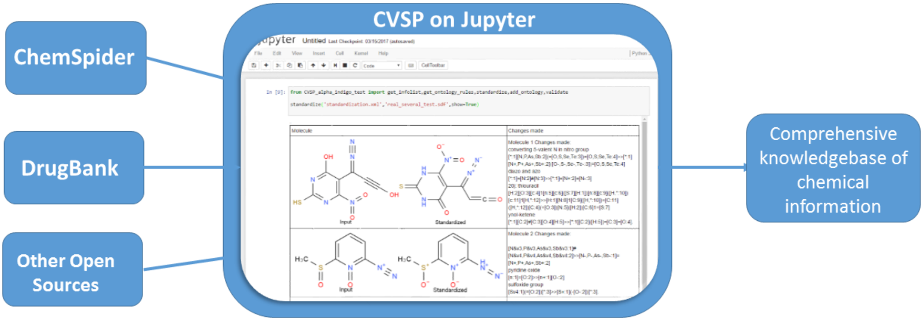
B. Sattarov [1,2]
A. Khayrullina [1,2]
V. Tkachenko [1]

USING CHEMICAL VALIDATION AND STANDARDIZATION PLATFORM TO VALIDATE AND STANDARDIZE PUBLICLY AVAILABLE CHEMICAL DATA

[1] SCIENCE DATA SOFTWARE, LLC, Rockville, MD, USA
[2] A.M. Butlerov Insitute of Chemistry, Kazan Federal University, Russia

Brois475@gmail.com

Due to the rapid development of publicly available online chemical databases and the fact that there is a wide spectrum of issues associated with chemical structure representation, it seemed reasonable to develop a freely available platform for the processing of chemical compound datasets. The first version of Chemistry Validation and Standardization Platform (CVSP) was created to support Chemistry Registry System for OpenPHACTS [1]. While CVSP proven to be an extremely useful in indentifying problems with chemical compounds in a variety of openly available datasets including HMDB, DrugBank, ChEBI, ChEMBL, PDB, MeSH and SureChEMBL, its use was also limited due to affinity to Microsoft platform, reliance on commercial cheminformatics toolkits and complex way of system deployment. To address some of these issues we have introduced Jupyter-based [2] version of CVSP [3] (Chemical Standardization and Validation Platform) written in Python 3. In combination with our chemical web-scrapping python library the platform delivers very convenient tool for building a comprehensive knowledgebase of chemical information.


ChemSpider
DrugBank
Other Open Sources
CVSP on Jupyter
Jupyter Untitled Last Checkpoint: 03/15/2017 (autosaved)
File Edit View Insert Cell Kernel Help
Python
Code
CellToolbar
Molecule
Changes made
Input
Standardized
Input
Standardized
Comprehensive knowledgebase of chemical information

```
In [9]: from CVSP_alpha_indigo_test import get_infolist,get_ontology_rules,standardize,add_ontology,validate

        standardize('standardization.xml','real_several_test.sdf',show=True)
```

```
Molecule 1 Changes made:
converting 5-valent N in nitro group
diazo and azo
20): thiouracil
ynol-ketene
Molecule 2 Changes made:
pyridine oxide
sulfoxide group
```

Standardization script allows user to automatically standardize all molecules in the dataset and Validation script displays warnings about possible inconsistencies introduced at a stage of creating chemical structure or transcoding data between various formats. CVSP gives user information about "suspicious" fragments of the molecule that might require standardization. This is a very convenient feature, especially when user needs to check large datasets, but wants to standardize the molecule manually. CVSP both validates and standardizes chemical structure representations according to sets of systematic SMIRKS-based rules in XML format, which can be added and edited manually. Initial set of validation rules consisted of the structure depiction rules found in the Substance Registry System document issued by the Food and Drug Administration and IUPAC Blue Book.

Platform can also be used to relate molecules to different ontological classes of chemical compounds. We used the ontological principles of the ChEBI (Chemical Entities of Biological Interest)**.** This can be extremely useful for on-the-fly classification of organic compounds and will be published as RESTful services suite.

1. http://www.openphacts.org/ (accessed December 2016).
2. http://jupyter.org/ (accessed December 2016).
3. Karapetyan K. et al. *Journal of cheminformatics*, 2015, **7** **(1)**: 30.