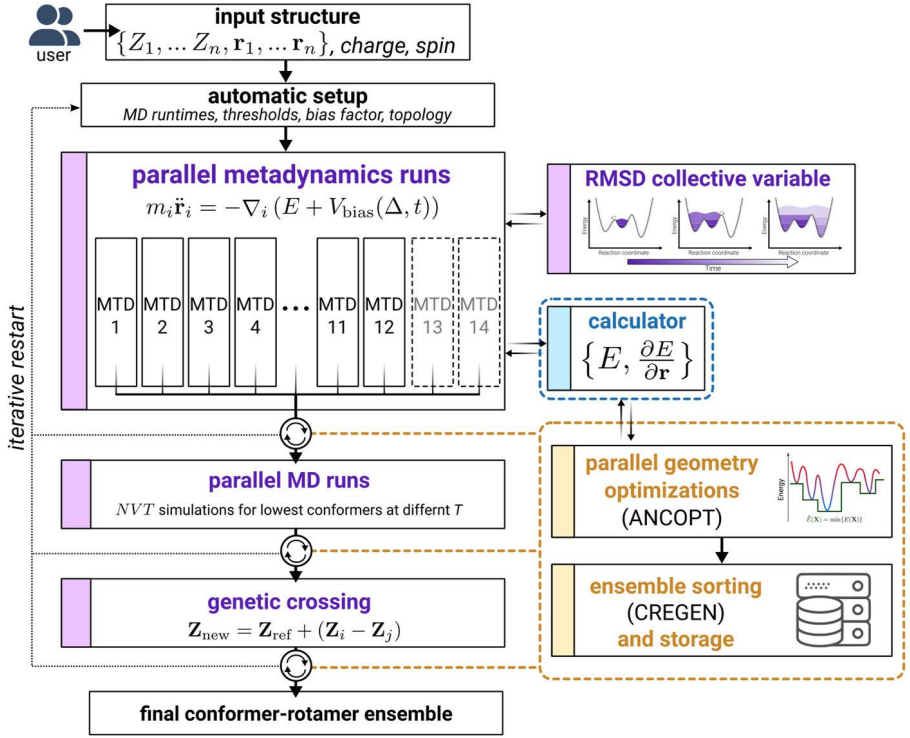
2. Müller M, et al. ωB97X-3c: A composite range-separated hybrid DFT method with a molecule-optimized polarized valence double-ζ basis set. The Journal of Chemical Physics. 2023 Jan 7;158(1).

