Version 1.0

Chemistry Validation and Standardization Platform (CVSP)

What is CVSP?

CVSP (Chemistry Validation and Standardization Platform) is a platform which allows you to upload chemical structure files which are then validated and optionally standardised in preparation for publication or submission to a chemical database.

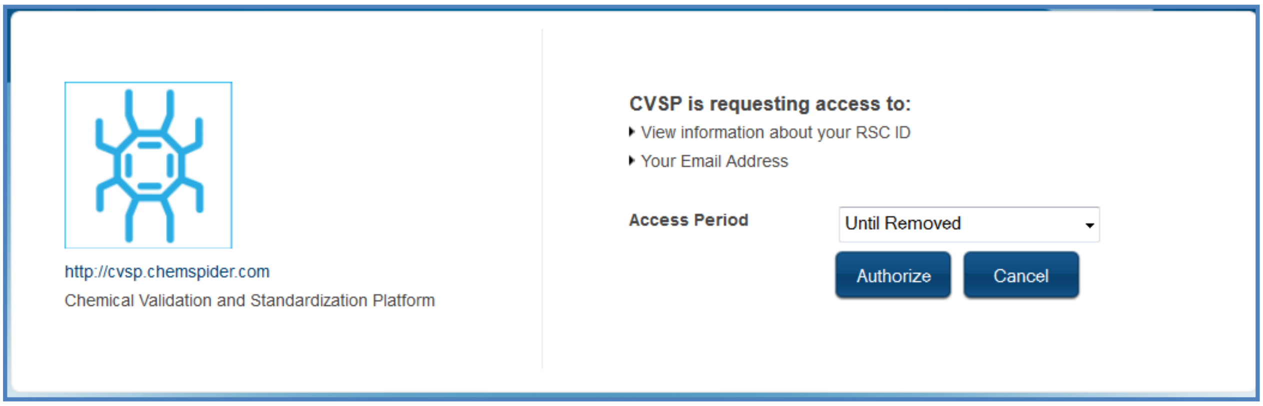
The CVSP validation rules are used to review the structures in your file and alert you to any chemical, formatting and layout issues. Standardisation amends the structures in your file by applying certain transformations either to fix some of the validation issues or to unify records according to standardization rules, e.g. de-aromatization, tautomer canonicalization, etc. You can choose to only validate the structures in your file, or to validate and standardise them.

How do I use it?

To use CVSP, you will need to log in to the site which is achieved by using RSC ID. For more information on how to register for an RSC ID, please view the RSC ID FAQ. At the top of the page you will see a black bar that contains an option to sign in.

Linking your RSC ID

When you are logged in to your RSC ID and you come to the CVSP site, you will be asked to allow the site to access your RSC ID.



http://cvsp.chemspider.com

Chemical Validation and Standardization Platform

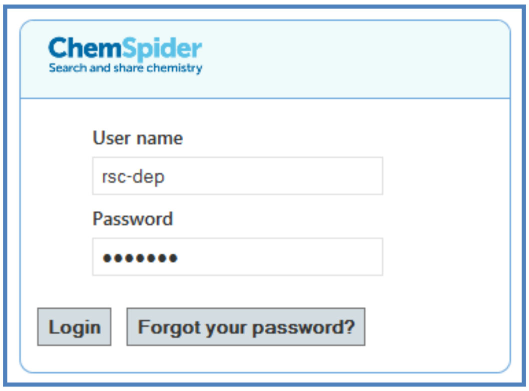
CVSP is requesting access to:

- View information about your RSC ID
- Your Email Address

Access Period: Until Removed

Once you have authorised this, you will be asked if you already have a ChemSpider account. If you need to register a new account, click “No” and a new ChemSpider account will be created for you.

If you already have an account, click “Yes”. You will be directed to a login screen.



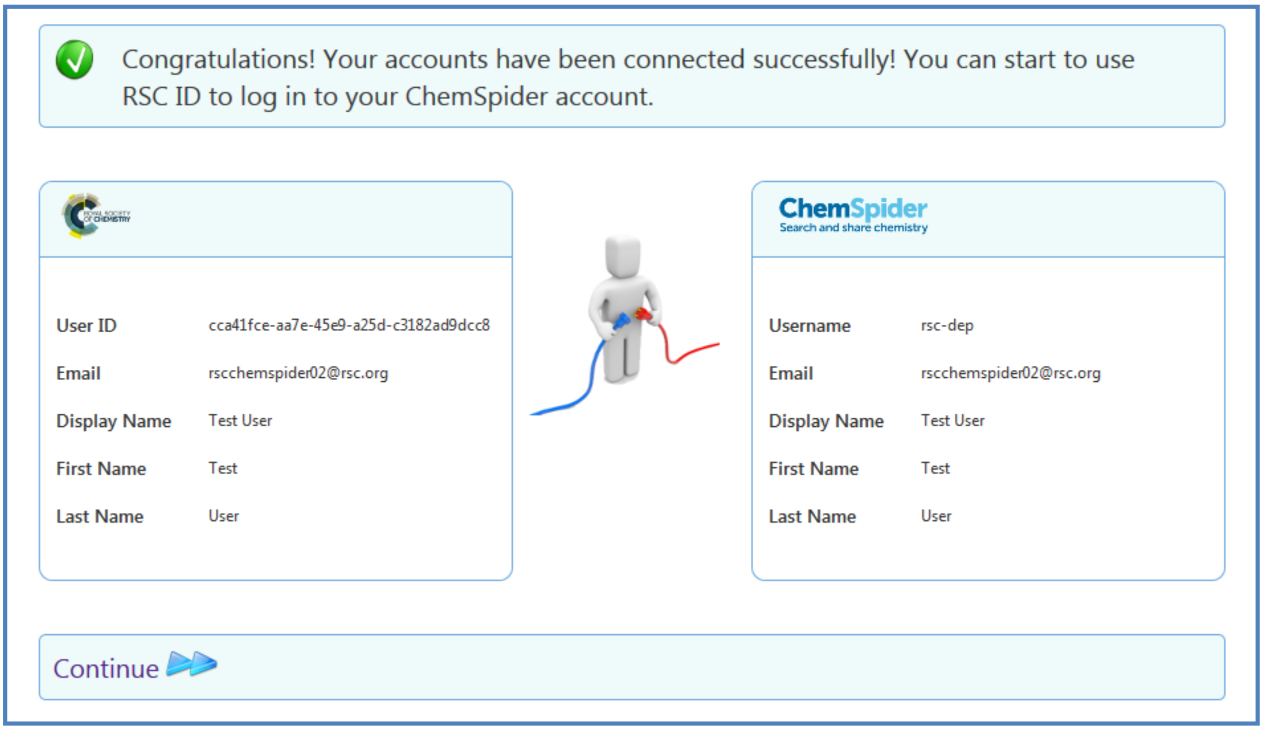
User name

rsc-dep

Password

•••••••

Fill in your ChemSpider user name and password, and click **Login**. Then on the next screen click **Next** to confirm. Your ChemSpider account is now linked to your RSC ID, and you should now be able to upload files.



