# Advancing Molecular Conformation Generation

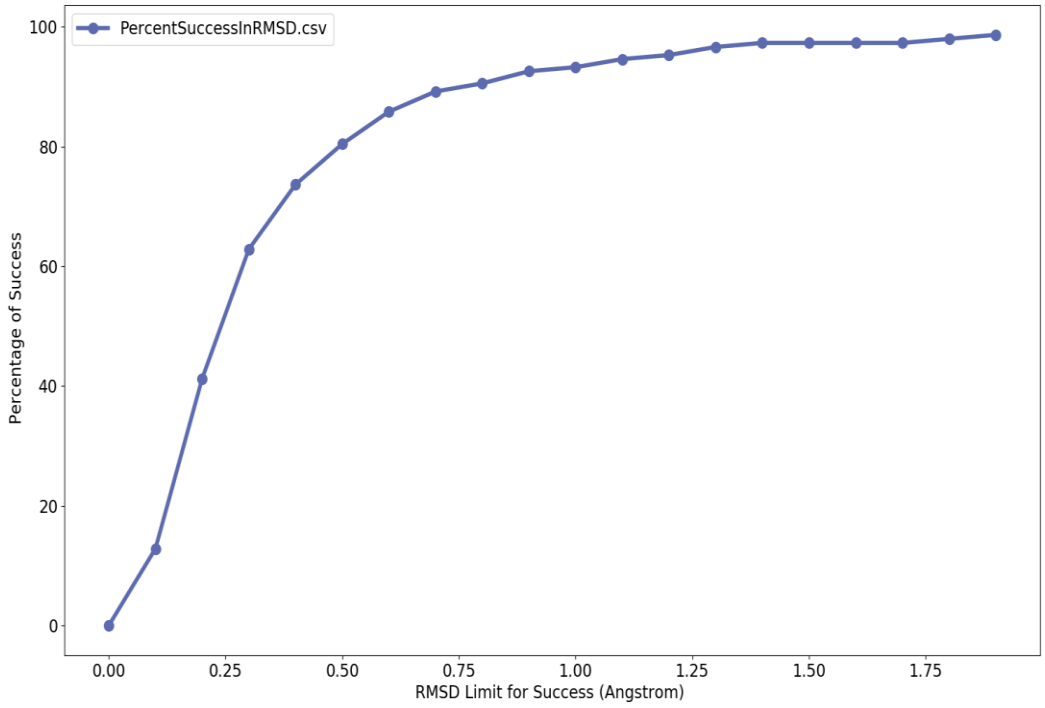
(Robustness with *ab initio* DFT-D4 Ranking, Ease of Use on Inexpensive Azure Spot Instances)

In the ever-evolving landscape of molecular research, the precision and efficiency of conformation generation play a pivotal role. At QuantumFuture, we are proud to introduce our groundbreaking strides with a robust and cost-effective solution for molecular conformation generation. Leveraging the power of the right combination of advanced sampling, QM semiempirical geometry optimizations and extremely efficient accurate *ab initio* DFT-D4 ranking on inexpensive Azure spot instances, our approach opens new avenues for advancements in drug discovery and molecular design.

To validate the robustness of our conformation generation code, we conducted an extensive benchmarking exercise utilizing a representative test set comprising 150 FDA-approved drugs. We randomly selected 25 FDA-approved drugs from the Crystallography Open Database (COD) for all six categories defined by the number of rotatable bonds (2, 3, 4, 5, 6, 7). Notably, torsions of methyl groups were excluded in the counts of the rotatable bonds. Employing our innovative qfconfsearchDFT program our analysis delved into the frequency of obtaining conformations close to experimental structures, the proximity of our computed conformations to the experimental counterparts, their rankings among generated conformations, and the strain energies. The proximity of molecular conformations to vacuum conformations, measured by the RMSD of non-hydrogen (heavy) atoms, provided a robust benchmark against the highest quality small molecule crystal structures.

