Post-mortem notes

I'll summarize here a couple of the issues that I struggled with. I tried to find solutions to these, first by reading carefully the documentation, and if unsuccessful, by trying to get a deeper understanding of the core source of the problem, and if not in my power to change, trying to circumvent it.

Data set issues

While doing the randomized testing (given the 134K chemical space of molecular structures), I realized that the spec for the XYZ format as presented in the paper was slightly incomplete.

For example, in the case of gdb dataset # 014650, the claim that the number of harmonic vibrational frequencies is equal to (3 * n_a − 5) or (3 * n_a − 6), in cm^{-1} does not hold (n_a being the number of atoms). There are basically 2 sets of 45 harmonic vibrational frequencies for this specific dataset, which is basically a multiplier of 2 of (3 * n_a − 6) for this molecular structure which has n_a = 17 (this is C6H9NO – polyvinylpyrrolidone.)

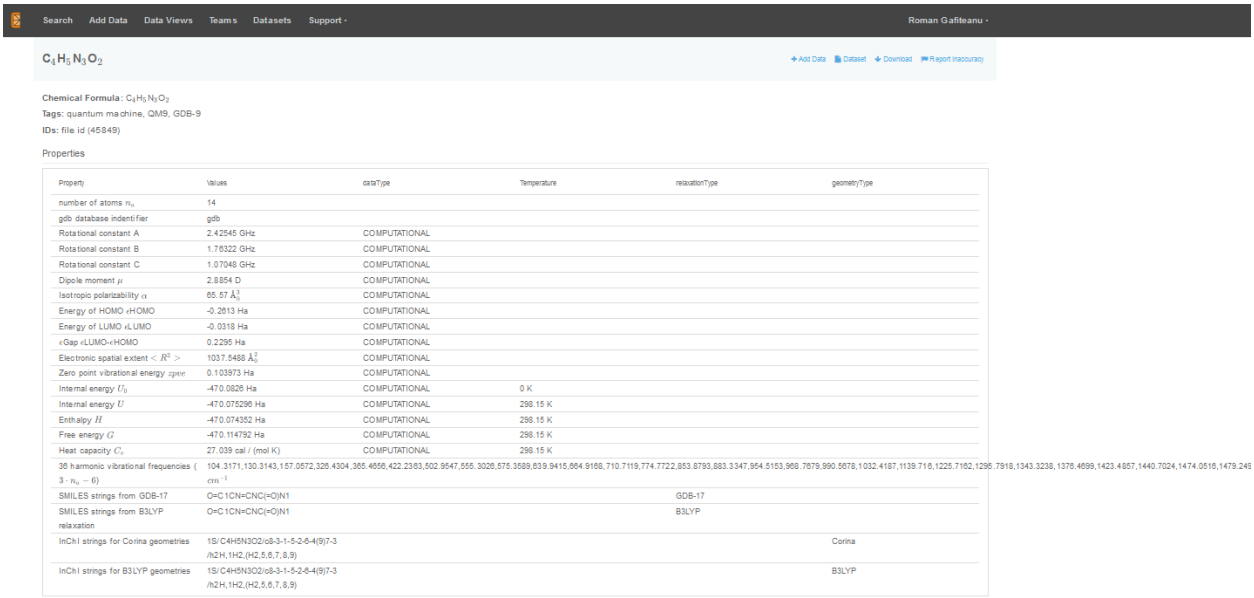
I accounted for this in the logic of the parser. You can see it in the sect4_properties() module in 2_xyz_converter.py file.

