



How to add your reactions to generate a Chemical Space

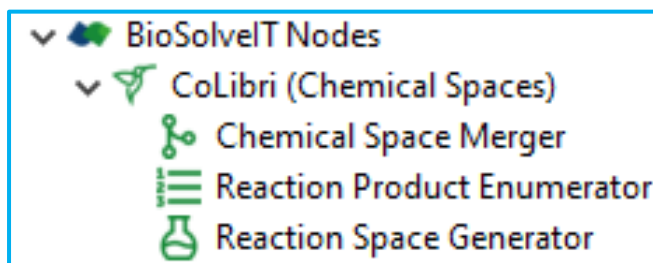
based on BioSolveIT's KNIME nodes

Introduction to CoLibri™



This tutorial is supposed to show how “normal” drawings of reactions can be easily edited to yield precise reaction definitions that can be handled by BioSolveIT’s CoLibri tools.

- ◆ In the first section we will show you the step-by-step addition and processing of a reaction using an esterification as very easy example. This part includes as well the formal parts regarding the usage and configuration of the KNIME nodes.

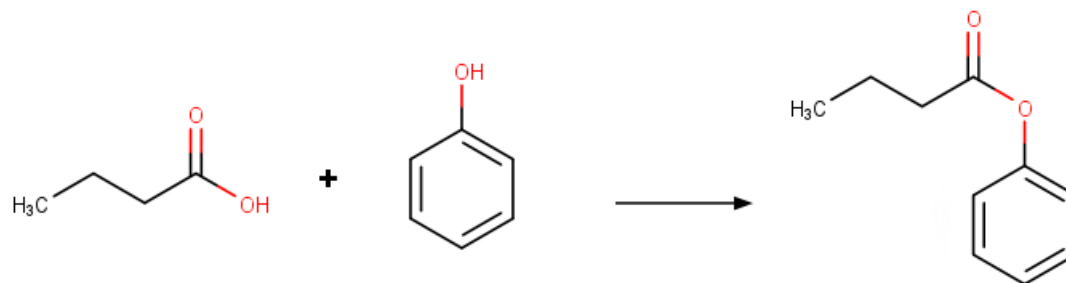


- ◆ In the second part four examples are shown which are real parts of BioSolveIT’s KnowledgeSpace. These shall show you more cases of the reaction adoption to the tools.
- ◆ The 3rd part deals with some tips and tricks to facilitate debugging or prettify the results.



Drawing of Reactions

In the very first step we define a reaction for CoLibri in the KNIME system. A simple esterification, where a carboxylic acid and an aromatic alcohol form an ester will be used as showcase:



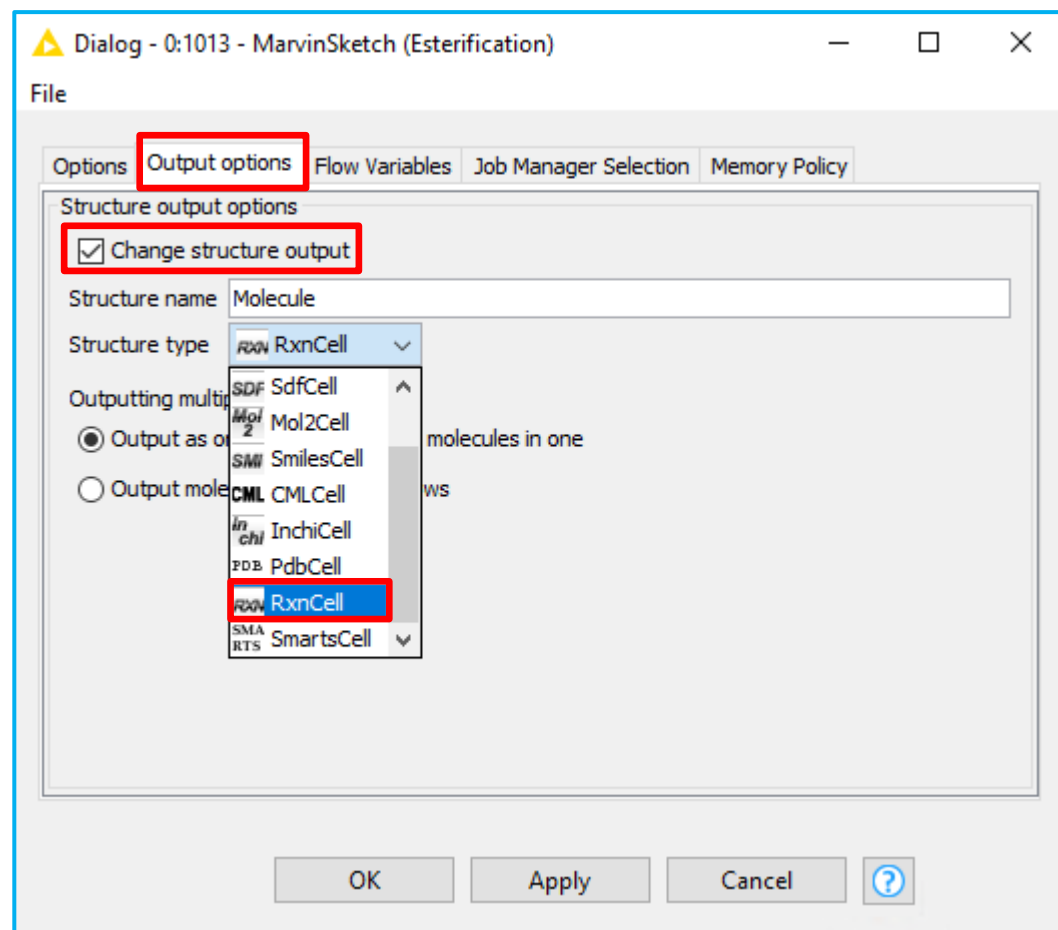
We recommend to use a recent update of the Marvin Sketch node to draw the reaction.



Set up Marvin Sketch

By default, Marvin uses its own format, but CoLibri requires the reactions in rxn format:

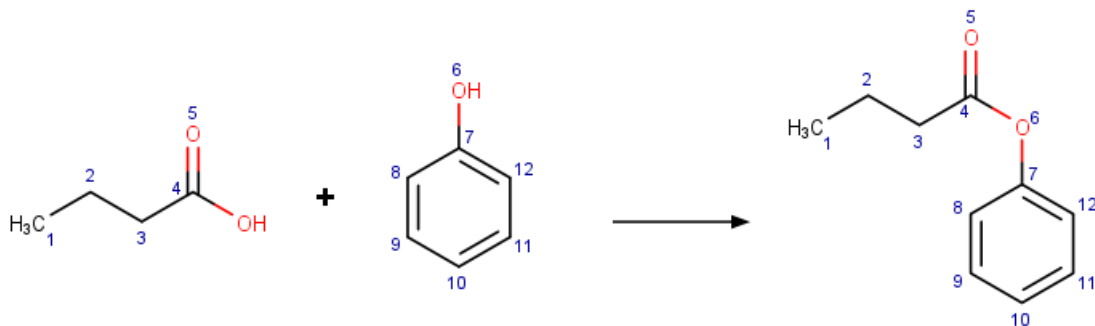
- ◆ Open the node configuration
- ◆ Go to the “Output option” tab
- ◆ Check “Change structure output”
- ◆ Choose “RxnCell” from dropdown



Mapping of Atoms

CoLibri needs a correspondence between the atoms on the educt side and the related atoms in the product molecule. These “mapping atoms” need to be defined within the Marvin Sketch node:

- ◆ Right-click on an atom
- ◆ Select “Map” and a number
- ◆ Right-click on corresponding atom in the product
- ◆ Select “Map” and the same number