Congratulations! Your accounts have been connected successfully! You can start to use RSC ID to log in to your ChemSpider account.

User ID	cca41fce-aa7e-45e9-a25d-c3182ad9dcc8
Email	rscchemspider02@rsc.org
Display Name	Test User
First Name	Test
Last Name	User

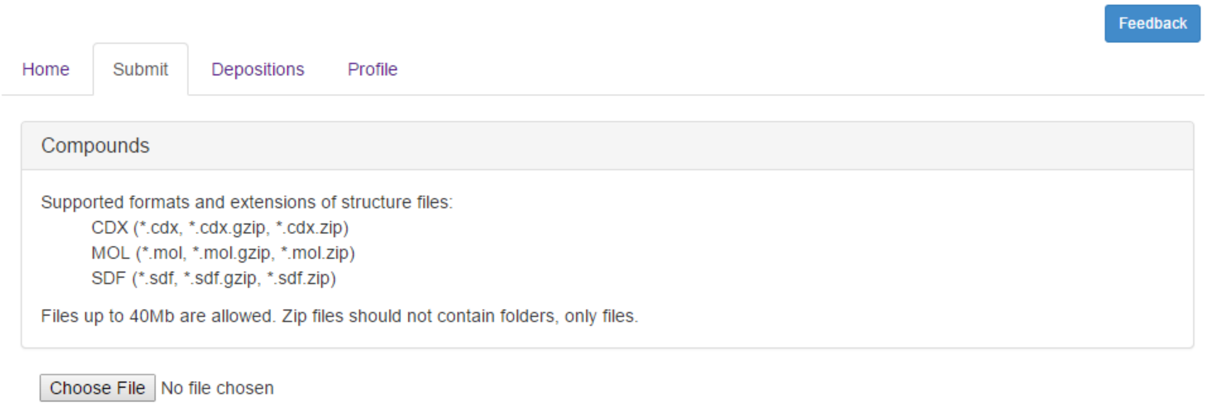
Username	rsc-dep
Email	rscchemspider02@rsc.org
Display Name	Test User
First Name	Test
Last Name	User

Click **Continue** to return to CVSP.

Once logged in you will see a home page with 4 tabs: **Home (**the page that you are on when you log in), **Submit (**the interface for uploading your data), **Depositions (**review and manage previously submitted files and check progress of files that are being processed), **Profile (**inspect the default rules, clone and edit the default rules, or create your own rules from scratch).

Uploading a file

When you have logged in to your CVSP account, you can use the **Submit** tab to access the interface to upload new file for processing.



Compounds

Supported formats and extensions of structure files:
CDX (*.cdx, *.cdx.gzip, *.cdx.zip)
MOL (*.mol, *.mol.gzip, *.mol.zip)
SDF (*.sdf, *.sdf.gzip, *.sdf.zip)

Files up to 40Mb are allowed. Zip files should not contain folders, only files.

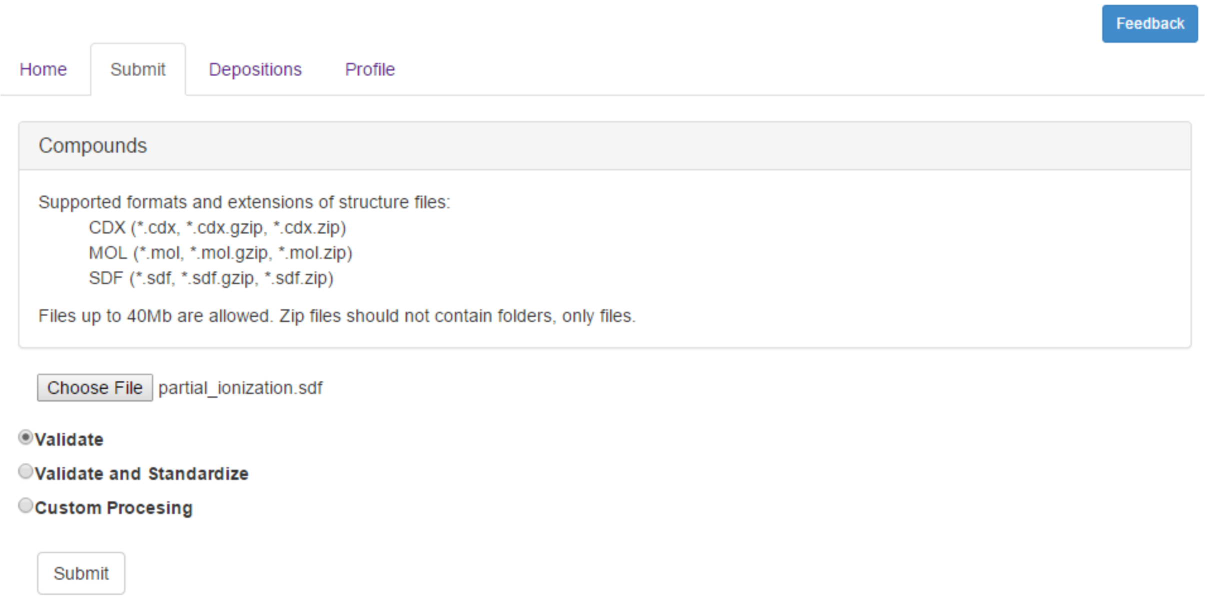
Choose File No file chosen

The system accepts .cdx, .mol and .sdf files. You can also upload .zip or .gz files containing one or more structure files. Use the choose file button to select the file which you want to process.

- **Please Note:** if uploading a compressed files please make sure that your file names conform to the expected structure
- **Please note**: that no folder(s) are allowed in Zip files.

We allow users to upload files up to a maximum file size of 40Mb through the web interface. If your file is over this size, you will need to apply for FTP access. Email chemspiderdev-at-rsc.org with your name, user name and approximate disk space required.

One you have selected a file, the page will update and display processing options. There are 3 options to select from: Validation (1), Validation and Standardization (2), and Custom Processing (3). Options (1) and (2) are using default CVSP rules for validation and standardization. Option (3) allows users to select their own rule sets (see more in the rule set section).



