

Testing InChI v1.07.0

**Jan C. Brammer,
RWTH Aachen**

24.07.2024

Pub  hem

Test infrastructure

- InChI 1.07.0 compiled with GCC 14.1.0
- Debian bookworm
- 16 physical cores
- InChI API's `MakeINCHIFromMolfileText`, `GetINCHIKeyFromINCHI` (without parameters)
- <https://github.com/IUPAC-InChI/InChI/tree/main/INCHI-1-TEST>

PubChem Datasets

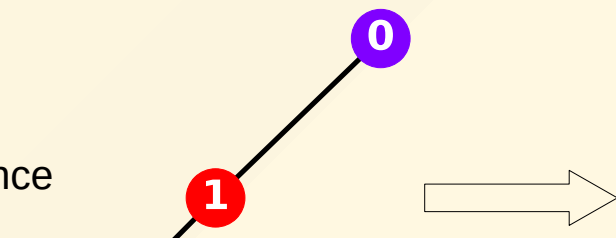
<https://ftp.ncbi.nlm.nih.gov/pubchem/>

	Compound	Compound 3D	Substance
download ^a	Oct 13 2023	Oct 25 2023	Oct 23 2023
size in GB (gzip) ^b	99	37	81
N SDF ^c	338	1,103	895
N structures ^d	114,726,411	23,487,296	306,711,305

Regression

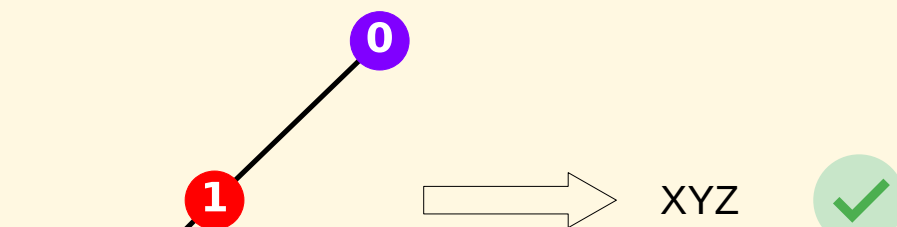
Are InChIs stable across version 1.06 and version 1.07?

Reference run



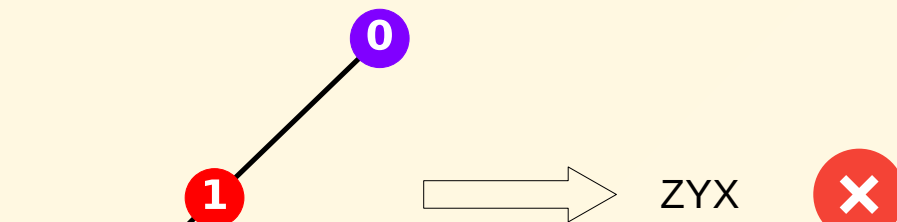
XYZ

1st run



XYZ

2nd run



ZYX

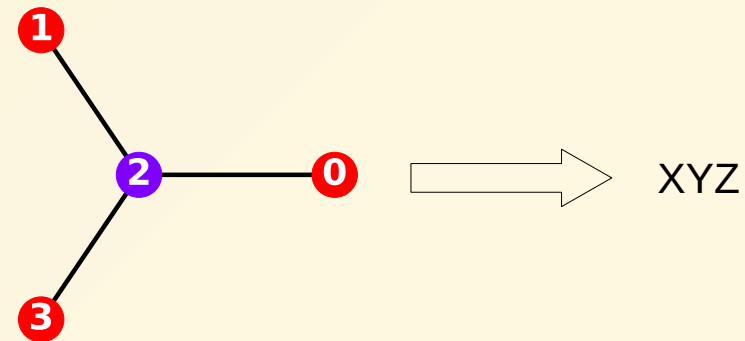
Regression Results

	Compound	Compound 3D	Substance
N structures ^e	114,726,411	23,487,296	306,711,305
N structures passed ^f	114,726,411	23,487,296	306,711,303
N structures failed ^g	0	0	2
percentage failed ^h	0	0	0.00000064
run-time total ⁱ	402 min (6 hrs, 42 min)	106 min (1 hr, 46 min)	585 min (9 hrs, 45 min)
avg run-time per structure ^j	0.21 ms	0.27 ms	0.10 ms

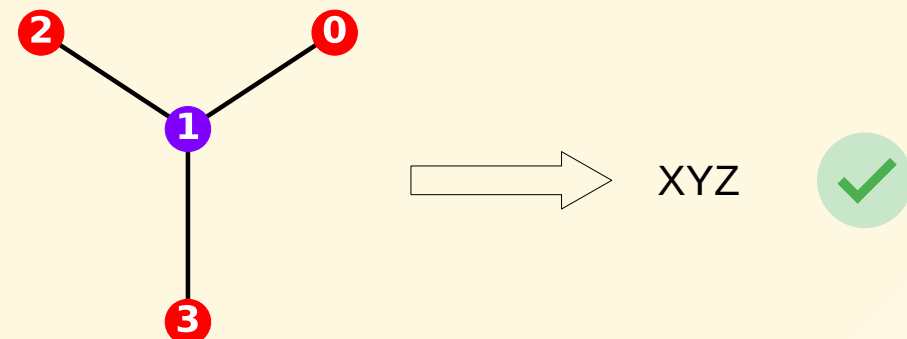
Invariance

Are InChIs canonical?