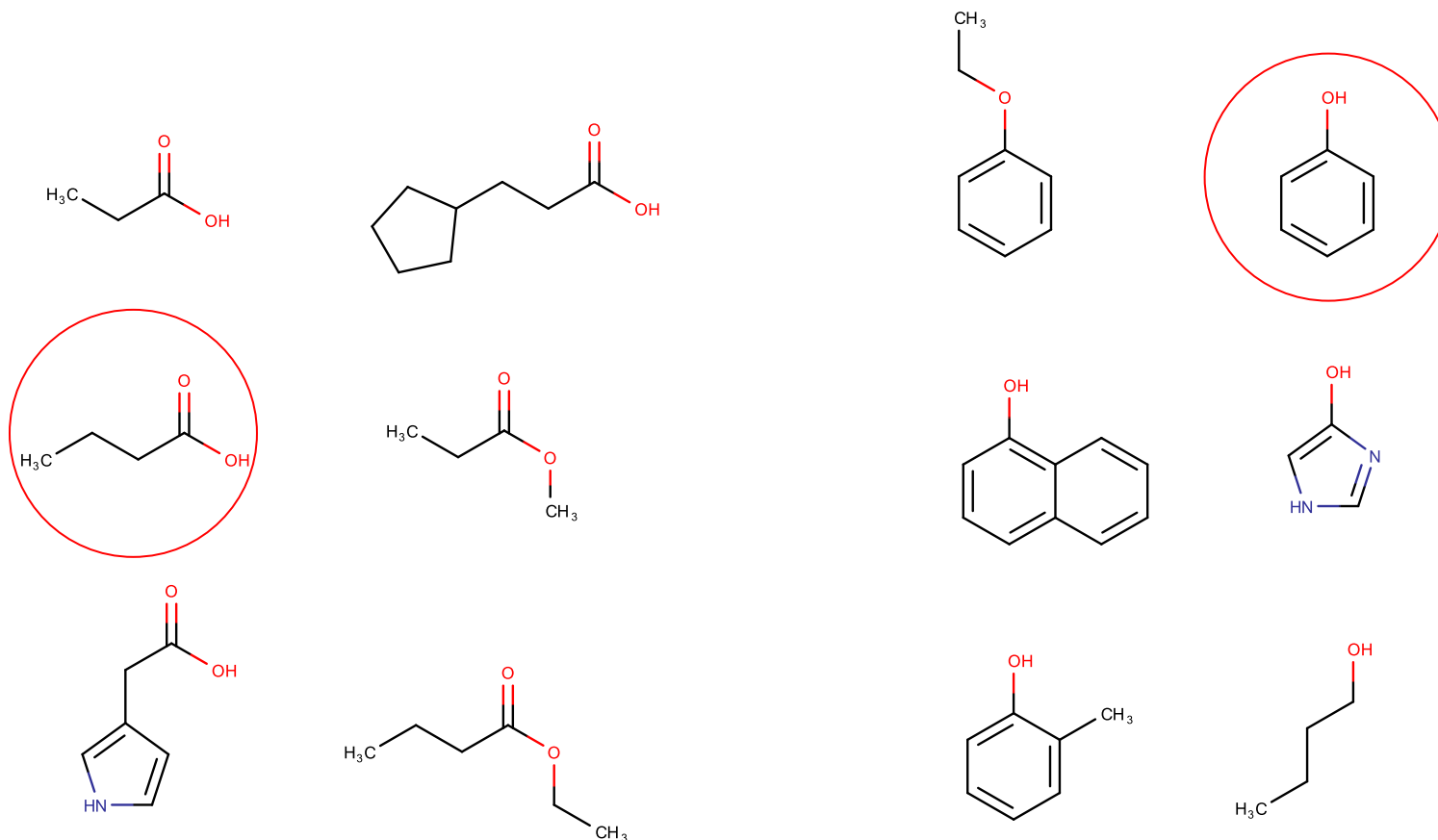


- ◆ Add mapping numbers only to those atoms that occur on both sides. Atoms without mapping number on the reagent side will be removed (e.g. leaving groups)