Conformational Search

CREST – *Conformer–Rotamer Ensemble Sampling Tool*

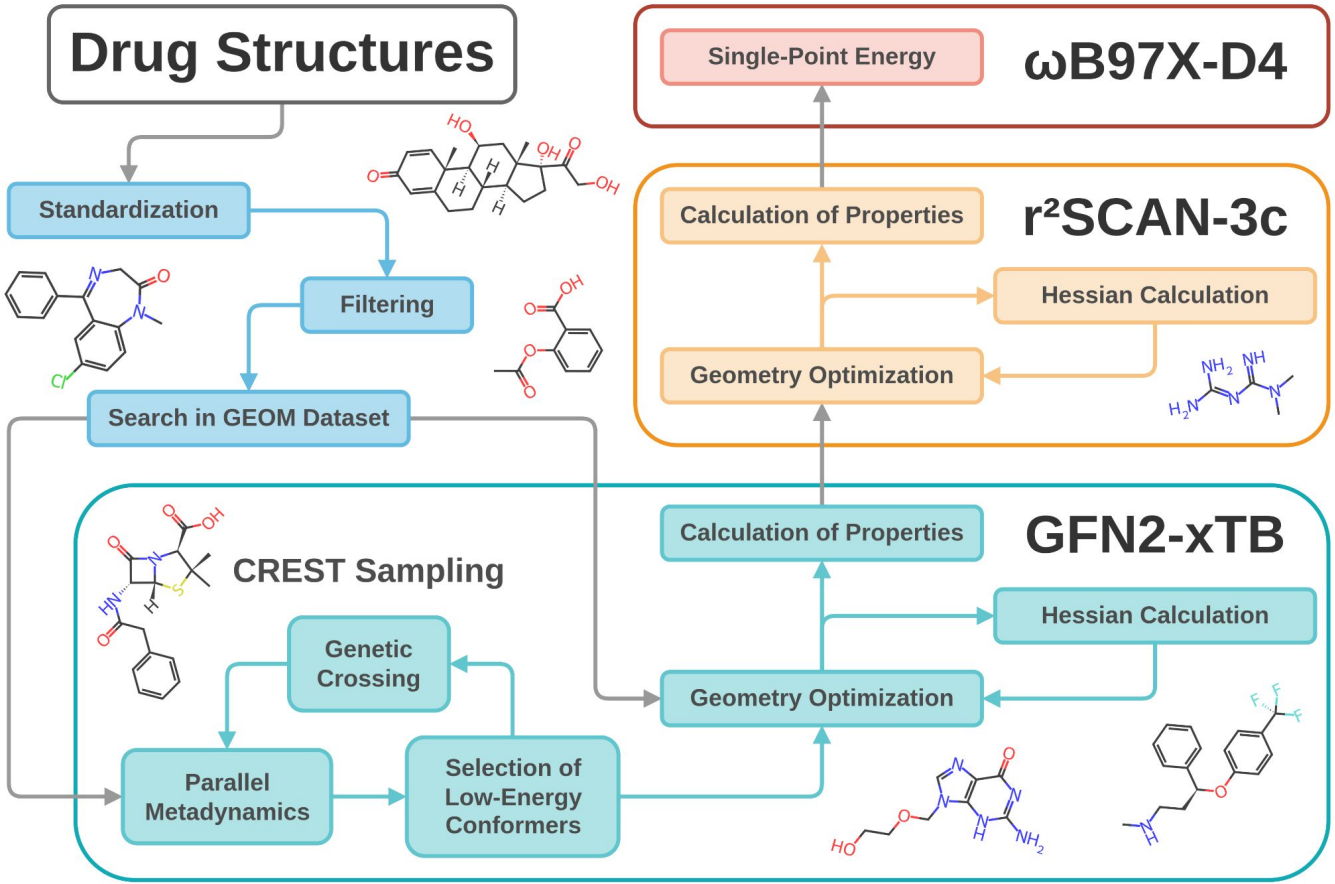
GEOM – *Geometric Ensemble Of Molecules*