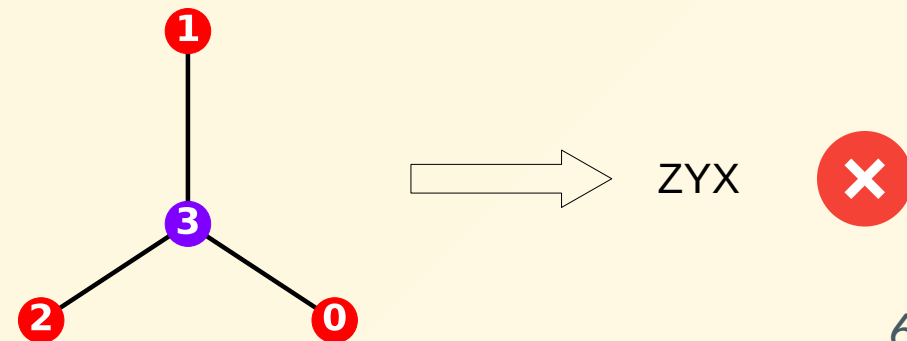
1st
permutation



2nd
permutation



3rd
permutation



Invariance Results

	Compound	Compound 3D	Substance
N structures ^k	114,726,411	23,487,296	306,711,305
N structures missing ^l	0	0	16,932,378
N structures error ^m	n/a	0	21
N structures passed ⁿ	n/a	23,487,290	289,776,775
N structures failed ^o	n/a	6	2,131
percentage failed ^p	n/a	0.000026	0.000735
run-time total ^q	n/a	389 min (6 hrs, 29 min)	4,063 min (2 days, 18 hrs, 43 min)
avg run-time per structure ^r	n/a	0.98 ms	0.84 ms

Example Invariance Failure Substance

<https://pubchem.ncbi.nlm.nih.gov/substance/140565978>

10 permutations resulted in 2 InChI variants.

Below, the identical /c and /p layers are omitted to make difference in /h layer more salient.

```
InChI=1S/C55H53N3O14S2/h7,9-10,13-14,16-25,27-30H,6,8,11-12,15,26,31H2,1-5H3,(H5-,56,59,61,62,63,64,65,66,67,68,69)
```

```
InChI=1S/C55H53N3O14S2/h7,9-10,13-14,16-25,27-30H,6,8,11-12,15,26,31H2,1-5H3,(H5-,56,59,60,61,62,63,64,65,66,67,68,69)
```


Example Invariance Failure Compound 3D

<https://pubchem.ncbi.nlm.nih.gov/compound/6401426>

10 permutations resulted in 2 InChI variants.

```
InChI=1S/C10H13N6O2/c1-3-7-6(2)14-16-5-12-15(4-8(17)13-11)10(16)9(7)18/h5,11,18H,3-4H2,1-2H3/q-1/p+1  
InChI=1S/C10H14N6O2/c1-3-7-6(2)14-16-5-12-15(4-8(17)13-11)10(16)9(7)18/h5,11,14,18H,3-4H2,1-2H3
```

Details PubChem Datasets

a)

```
find . -type f -name "*.gz" -exec du -b {} + | awk '{ total += $1 } END {  
print total / 1024 / 1024 / 1024 " GB" }'
```

b) according to `.listing` file from PubChem FTP download

c) `ls *.sdf.gz | wc -l`

d)

```
totalCount=0; for file in ./*.sqlite; do count=$(sqlite3 "$file" "SELECT  
COUNT(*) FROM results;"); totalCount=$((totalCount + count)); done; echo  
$totalCount
```

Details Regression

e) see a)

f) $N \text{ structures} - (N \text{ structures missing} + N \text{ structures error} + N \text{ structures failed})$

g) `grep -o "test failed" ./<log-name>.log | wc -l`

h) $N \text{ structures failed} / (N \text{ structures} - (N \text{ structures missing} + N \text{ structures error})) * 100$

i) last timestamp - first timestamp from logs

j) $\text{run-time total} / (N \text{ structures passed} + N \text{ structures failed}) * 60000$

Details Invariance

k) see a)

l) `grep -o "test didn't run" ./<log-name>.log | wc -l`; empty molfiles; see e.g.,

<https://pubchem.ncbi.nlm.nih.gov/rest/pug/substance/sid/2167/record/SDF>

m) `grep -o "RuntimeError" ./<log-name>.log | wc -l`; InChI failed to process molfiles

n) see f)

o) see g)

p) see h)

q) see i)

r) see j)