Test set for Esterification

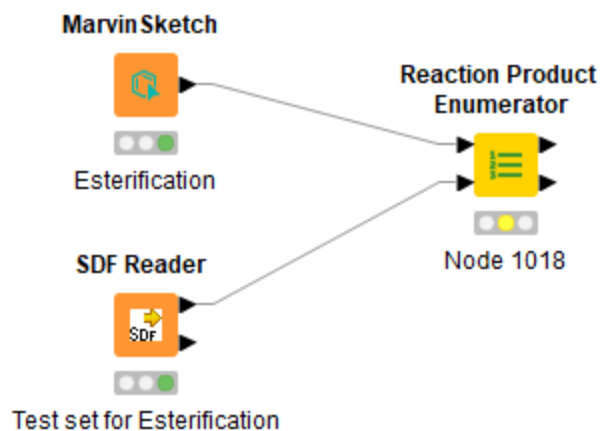
- ◆ A test set including matching and non-matching molecules.
- ◆ Exact matches on our reaction definition are highlighted in red.



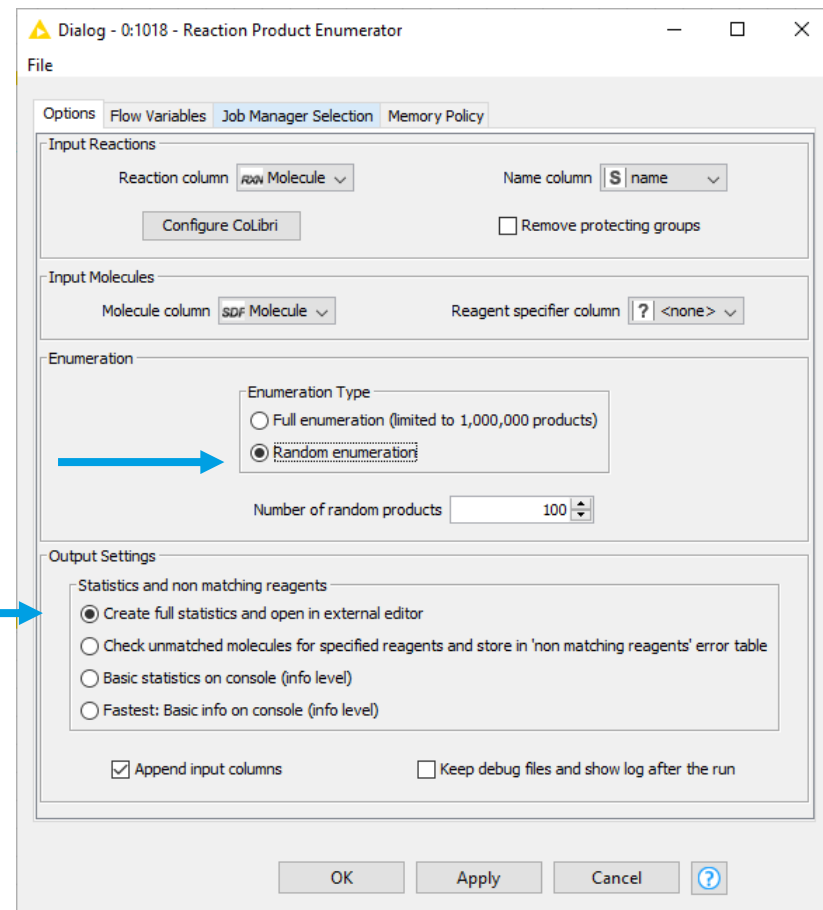
Run Reactions with CoLibri

We start with a minimum workflow including CoLibri's "Reaction Product Enumerator" node. This node generates some products and statistics that helps debugging when setting up new reactions.