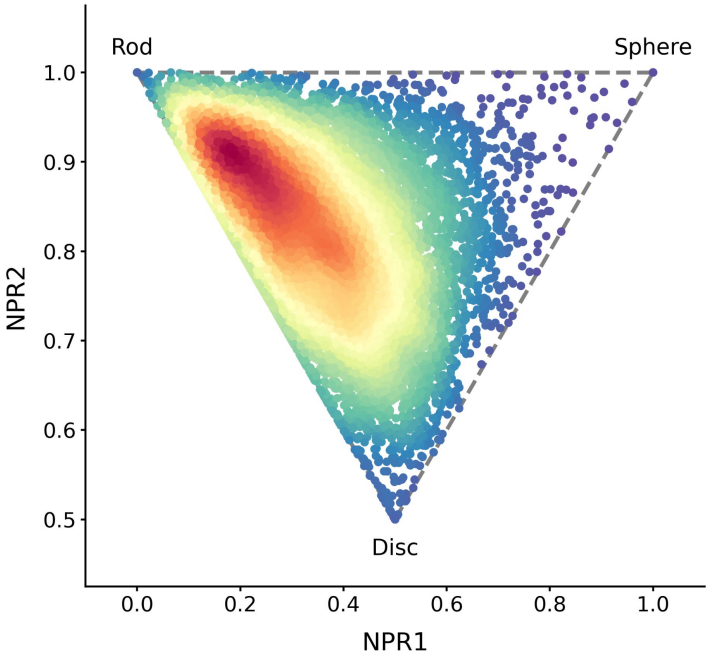


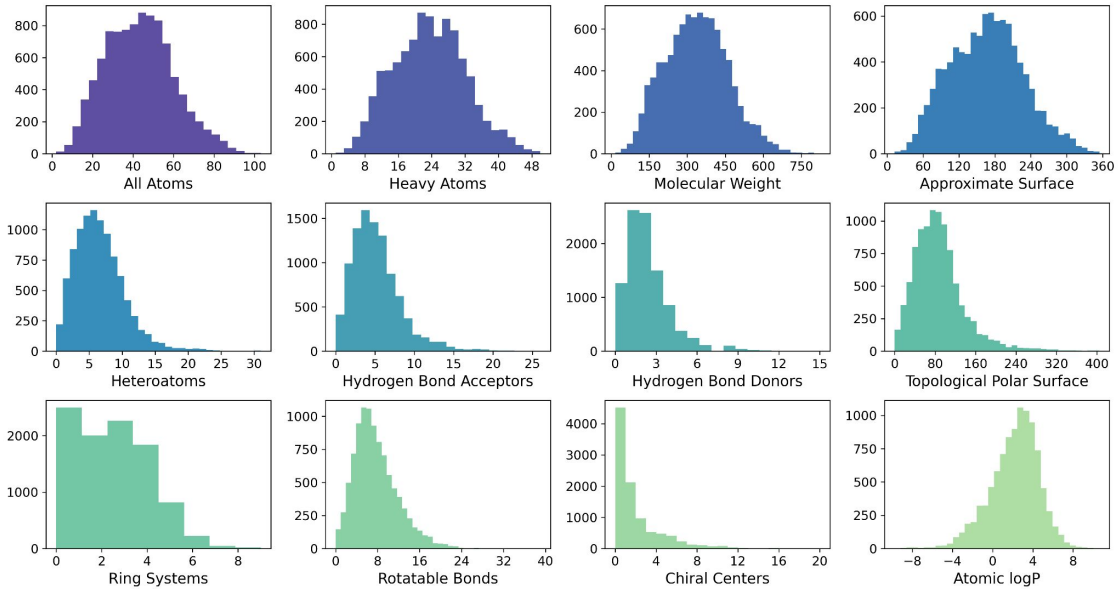
9,916 (83%) molecules successfully passed conformational search and DFT optimization.

Complete Workflow Overview



Diversity of Molecules in QURES Dataset





Comparison of QURES with Other Datasets

Dataset	Unique Structures	Max. Heavy Atoms	Atom Types	SE Method	DFT Method	Geometry	Wave Function	Thermo chemistry
QURES	9.9k	50	H, C, N, O, F, P, S, Cl, Br	GFN2-xTB	r^2SCAN-3c & ωB97X-D4/def2-TZVP	DFT	+	+
QM9 [1]	134k	9	H, C, N, O, F	−	B3LYP/6-31G(2df,p)	DFT	−	+
QMugs [2]	666k	100	H, C, N, O, F, P, S, Cl, Br, I	GFN2-xTB	ωB97X-D/def2-SVP	SE	+	−
PubChemQC [3]	86M	80	H, C, N, O, F, Na, Mg, Si, P, S, Cl, K, Ca, Br, I, etc.	PM6	B3LYP/6-31G*	SE	+	−
∇DFT [4]	1M	27	H, C, N, O, F, Cl, Br	−	ωB97X-D/def2-SVP	MM	+	−

1. Ramakrishnan R, et al. Quantum chemistry structures and properties of 134 kilo molecules. Scientific data. 2014 Aug 5;1(1):1-7.

2. Isert C, et al. QMugs, quantum mechanical properties of drug-like molecules. Scientific Data. 2022 Jun 7;9(1):273.

3. Nakata M, et al. PubChemQC B3LYP/6-31G*//PM6 data set: The electronic structures of 86 million molecules using B3LYP/6-31G* calculations. Journal of Chemical Information and Modeling. 2023 Sep 7;63(18):5734-54.

4. Khrabrov K, et al. nabladft: Large-scale conformational energy and hamiltonian prediction benchmark and dataset. Physical Chemistry Chemical Physics. 2022;24(42):25853-63.