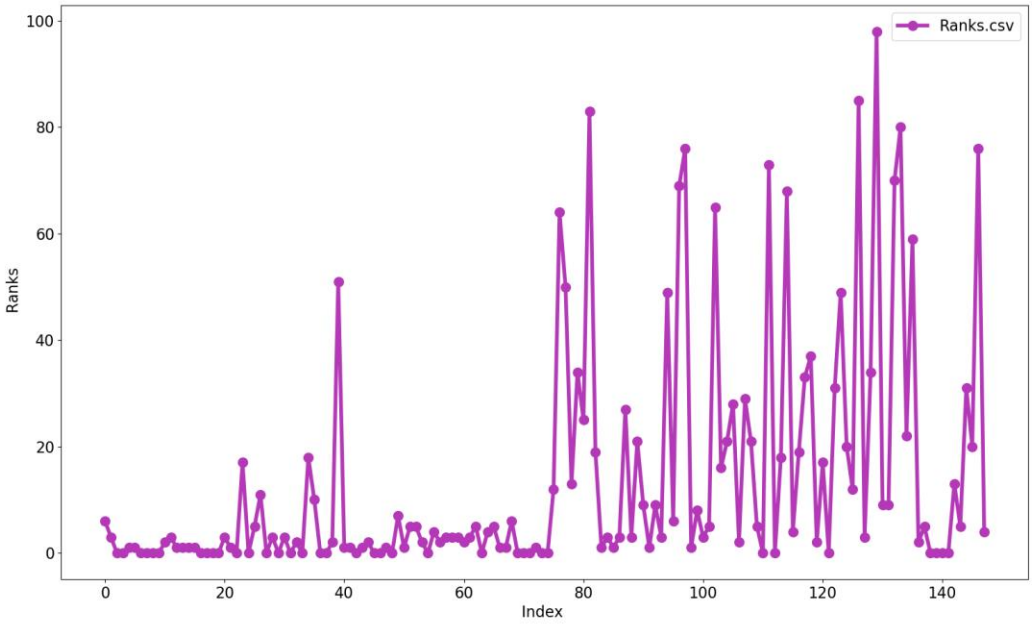
Central to our assessment is the definition of success through heavy atom RMSD values. Success is gauged by the ratio of cases where our found conformations are closer to the experimental structure than a predefined RMSD success limit and then this systematic evaluation allows us to measure the success rate as the function of the RMSD success limits, providing a nuanced understanding of the robustness of our methodology.

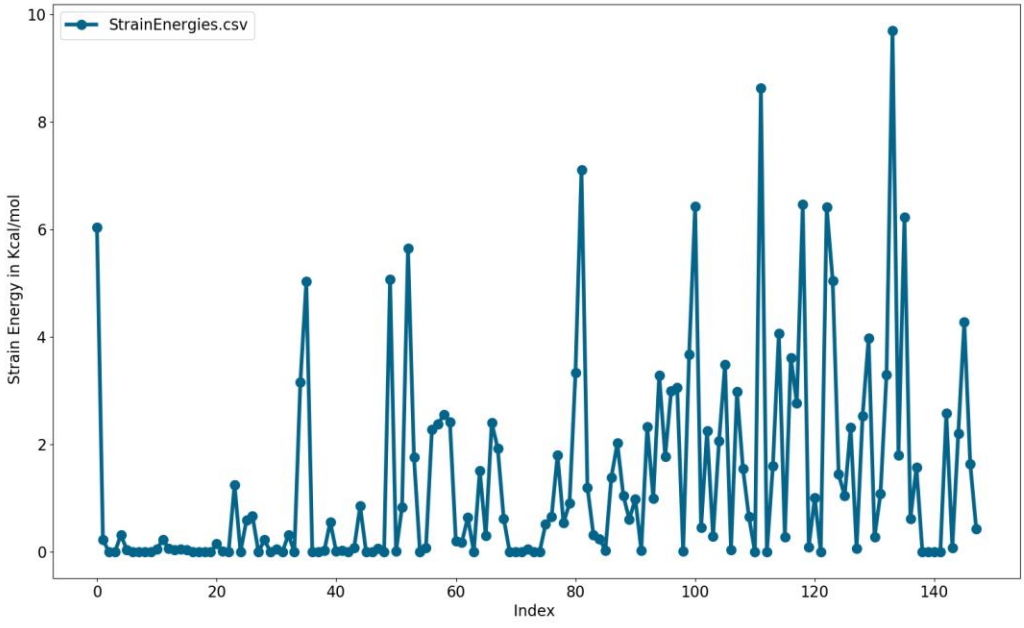
The success rates in percentage are shown in the plot below:




The results show that over 62% of drug molecules exhibit heavy atom RMSD below 0.3 Angstrom, surpassing 80% below 0.5 Angstrom, and an impressive 93% below 1 Angstrom. These figures bear significance when contrasted with traditional cheminformatics and force field-based approaches, often benchmarked the success rates with 1.5 or even at 2 Angstrom RMSD. The pivotal importance of achieving sub 0.5 Angstrom RMSD lies in ensuring accuracy in subsequent predictions, be it in docking, molecular dynamics simulations, or organic crystal structure predictions. Larger difference than 0.5 Angstrom in RMSD almost certainly guarantees to have different heavy atom conformation than the experimental one indicating not as robust scheme in conformation generations as we would like to have which makes the conformation based subsequent projects likely much less accurate.

The DFT-D4 (rev-TPSS, 6-311++G**, QF optimized D4 parameters) energy rankings and the strain energies are shown below.


Ranks.csv
Ranks
Index
0
20
40
60
80
100
120
140


StrainEnergies.csv
Strain Energy in Kcal/mol
Index
0
2
4
6
8
10
20
40
60
80
100
120
140

The index starts with the first 25 drug molecules using 2 rotatable bonds and takes the next 25 batches with 3 rotatable bonds and so on until finishes with the last drug molecule of the batch of 7 rotatable bonds. Unfortunately, there were two drug molecules where the freely available tools were not able to produce valid sdf file from the original cif file and we just dropped those two cases reducing our test set to 148 FDA approved drug structures. Both have 4 rotatable bonds and therefore this category has only 23 instead of 25 molecules.

| Number of Rotational Bonds | Average RMSD (Angstrom) | Average Rank | Average Strain Energy (Kcal/mol) |
|---|---|---|---|
| 2 | 0.229415 | 1.64 | 0.341 |
| 3 | 0.205316 | 4.72 | 0.667 |
| 4 | 0.327559 | 2.43 | 1.122 |
| 5 | 0.323427 | 23.24 | 1.487 |
| 6 | 0.590802 | 20.24 | 2.353 |
| 7 | 0.475876 | 28.24 | 2.089 |

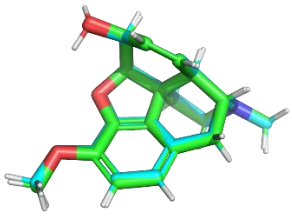
Both plots and the table above show that the ranking and the strain energies are usually very low and there are numerous cases when they are exactly zero which means that the lowest energy vacuum conformation is the closest one to the experimental structures. After about 3-4 rotatable bonds the situation gets a bit more complex, and both the rankings and the strain energies are usually larger and more volatile. We found it amazing that even for the drugs with 7 rotatable bonds the lowest energy vacuum conformations are the closest to the experimental structure for numerous examples. Since we know that accurate *ab initio* ranking of the conformations are very important, we have repeated the entire project by using our most accurate revSCAN functional with def2-TZVP basis set.

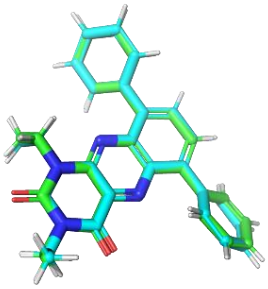
There is no strict rule about that how low the ranking of the vacuum conformations should be with the matching experimental structure but usually lowering the rankings indicates the increase of accuracy at least on average when enough structures are considered. This is exactly what we found here as well. The table below shows the results with revSCAN functional, def2-TZVP basis set and with our new D4 optimized VDW parameters. Perhaps the largest difference is that the average ranking went down from 28 to 22 for the most flexible category having 7 rotatable bonds.