Compounds

Supported formats and extensions of structure files:
CDX (*.cdx, *.cdx.gzip, *.cdx.zip)
MOL (*.mol, *.mol.gzip, *.mol.zip)
SDF (*.sdf, *.sdf.gzip, *.sdf.zip)

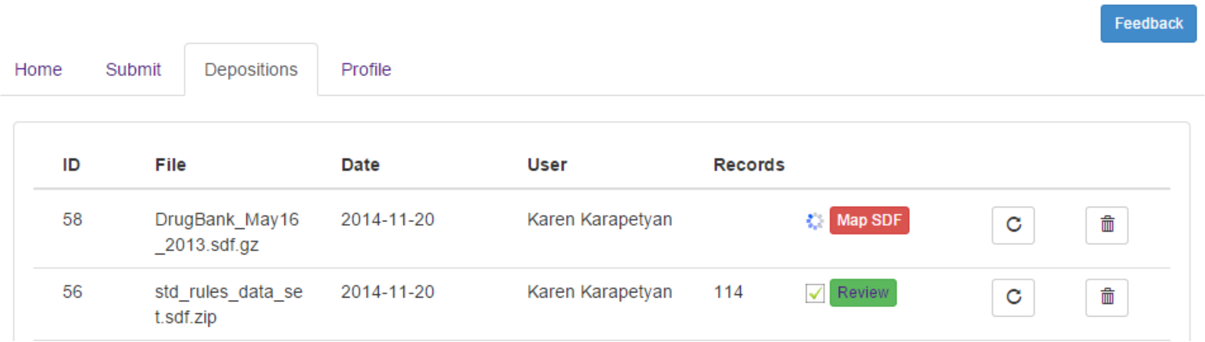
Files up to 40Mb are allowed. Zip files should not contain folders, only files.

Choose File partial_ionization.sdf

● Validate
○ Validate and Standardize
○ Custom Procesing

Hit the **Submit** button to start the processing.

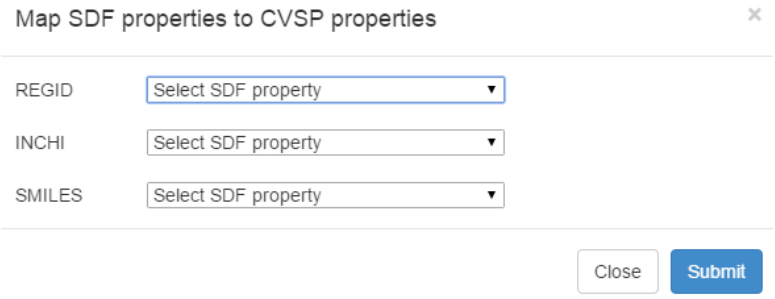
You will automatically be taken to the "Depositions" page where you will be able to see any file you have uploaded, as well as any submissions that have been made public by administrators (for demonstration purposes).



ID	File	Date	User	Records			
58	DrugBank_May16 _2013.sdf.gz	2014-11-20	Karen Karapetyan		Map SDF	C	🗑
56	std_rules_data_se t.sdf.zip	2014-11-20	Karen Karapetyan	114	✓ Review	C	🗑

If you supplied a correctly named sdf file (*.sdf or *.sdf.gz, or *.sdf.zip) and CVSP has found SDF properties then shortly after the submission you should see **Map SDF** button that allows you to map certain fields from your submitted SDF file to the internal CVSP fields to allow further validation tests, the internal fields are as follows:

- **Regid** – your unique identifier for the record
- **SMILES** – your calculated SMILES string (if you map this field CVSP will compare the SMILES you provide with the structure that you provide)
- **InChI** – your calculated standarrd InChI string (as with the SMILES field; if you map this field CVSP will compare the InChI you provide against the structure)

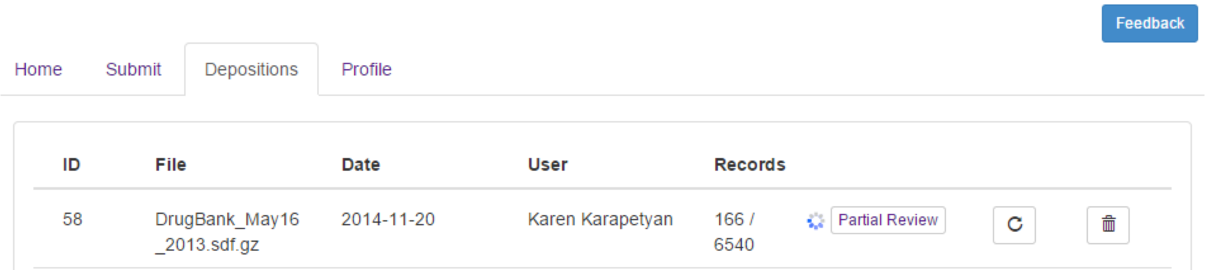

Map SDF properties to CVSP properties

REGID: Select SDF property

INCHI: Select SDF property

SMILES: Select SDF property

The amount of time it will take to process your file will depend on the size of the file and any files that are already in the processing queue. The depositions page will refresh and as an indication of progress it will show the number of records in the submitted file and the number of those that have been proccessed.



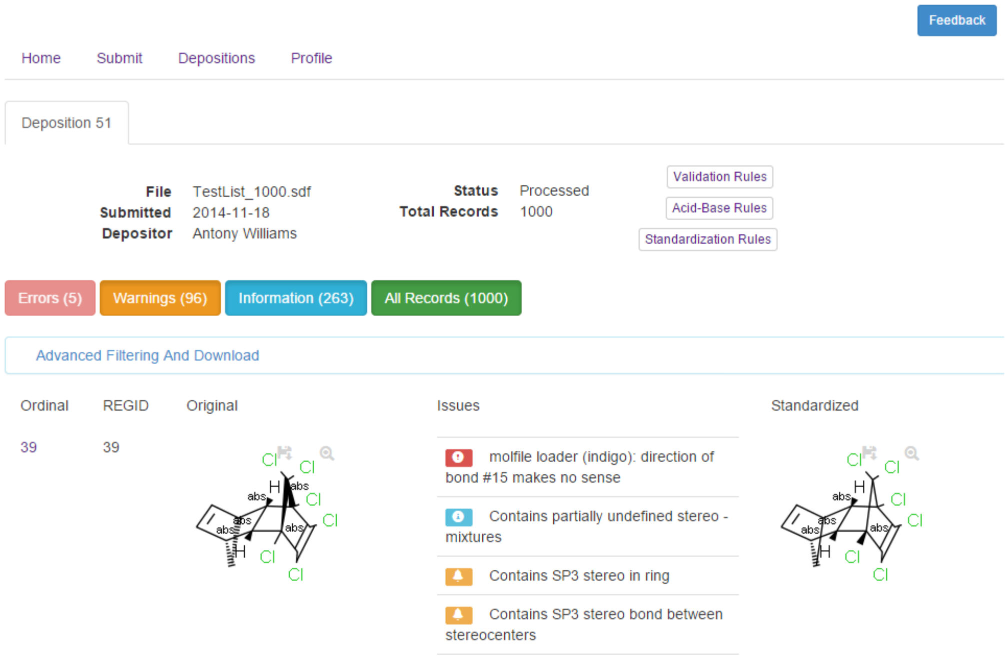
ID	File	Date	User	Records			
58	DrugBank_May16 _2013.sdf.gz	2014-11-20	Karen Karapetyan	166 / 6540	Partial Review	C	🗑