- ◆ Set up a workflow like this:



- ◆ Configure "Reaction Product Enumerator":



Start Debugging

Unfortunately the “Reaction Product Enumerator” fails:
Check out the stats file:

Reaction details:

Name: esterification

Number of reagents: 2

Reagent 1 smarts pattern (1): [\$([#8]):5]=[\$([#6]):4](-[\$([#6]):3]-[\$([#6]):2]-[\$([#6]):1])-\$([#8])]

Reagent 2 smarts pattern (1): [\$([#6]):7]=1(-[\$([#6]):8]=[\$([#6]):9]-[\$([#6]):10]=[\$([#6]):11]-[\$([#6]):12]1)-[\$([#8]):6]

Number of new cores: 0

Product smarts pattern (1): [O:6](-[c:7]-1=[c:12]-[c:11]=[c:10]-[c:9]=[c:8]1)-[C:4](-[C:3]-[C:2]-[C:1])=[O:5]

Matching details:

Number of matches for reagent 1: 3

Molecule: O=C(O)CCC Row1

Molecule: O=C(O)CCC1CCCC1 Row2

Molecule: O=C(OCC)CCC Row4

Number of non matches for reagent 1: 9

Number of matches for reagent 2: 0

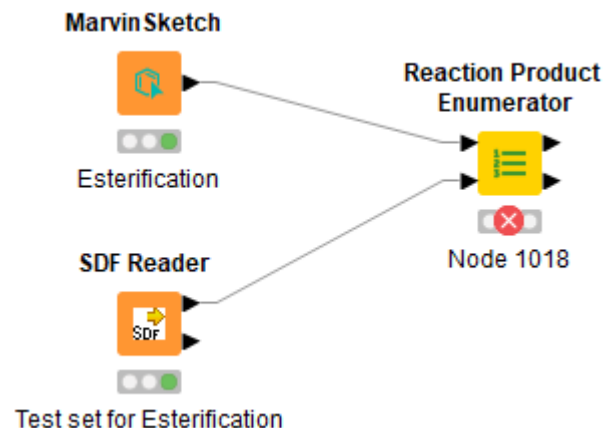
Number of non matches for reagent 2: 12

Fragments statistic:

Fragments for reagent 1: 2

Fragments for reagent 2: 0

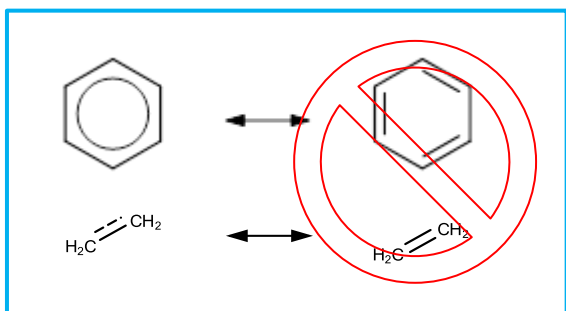
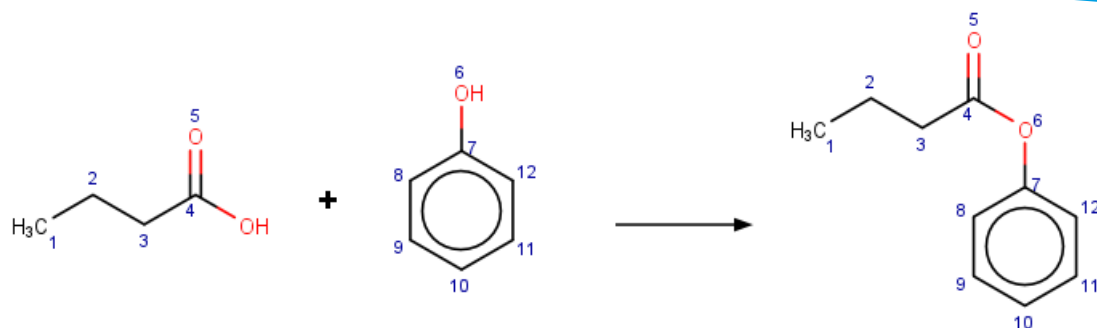
=> No matches on reagent 2! This seems strange, because we have drawn the matching reagent. But it is a question of the details in the drawing...



Define Aromaticity in Educt and Product

When using the RxnCell format in Marvin Sketch, it processes compounds with alternating bonds exactly like they were sketched. The sd-file reagent input is interpreted to have aromatic bonds. Our reaction definition need to be adjusted:

- ◆ Right-click on an aromatic bond
- ◆ Select "Type" => "Aromatic"



Don't draw alternating single and double bonds!