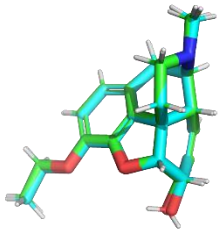
| Number of Rotational Bonds | Average RMSD (Angstrom) | Average Rank | Average Strain Energy (Kcal/mol) |
|---|---|---|---|
| 2 | 0.229415 | 1.40 | 0.321 |
| 3 | 0.205316 | 4.12 | 0.624 |
| 4 | 0.327559 | 2.22 | 1.222 |
| 5 | 0.323427 | 19.96 | 1.581 |
| 6 | 0.590802 | 19.36 | 2.299 |
| 7 | 0.475876 | 22.00 | 1.891 |

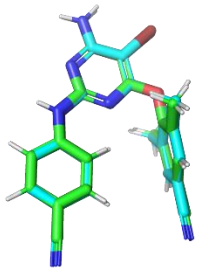
Note that the average ranks are even better than it shows in the tables above for larger molecules. The qfconsearchDFT.x application utilizes a new, experimental, and very fast alternative to the geometric RMSD deduplications when the molecule has over 40 atoms and the current parameters are very conservative and therefore sometimes keeps some conformations that are almost identical.

A few pictures of amazing overlaps of the experimental and theoretical conformations are show below (just for fun purposes) for Codeine, Ethylmorphine, 1,3-diethyl-6,9-diphenylalloxazine and Etravirine achieving very low 0.0706947, 0.0644524, 0.124212 and 0.118875 Angstrom heavy atom RMSDs.