The **Partial Review** button will be available to review the records that are already processed. Once all records processed it will be replaced by green **Review** button.

Reviewing your processed file

When submission file has been processed (or partially processed), user can navigate to deposition details page by clicking on **Review** or **Partial Review** button. At the top of the page basic information about the file and the submission are displayed – name of the file, date of submission, submitter’s name, submission status, total records in the file. User also can see the rules that were applied to the submission: validation rules, acid-base rules, standardization rules (only if “Validate and Standardize” option was selected). By clicking on the rule links user can navigate to the rule content pages.

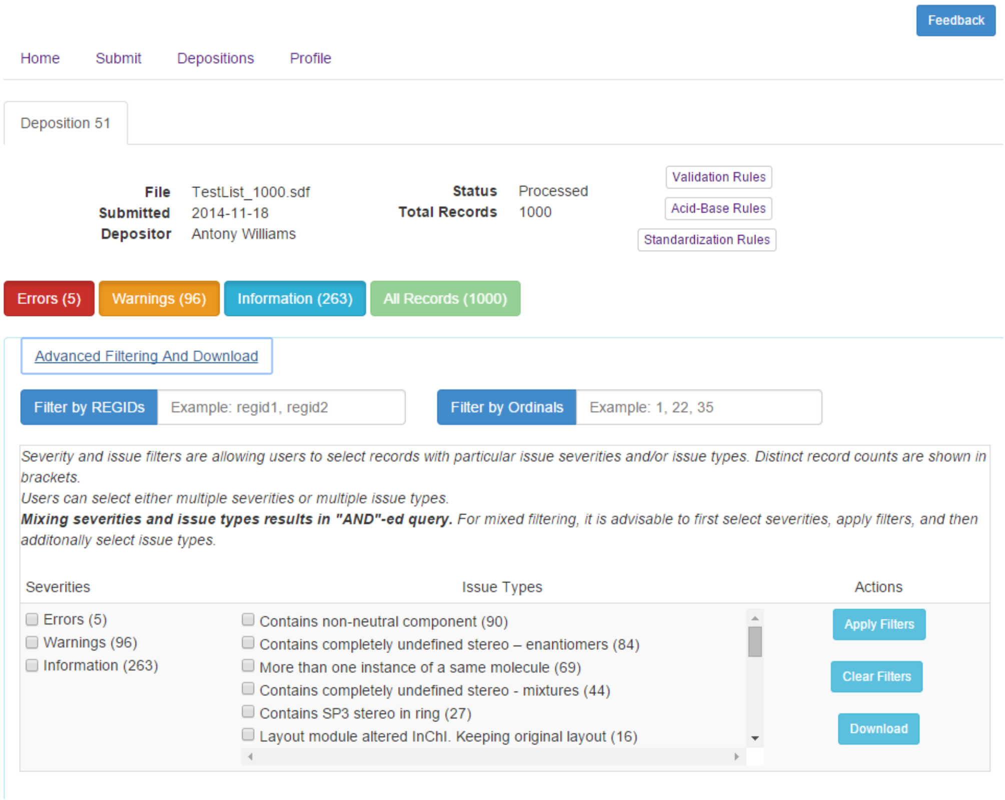
By default, when navigated to deposition details web page only records containing errors will be shown (red “Errors” button is inactive and pressed). By clicking on yellow ”Warning” button user may apply warnings filter and display only records containing warninig issues. Accordingly, blue “Information” button would apply filter to only display records containing informational issues, and green “All Records” will show all records (thus removing all filters). Colored buttons show the count of unique records that a particular severity has been found in.



File	TestList_1000.sdf	**Status**	Processed
Submitted	2014-11-18	**Total Records**	1000
Depositor	Antony Williams		

Ordinal	REGID	Original	Issues	Standardized
39	39		❗ molfile loader (indigo): direction of bond #15 makes no sense ℹ Contains partially undefined stereo - mixtures 🔔 Contains SP3 stereo in ring 🔔 Contains SP3 stereo bond between stereocenters	

Alternatively, you can use the **Advanced Filtering and Download** option to review specific subsets of data filtered by severity, issue types, REGIDs, etc. and to download them.



File TestList_1000.sdf
Submitted 2014-11-18
Depositor Antony Williams
Status Processed
Total Records 1000

Filter by REGIDs: Example: regid1, regid2
Filter by Ordinals: Example: 1, 22, 35

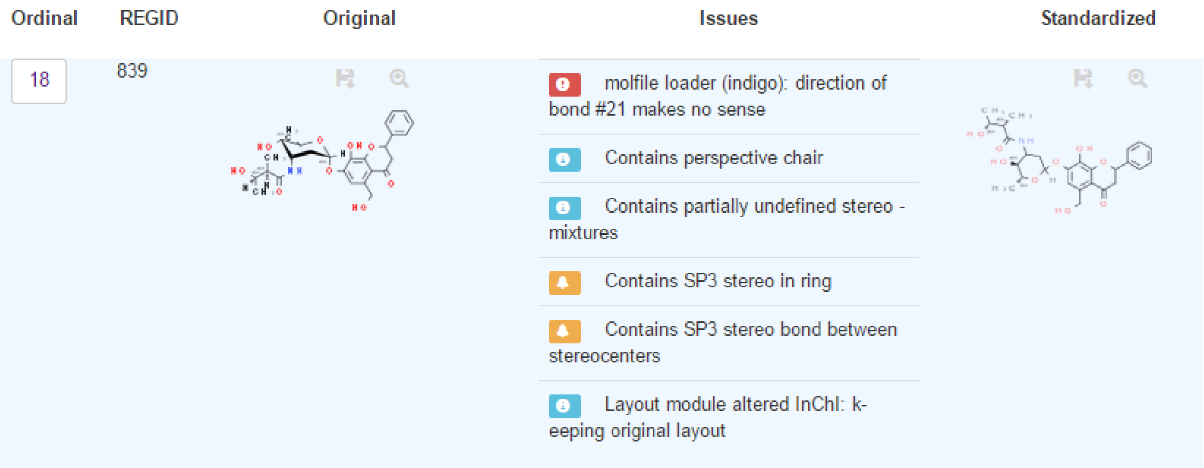
Severity and issue filters are allowing users to select records with particular issue severities and/or issue types. Distinct record counts are shown in brackets.
Users can select either multiple severities or multiple issue types.
***Mixing severities and issue types results in "AND"-ed query.** For mixed filtering, it is advisable to first select severities, apply filters, and then additonally select issue types.*

