1 Data curation

Manually curated

Chemistry olympiads

University exams

University exercise sheets

Semi-programatically curated

Property databases

Name to SMILES

Programatically curated

Number NMR peaks

Number isomers

2 Semantic annotation

Reactions

`[START_RXNSMILES] [END_RXNSMILES]`

Molecules

`[START_SMILES] [END_SMILES]`

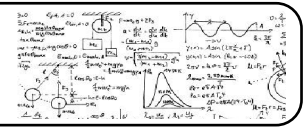
`\ce{C6H6}`

Units

`\pu{m^{-3}}`

Equations

`$a^2 + b^2 = c^2$`



3 Review

Manual inspection

Factual correctness

Clarity and phrasing

Error analysis

Automatic checks

Schema

Invariance to shuffling

Spelling