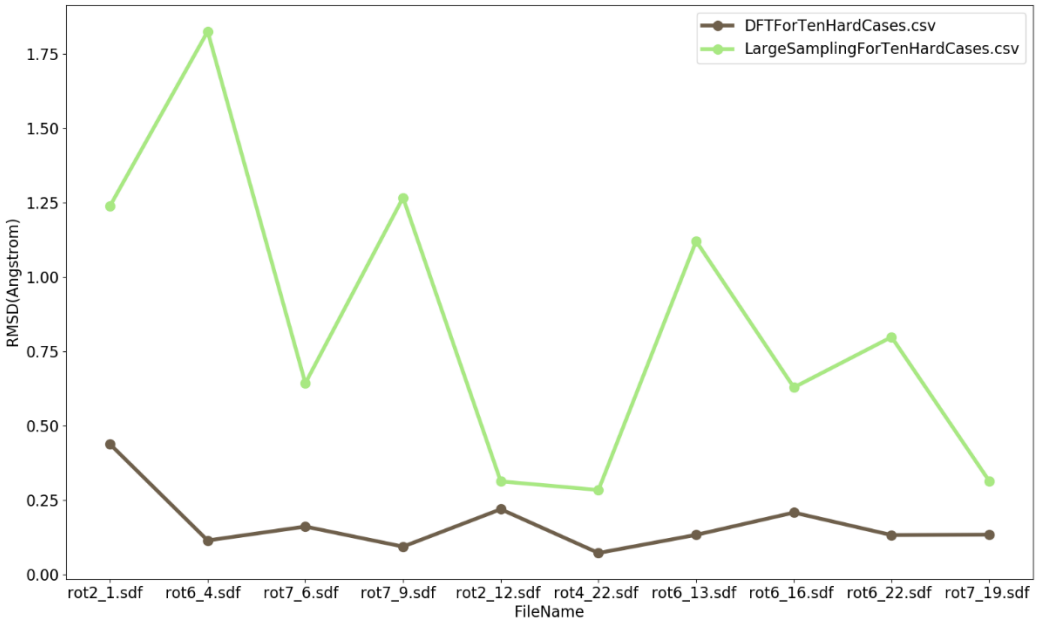








Acknowledging the pursuit of perfection, we recognize room for improvement in our scheme. Our initial results were achieved using default parameters, with only 10 structures exhibiting RMSD larger than 1.0 Angstrom. Rigorous efforts were invested in increasing sampling size and increasing the number of conformations for reranking conformers with DFT-D4, resulting in some further improvements and bringing the RMSD below 1 Angstrom for 6 out of the 10 problematic cases. We have also made DFT-D4 geometry optimizations for those 10 examples starting the local optimizations from the experimental structures to find out how far one of the nearest DFT-D4 vacuum local minimum geometry is located from the experimental structures. Both sets of results are shown on the plot below.



The plot underscores the remarkable closeness of *ab initio* DFT-D4 vacuum-optimized geometries to experimental structures. While our current scheme balances semi-empirical QM geometries and DFT-D4 re-ranking, the potential for even more accurate geometries and relative energies is apparent through full DFT-D4 geometry optimizations. This solution is a bit too costly on one workstation and therefore it is not available in one shot automatically in our current qfconfsearchDFT application (it can be done with the combination of qfdft and qfconfsearchDFT), however we are putting together a new automatic solution for this purpose on azure soon.

This leads to our final topic about the ease of use of our applications on Linux based workstations and on azure. On any supported Linux based workstation one can perform such calculations as easily as

```
qfconfsearchDFT.x --Input myInputMol.sdf
```

or

```
qfconfsearchDFT.x --Input myInputMol.smi
```

where all application parameters are used with their default values. This command can be obviously inserted in a loop of a shell or python script and perform conformation searches for unlimited molecules one after another. The application parameters can be modified by using simple command line options. For instance, requesting to use the very recent modern R2SCAN functional with the def2-TZVP basis set can be achieved like

```
qfconfsearchDFT.x --Input myInputMol.sdf --BasisSet def2-TZVP --Functional R2SCAN
```

We have also developed a new application to be able to perform qfdft, qfconfsearchDFT and qfLowerLevel calculations on a large scale on azure utilizing cheap spot instances with up to 90% discounts compared to on demand prices. The usage is simple here as well, for example:

```
qfazurelaunch.x --azureRegion westus --azureInstanceType Standard_F16s_v2 --azureMaxSpotInstances 50 --azureProjectName MyProject qfconfsearchDFT.x --BasisSet def2-TZVP --Functional R2SCAN
```

Where the only requirements are that all molecular input files must be tarred in the MyProject.tar.gz and a valid QuantumFuture.lic license file needs to exist in the directory. That's it! Everything is taken care of automatically and all the results can be downloaded with another simple command. This application will automatically install all necessary QF applications and libraries, creates a temporary encrypted blob storage on azure, upload the input files to the encrypted storage space, starts 50 computational nodes with 16 vCPUs each in azure batch, perform all the calculations and automatically scales down the number of nodes as the project progresses and eventually to zero at the conclusion of the project. In order to have maximum privacy and security all users utilize their own secure azure account(s) so our company does NOT provide any service during the calculations except the applications themselves, the license file and customer support if needed. We use exclusively inexpensive spot instances with great discounts to make the calculations as affordable as possible. One of the catches by using spot instances is that, based on demand changes in capacity, we can lose the nodes in the middle of calculations. For this reason, we have developed the technology to be able to automatically continue the calculations without much repetition when we get back the node(s) and we can continue the calculations the same inexpensive way as before. This technology, in combination with our extremely fast DFT program (see

https://bettermolecularmodelling.com/qffileexchange/QF23Beta/QFDFT22BetaFlyer.pdf

and

https://bettermolecularmodelling.com/qffileexchange/QF23Beta/CheaperThanFreeQFDFT.mp4
)

provides robust QM based conformations searches with *ab initio* DFT based energy rankings in a very uniquely efficient and affordable manner.

Merry Computing from QuantumFuture!