Severities	Issue Types	Actions
□ Errors (5)	□ Contains non-neutral component (90)	Apply Filters
□ Warnings (96)	□ Contains completely undefined stereo – enantiomers (84)	
□ Information (263)	□ More than one instance of a same molecule (69)	Clear Filters
	□ Contains completely undefined stereo - mixtures (44)	
	□ Contains SP3 stereo in ring (27)	Download
	□ Layout module altered InChI. Keeping original layout (16)	

- **Filter by REGID** and **Filter by Ordinals** allow you to specify comma-separated REGIDs or ordinals of the records that you wish to filter out. These 2 filters are independent of the Issues/Severities filters and will overrule other filters when invoked
- Each issue has a single assigned severity, e.g. issue type "Contains SP3 stereo in ring" belongs to "Warning" severity type. So by selecting one or more severities and applying filter ("Apply Filter" button) the values in "Issue Types" column will be refreshed to display the relevant issues
- Severities filters allow user to filter records having either of the selected severities
- Issues filters allow user to filter records having either of the selected issues
- All severities and issues have unique record counts included in brackets
- A single record may have multiple issues and multiple severities linked ot those issues
- To download subset of records with particular severities and/or issues select the desired filters, then click the **Apply Filters** button. You can then use the **Download** button to instruct CVSP to download SDF file with the filtered records. Users will be redirected to a new auto-refreshing web page that will eventually show the download link. The time required to generate the download file depends on the size of the file. The downloaded SDF file will include validation issues as SDF properties
- To download all records with validation messages hit "Clear Filters" button, again open "Advanced Filtering and Download" section and click "Download" button

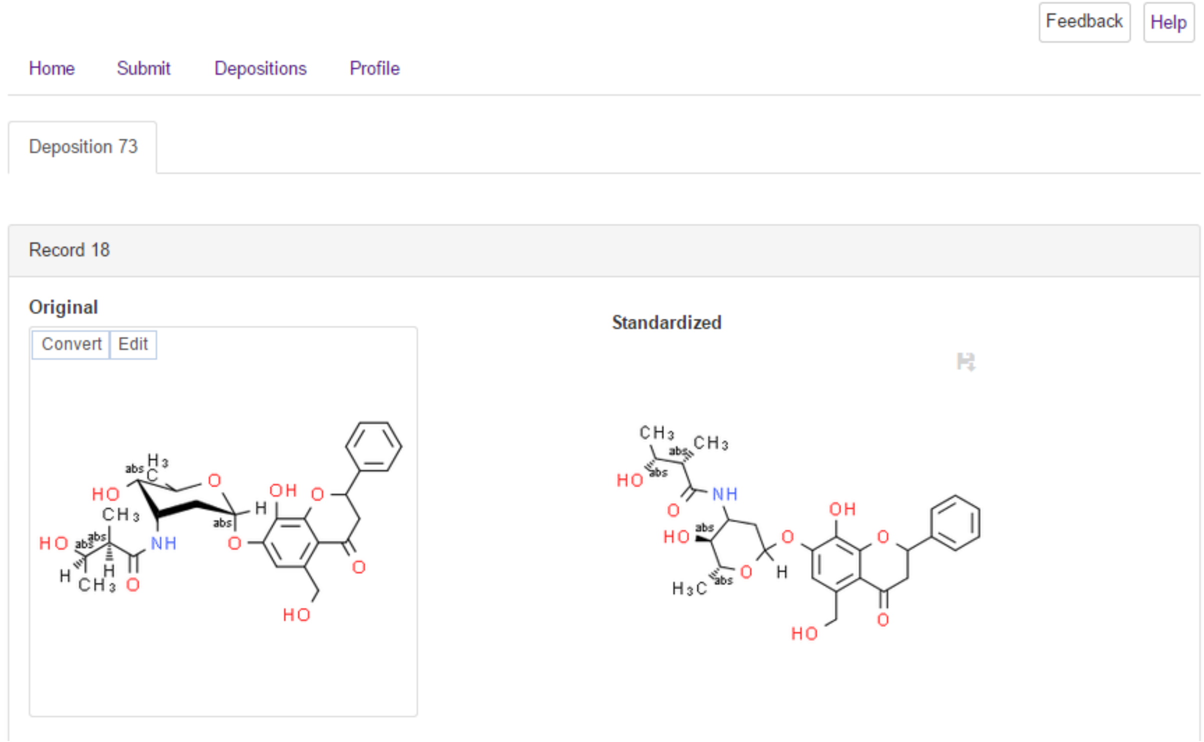
Each record in deposiiton details page has several columns: the **ordinal number** (the ordinal number of the record in the original file), **REGID** (not empty if user mapped this field at submission), **Original** (image of the original original record), **Issues** (a list of all of the issues found for that record).

Color coded icons at each issue indicate the severity associated with the issue: red - error, yellow - warning, blue – information. If user chose the “Validate and Standardize” option at submission time then **Standardized** column would show the structure after standardisation rules have been applied. Both original and standardized images have “Save” and “Zoom” buttons.



Ordinal	REGID	Original	Issues	Standardized
18	839		❗ molfile loader (indigo): direction of bond #21 makes no sense ℹ Contains perspective chair ℹ Contains partially undefined stereo - mixtures 🔔 Contains SP3 stereo in ring 🔔 Contains SP3 stereo bond between stereocenters ℹ Layout module altered InChI: k-eeping original layout	

Ordinal number button will take user to record page where user may revise the molecule and resubmit for processing.



This guide should have given you all the information you need to know to use CVSP’s basic features. Please email us at chemspiderdev@rsc.org if you have any questions or if anything isn’t clear.

