

## 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  $\text{cm}^{-1}$  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  $\text{cm}^{-1}$ )” 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 citration into the same dataset, at the same time, but the harmonic vibrational

	Search	Add Data	Data Views	Teams	Datasets	Support	Roman Gafteanu
--	--------	----------	------------	-------	----------	---------	----------------

  

C <sub>4</sub> H <sub>5</sub> N <sub>3</sub> O <sub>2</sub>	+ Add Data	Dataset	Download	Report Inaccuracy
---	------------	---------	----------	-------------------

Chemical Formula: C<sub>4</sub>H<sub>5</sub>N<sub>3</sub>O<sub>2</sub>  
Tags: quantum machine, QM9, GDB-9  
IDs: file id (45849)

Properties

Property	Values	dataType	Temperature	relaxationType	geometryType
number of atoms n <sub>a</sub>	14				
gdb database identifier	gdb				
Rotational constant A	2.42545 GHz	COMPUTATIONAL			
Rotational constant B	1.76322 GHz	COMPUTATIONAL			
Rotational constant C	1.07048 GHz	COMPUTATIONAL			
Dipole moment μ	2.8854 D	COMPUTATIONAL			
Isotropic polarizability α	65.87 Å <sup>3</sup>	COMPUTATIONAL			
Energy of HOMO «HOMO	-0.2013 Ha	COMPUTATIONAL			
Energy of LUMO «LUMO	-0.0318 Ha	COMPUTATIONAL			
«Gap «LUMO«HOMO	0.2295 Ha	COMPUTATIONAL			
Electronic spatial extent < R <sup>2</sup> >	1037.5498 Å <sup>2</sup>	COMPUTATIONAL			
Zero point vibrational energy zpvib	0.103973 Ha	COMPUTATIONAL			
Internal energy U <sub>0</sub>	-470.0828 Ha	COMPUTATIONAL	0 K		
Internal energy U	-470.075296 Ha	COMPUTATIONAL	298.15 K		
Enthalpy H	-470.074352 Ha	COMPUTATIONAL	298.15 K		
Free energy G	-470.114792 Ha	COMPUTATIONAL	298.15 K		
Heat capacity C <sub>v</sub>	27.039 cal / (mol K)	COMPUTATIONAL	298.15 K		
45 harmonic vibrational frequencies (3 · n <sub>a</sub> - 6)	104.3171, 130.3143, 157.0572, 326.4304, 365.4656, 422.2393, 502.9547, 555.3026, 575.3589, 639.9415, 664.9106, 710.7119, 774.7722, 853.8793, 883.3347, 954.5153, 966.7079, 990.5678, 1032.4187, 1139.716, 1226.7162, 1296.7918, 1343.3238, 1376.4699, 1423.4857, 1440.7024, 1474.0502, 1479.3186, 1500.1279, 1500.1279 cm <sup>-1</sup>				
SMILES strings from GDB-17	O=C1CN=CNC(O)N1			GDB-17	
SMILES strings from B3LYP relaxation	O=C1CN=CNC(O)N1			B3LYP	
InChI strings for Corina geometries	1S/C4H5N3O2/c8-3-1-5-2-6-4(9)/t-3/r2H,1H2,(H2.5.6.7,8.9)				Corina
InChI strings for B3LYP geometries	1S/C4H5N3O2/c8-3-1-5-2-6-4(9)/t-3/r2H,1H2,(H2.5.6.7,8.9)				B3LYP

  

SF6	Search	Add Data	Data Views	Teams	Datasets	Support	Roman Gafteanu
-----	--------	----------	------------	-------	----------	---------	----------------

  

C <sub>7</sub> H <sub>8</sub> O <sub>2</sub>	+ Add Data	Dataset	Download	Report Inaccuracy
--	------------	---------	----------	-------------------

Chemical Formula: C<sub>7</sub>H<sub>8</sub>O<sub>2</sub>  
Tags: quantum machine, QM9, GDB-9  
IDs: file id (43220)

Properties

Property	Values	dataType	Temperature	relaxationType	geometryType
number of atoms n <sub>a</sub>	17				
gdb database identifier	gdb				
Rotational constant A	2.922 GHz	COMPUTATIONAL			
Rotational constant B	1.82745 GHz	COMPUTATIONAL			
Rotational constant C	1.61472 GHz	COMPUTATIONAL			
Dipole moment μ	3.3835 D	COMPUTATIONAL			
Isotropic polarizability α	71.13 Å <sup>3</sup>	COMPUTATIONAL			
Energy of HOMO «HOMO	-0.2396 Ha	COMPUTATIONAL			
Energy of LUMO «LUMO	-0.0151 Ha	COMPUTATIONAL			
«Gap «LUMO«HOMO	0.2245 Ha	COMPUTATIONAL			
Electronic spatial extent < R <sup>2</sup> >	887.2623 Å <sup>2</sup>	COMPUTATIONAL			
Zero point vibrational energy zpvib	0.138505 Ha	COMPUTATIONAL			
Internal energy U <sub>0</sub>	-421.810114 Ha	COMPUTATIONAL	0 K		
Internal energy U	-421.80374 Ha	COMPUTATIONAL	298.15 K		
Enthalpy H	-421.802795 Ha	COMPUTATIONAL	298.15 K		
Free energy G	-421.840687 Ha	COMPUTATIONAL	298.15 K		
Heat capacity C <sub>v</sub>	26.505 cal / (mol K)	COMPUTATIONAL	298.15 K		
SMILES strings from GDB-17	O=C1C2OC3CC2C13			GDB-17	
SMILES strings from B3LYP relaxation	O=C1[C@H]2COC[C@@H]3C[C@H]2[C@H]13			B3LYP	
InChI strings for Corina geometries	1S/C7H8O2/c8-7-4-2-9-5-1-3(4)(6)/r3-9H,1-2H2				Corina
InChI strings for B3LYP geometries	1S/C7H8O2/c8-7-4-2-9-5-1-3(4)(6)/r3-9H,1-2H2/t3-,4+,5-,6-/m1/s1				B3LYP
46 harmonic vibrational frequencies (3 · n <sub>a</sub> - 6)	135.3918, 288.5454, 323.7135, 409.4626, 478.4888, 516.0875, 526.4996, 623.8517, 743.2186, 767.9746, 785.0458, 824.3617, 840.1016, 864.1915, 922.0087, 960.9723, 978.4645, 1017.5927, 1032.0581, 1035.2923, 1069.4616, 1088.3929, 1106.2664, 1141.7915, 1192.231 cm <sup>-1</sup>				

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 uploaded on citrination.com, it wouldn't do it. I tried to follow the spec for the [Python Citrination Client](#) to no avail.