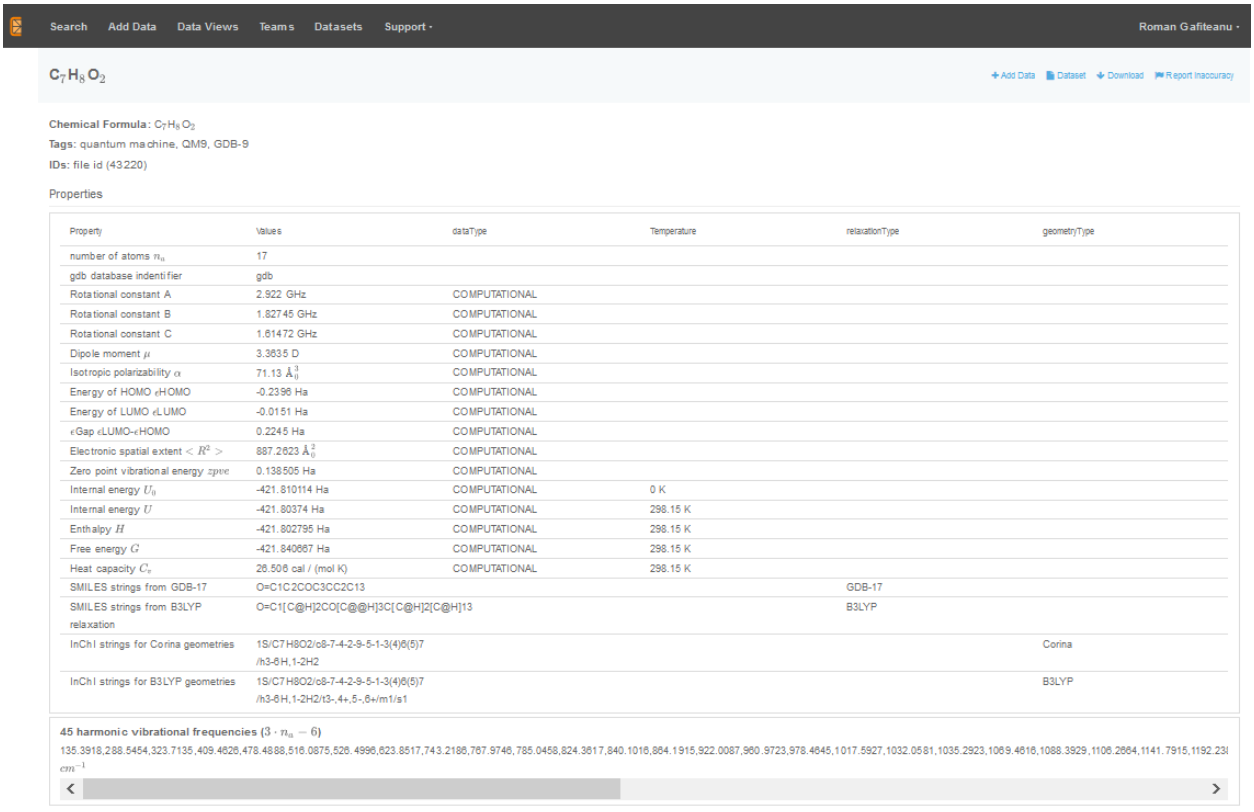
Default ingester issues

XYZ files have the "Element type, coordinate (*x*, *y*, *z*, in Å), Mulliken partial charges (in *e*) on atoms" and the "Harmonic vibrational frequencies (3 * n_a − 5 or 3 * n_a − 6, in cm^{-1})" sections - defined as sect3 and sect4 items in 2_xyz_converter.py file. Based on the structure of these types of data, I decided to store it in the PIF format as subsystem() to the ChemicalSystem(), each with its own name. You can see this by inspecting the sampling of converted PIF files that I provided on github, in the "GDB-9-molecules-pif" directory.

When I tried to make it part of the ChemicalSystem(), the formatting on the website turned out to be not appropriate, as well as inconsistent. I provide below two screen captures of two different chemical systems which were uploaded on citrination into the same dataset, at the same time, but the harmonic vibrational

frequencies ended up in different parts of the card. Similar things happened with element type, coordinates and Mulliken charges info, when I tried to make it part of the main system().





Finally, with respect to formatting, as it relates to special characters and symbols, I did not find in the documentation any specific information on how to deal with it (certainly I might have missed it.) I finally went to the “math mode” LaTex notation.

Ability to download files from Citrination

I might have been doing something wrong, but when I tried to programmatically download some of the files from the dataset that I created on citrination.com, it would do it. I tried to follow the spec for the Python Citrination Client (https://citrineinformatics.github.io/python-citrination-client/index.html) to no avail.