Personlise your validation

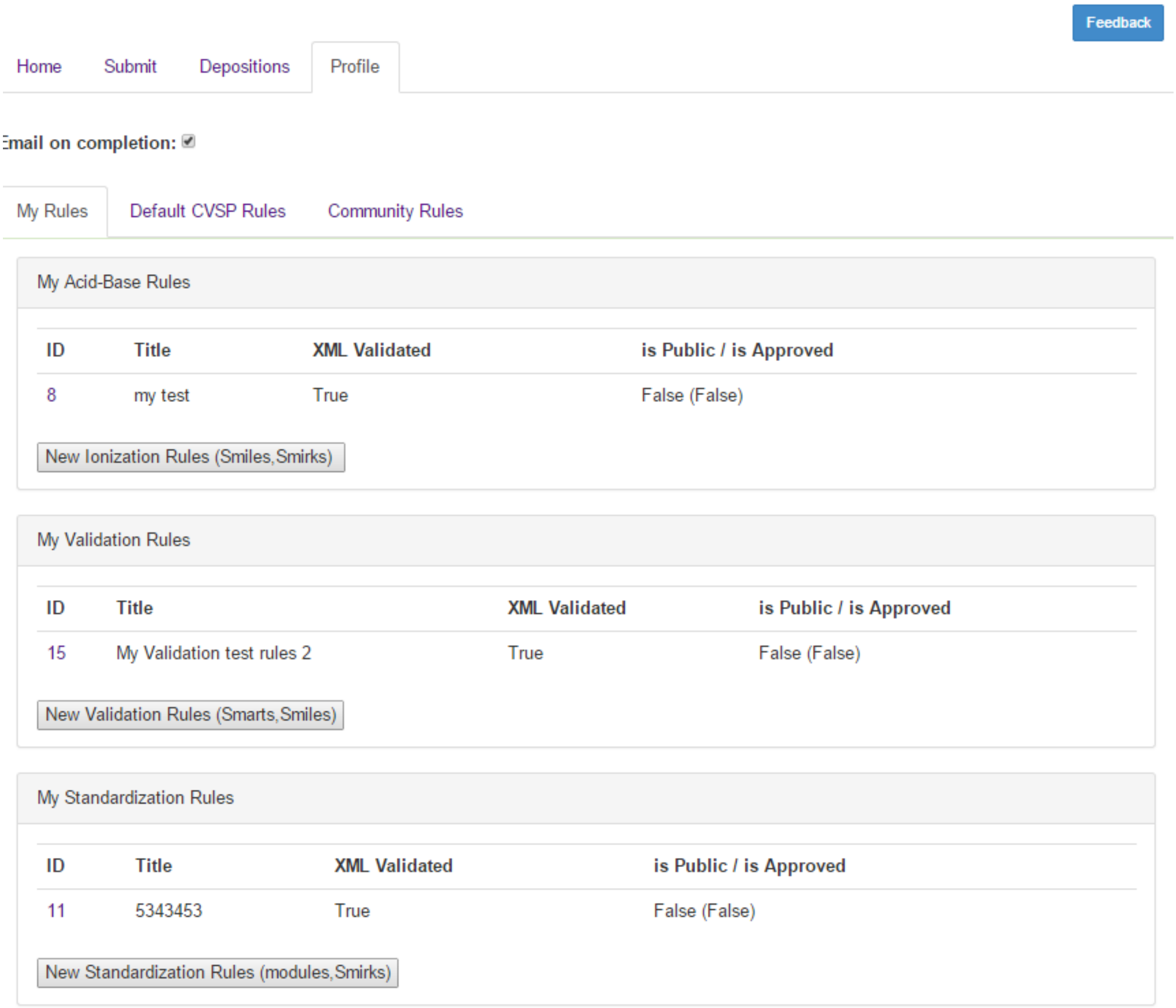
The Profile tab allows you to customize email notifications and interact with rule sets. There are 3 tabs: “My Rules”, “Default CVSP Rules”, and “Community Rules”.

My Rules. You can create your own private rules from scratch or clone the default or community rules and optionally modify them according to your own needs. Each rule has the following attributes: ID (rule identifier), title (a short descriptive title), “XML validated” flag (whether or not passed automated XML format validation), “Is Public/Is Approved” flags – “Is Public” flag shows whether or not user intended to make the rule set available for CVSP community and “Is Approved” flag shows whether or not user’s rule set has been approved to become public. New rule set buttons allow experienced users to create brand new rule sets from scratch. To avoid errors in XML format it is advisable to clone the existing default rule sets and then modify them.

Default CVSP Rules. CVSP has 3 default rule sets: validation (includes SMARTS patterns), acid-base rules (includes SMARTS and SMIRKS), and standardization rules (includes pre-defined CVSP modules and SMIRKS). The Default rules cannot be directly modified by users (though you can edit a cloned copy of the Default rules). It is recommended to at least revise the titles of the cloned rule sets for clarity.

Community Rules. These are the rules shared by community and approved by CVSP Admins. CVSP team is not responsible for the quality of user rules. Community rules are not modifiable by other users, but they can be cloned and revised.

Cloning. Cloned rules will appear under **My Rules** and will be available for selection on the **Submit** web page under “Custom Processing”. Cloned rules behave like user’s own rules and can be revised.



Feedback
Home
Submit
Depositions
Profile

Email on completion: ☑

My Rules
Default CVSP Rules
Community Rules

My Acid-Base Rules

ID	Title	XML Validated	is Public / is Approved
8	my test	True	False (False)

New Ionization Rules (Smiles,Smirks)

My Validation Rules

ID	Title	XML Validated	is Public / is Approved
15	My Validation test rules 2	True	False (False)

New Validation Rules (Smarts,Smiles)

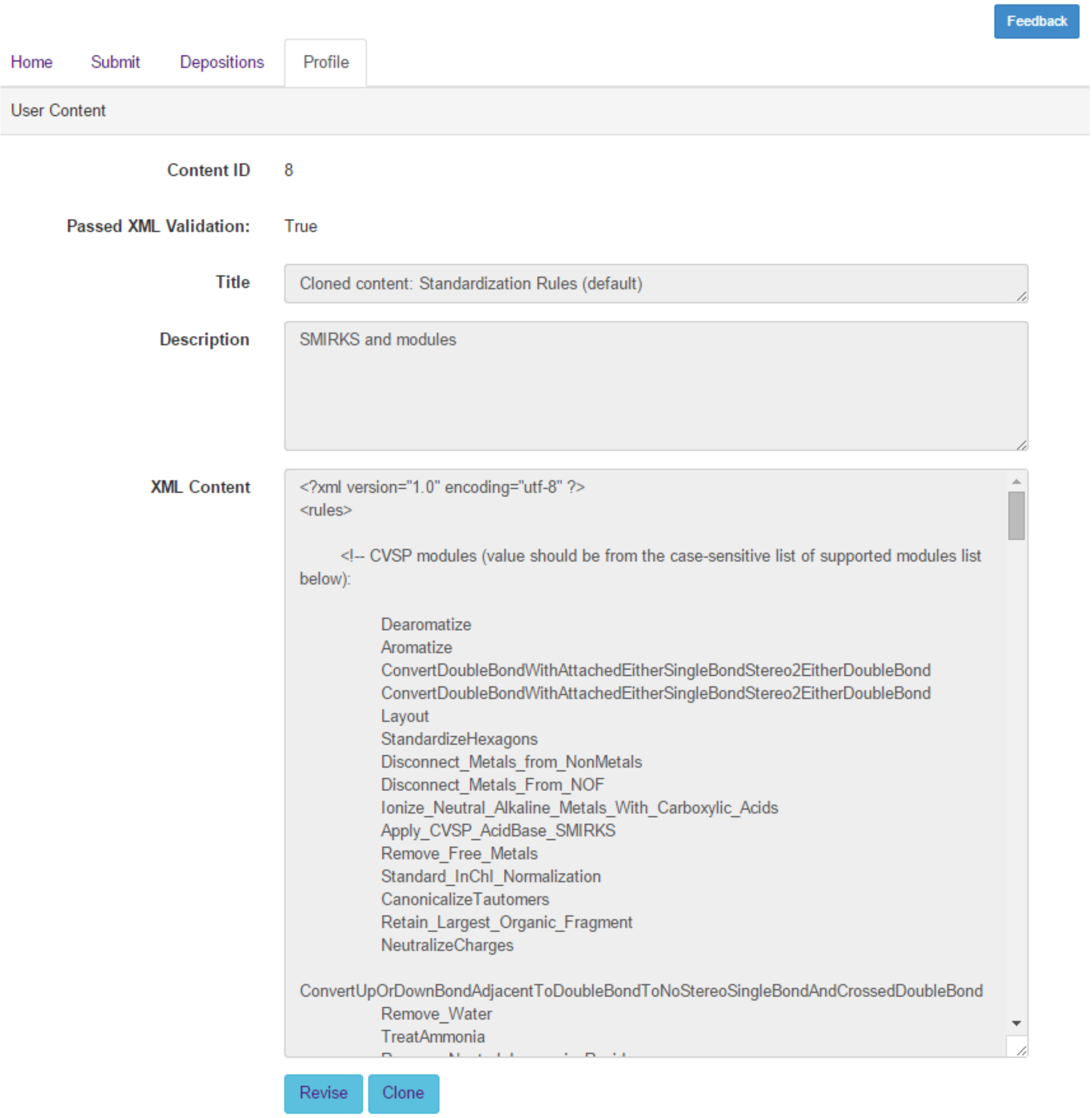
My Standardization Rules

ID	Title	XML Validated	is Public / is Approved
11	5343453	True	False (False)

New Standardization Rules (modules,Smirks)

Rule Sets

Each rule set has annotations presented below. If the rule set belongs to the user then system will allow to revise it. At each revision the XML format will be automatically validated ("Passed XML Validation" flag).



Content ID 8

Passed XML Validation: True

Title: Cloned content: Standardization Rules (default)

Description: SMIRKS and modules

XML Content:

```
<?xml version="1.0" encoding="utf-8" ?>
<rules>

      <!-- CVSP modules (value should be from the case-sensitive list of supported modules list
below):

            Dearomatize
            Aromatize
            ConvertDoubleBondWithAttachedEitherSingleBondStereo2EitherDoubleBond
            ConvertDoubleBondWithAttachedEitherSingleBondStereo2EitherDoubleBond
            Layout
            StandardizeHexagons
            Disconnect_Metals_from_NonMetals
            Disconnect_Metals_From_NOF
            Ionize_Neutral_Alkaline_Metals_With_Carboxylic_Acids
            Apply_CVSP_AcidBase_SMIRKS
            Remove_Free_Metals
            Standard_InChI_Normalization
            CanonicalizeTautomers
            Retain_Largest_Organic_Fragment
            NeutralizeCharges

ConvertUpOrDownBondAdjacentToDoubleBondToNoStereoSingleBondAndCrossedDoubleBond
            Remove_Water
            TreatAmmonia
```

Most of the validation, acid-base, and standardization rules were developed from a subset of rules specified in the Food and Drug Administration's Substance Registration System User's Guide:

http://www.fda.gov/ForIndustry/DataStandards/SubstanceRegistrationSystem-UniqueIngredientIdentifierUNII/default.htm

We also incorporated validation rules relying on IUPAC's representation standards:

http://pac.iupac.org/publications/pac/pdf/2008/pdf/8002x0277.pdf

Validation Rules

Not all validation rules can be encoded as SMARTS patterns so many validation modules are built into the system and called with any validation rule sets. Rules that are encoded as SMARTS are included in validation XML rule set. The XML format of validation rules using SMARTS pattern is presented below.

```
<?xml version="1.0" encoding="utf-8" ?>
<rules>
    <moleculerules>
        <Warning message="Contains adjacent atoms with like charges" description="[-,--]~[-,--] or [+,++]~[+,++]">
            <or>
                <test name="SMARTStest" param="[-,--]~[-,--]"/>
                <test name="SMARTStest" param="[+,++]~[+,++]"/>
            </or>
        </Warning>
        <Warning message="Contains ethane" description="[CX4;H3][CX4;H3]">
            <test name="SMARTStest" param="[CX4;H3][CX4;H3]"/>
        </Warning>
        <Information message="Contains covalent metal-nitrogen bond" description="[{M}][#7]">
            <test name="SMARTStest" param="[{M}][#7]"/>
        </Information>
        <Information message="Contains aluminium–non-metal bond" description="[Al][{NM}]">
            <test name="SMARTStest" param="[Al][{NM}]"/>
        </Information>
        <Information message="Contains non-metal–transition metal bond" description="[{TM^Hg}][{NM}]">
            <test name="SMARTStest" param="[{TM^Hg}][{NM}]"/>
        </Information>
        <Information message="Contains ammonia where ammonium expected"
            description="[N;H3&#x26;X3] and [O,F,Cl,Br;H]">
            <and>
                <test name="SMARTStest" param="[N;H3&#x26;X3]"/>
                <test name="SMARTStest" param="[O,F,Cl,Br;H]"/>
            </and>
        </Information>
    </moleculerules>
</rules>
```

Each rule has to have severity (Error, Warning, or Information), message (short informative and unique message that will appear in CVSP as issue type), description (will be attached to message and be shown with record issues). Each rule has to include the test (usability cases are below) with the SMARTS pattern that user seeks to identify.

Some characters have to be encoded (“&” as “&#x26;”, “\” as “\\”). SMARTS may use metal abbreviations that the system will expand during the processing: “{M}” – metals; “{M_+0}“ –neutral metals, “{M_+1}“ – metals with charge +1, etc.; {NM}” – non-metals except carbon; “{NMExcOFN}" – non-metals except O, F, N; “{TM}” – transition metals; “{TM^Hg}” – transition metals except Hg; “{Hal}” – halogens; “{Hal_-1}” – anions of halogens; “{Pn}” - pnictogens.

GGA’s Indigo toolkit is used to run SMARTS queries on molecules. GGA claims that Indigo supports all of the features of “original“ DayLight SMARTS:

http://ggasoftware.com/opensource/indigo/concepts#molecules-and-query-molecules

http://www.daylight.com/dayhtml/doc/theory/theory.smarts.html

SMARTS viewer: http://smartsview.zbh.uni-hamburg.de/smartsview/view

SMIRKS tool: http://www.daylight.com/daycgi_tutorials/react.cgi

Acid-Base Rules

Acid-Base rules are used to determine protonation/ionisation states in structures with multiple acid-base centres the strongest acid is ionized before the weakest one. The default acid-base rule set includes 30+ rules. The XML format of the acid-base rule set is presented below.

```
<?xml version="1.0" encoding="utf-8" ?>

<acidgroups>

        <!-- OSO3H-->

        <acidgroup rank="10" acid="OS(=O)(=O)[O;H]" base="OS(=O)(=O)[O-]"

                acid2base="[*:1][O:2][S:3](=[O:4])(=[O:5])[O:6][H:7]>>[*:1][O:2][S:3](=[O:4])(=[O:5])[O-:6].[H+:7]"

                base2acid="[*:1][O:2][S:3](=[O:4])(=[O:5])[O-:6].[H+:7]>>[*:1][O:2][S:3](=[O:4])(=[O:5])[O:6][H:7]"/>

         <!-- SO3H-->

         <acidgroup rank="20" acid="[!O]S(=O)(=O)[O;H]" base="[!O]S(=O)(=O)[O-]"

                acid2base="[!O:1][S:2](=[O:3])(=[O:4])[O:5][H:6]>>[!O:1][S:2](=[O:3])(=[O:4])[O-:5].[H+:6]"

                 base2acid="[!O:1][S:2](=[O:3])(=[O:4])[O-:5].[H+:6]>>[!O:1][S:2](=[O:3])(=[O:4])[O:5][H:6]" />

</acidgroups>
```

Each acid-base rule has rank (the lower the rank the stronger the acid), SMILES for acid and the conjugated base (used for detecting acids and bases in molecule), and SMIRKS for acid2base and base2acid transformations (used in standardization module “Apply_CVSP_AcidBase_SMIRKS”, see below).

Standardization Rules

Standardization rule are allowing users to put together a standardization workflow that would apply rules to all the records in the submission. **Please note**: you should use the CVSP default standardization rules carefully as it does not apply tautomer canonicalization.

```
<?xml version="1.0" encoding="utf-8" ?>

<rules>

        <rule category="CVSP" type="module" value="Dearomatize"></rule>

        <rule category="INCHI" type="SMIRKS" value="[N-,P-,As-,Sb-,O-,S-,Se-,Te-:1][C:2]=[N+,P+,

                As+,Sb+,O+,S+,Se+,Te+:3]>>[N,P,As,Sb,O,S,Se,Te:1]=[C:2][N,P,As,Sb,O,S,Se,Te:3]" description="Example:
                (O-)-C=(O+) -> CO2H"/>

</rules>
```

30+ CVSP modules are available for users to include in their custom standardization rule sets:

Dearomatize, Aromatize, ConvertDoubleBondWithAttachedEitherSingleBondStereo2EitherDoubleBond, Layout, StandardizeHexagons, Disconnect_Metals_from_NonMetals, Disconnect_Metals_From_NOF, Ionize_Neutral_Alkaline_Metals_With_Carboxylic_Acids, Apply_CVSP_AcidBase_SMIRKS, Remove_Free_Metals, Standard_InChl_Normalization, CanonicalizeTautomers, Retain_Largest_Organic_Fragment, NeutralizeCharges, ConvertUpOrDownBondAdjacentToDoubleBondToNoStereoSingleBondAndCrossedDoubleBond, Remove_Water, TreatAmmonia, Remove_Neutral_Inorganic_Residue, Remove_Organic_Solvents, Remove_Gaseous_Molecules, Fold_Non_Stereo_Hydrogens, StripAmbiguousSp3Stereo, Fold_All_Hydrogens, Remove_SP3_Stereo, Reset_Symmetric_Stereo_Centers, Reset_Symmetric_CisTrans_Bonds, Convert_Double_Bond_to_Either, Convert_Either_Double_Bonds_To_Stereo, Remove_Allene_Stereo

To use any of the CVSP modules listed above user has to have the rule formatted as below (value has to exactly match the case-sensitive module name):

```
        <rule category="CVSP" type="module" value="Retain_Largest_Organic_Fragment"></rule>
```

To add a standardization rule in a form of SMIRKS it has to have the following:

```
        <rule category="YourCategory" type="SMIRKS" value="[N-,P-,As-,Sb-,O-,S-,Se-,Te-:1][C:2]=[N+,P+,

As+,Sb+,O+,S+,Se+,Te+:3]>>[N,P,As,Sb,O,S,Se,Te:1]=[C:2][N,P,As,Sb,O,S,Se,Te:3]" description="Example: (O-)-C=(O+) ->
CO2H"/>
```

CVSP will run standardization exactly in the order that the rules are listed in the XML file.

Submitting Feedback

CVSP is being provided as a service for the chemical community, and we are always seeking to improve it and make it more useful for you. We are always interested in hearing feedback about the site, and are happy to answer any questions you might have.

To submit feedback about the site as a whole, click the “Provide Feedback” button in the upper right corner of the page. In addition to user’s feedback text the feedback form additionally submits user name, browser version, and the URL from which the form has been initiated (thus capturing CVSP deposition identifier and issue filters).