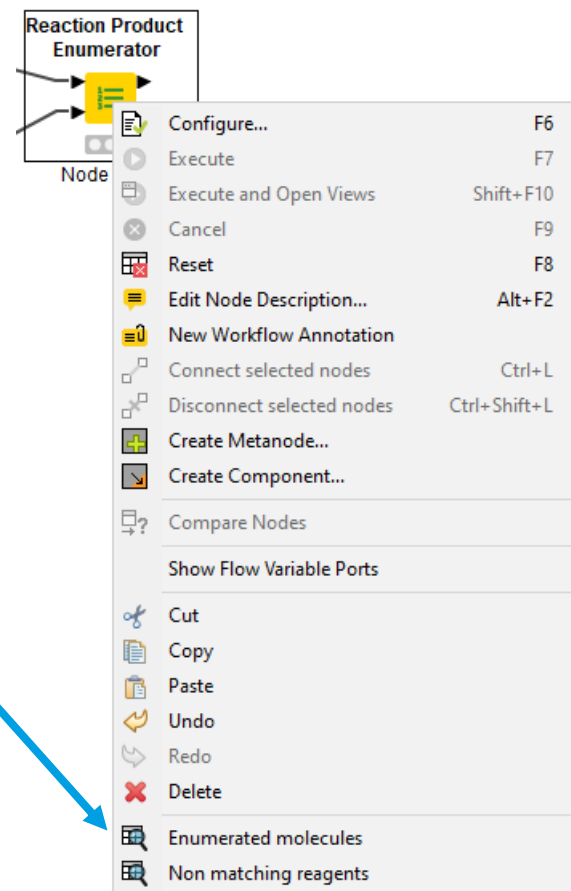
The screenshot shows the 'Type' menu in Marvin Sketch. The 'Aromatic' option is highlighted with a blue arrow pointing to it from the text above. The menu includes options for bond types (Single, Double, Triple, Aromatic), Topology, Reacting Center, Stereo Search, Arrange, Align, Delete, Group..., Add, Transformation, Format..., and Edit Properties...



Rerun "Reaction Product Enumerator"

Rerunning the node works! But we get 8 products instead of the expected 1. So we have many more matches than anticipated. Let's take a look at the enumerated products:

- ◆ Right-click on the "Reaction Product Enumerator"
- ◆ Choose "Enumerated molecules"



Matching details:

Number of matches for reagent 1: 3

Molecule: O=C(O)CCC Row1

Molecule: O=C(O)CCC1CCCC1 Row2

Molecule: O=C(OCC)CCC Row4

Number of non matches for reagent 1: 9

Number of matches for reagent 2: 4

Molecule: O(c1cccc1)CC Row6

Molecule: Oc1c2c(ccc1)cccc2 Row7

Molecule: Oc1c(cccc1)C Row8

Molecule: Oc1cccc1 Row9

Number of non matches for reagent 2: 8

Fragments statistic:

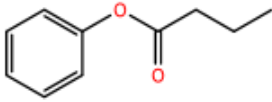
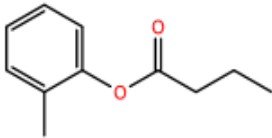
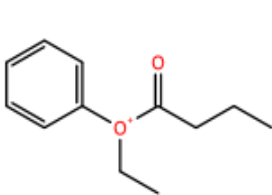
Fragments for reagent 1: 2

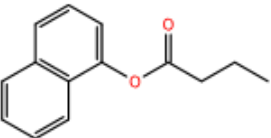
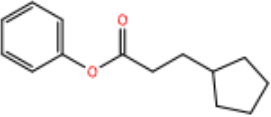
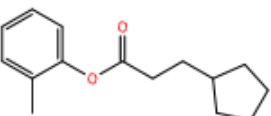
Fragments for reagent 2: 4

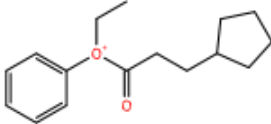
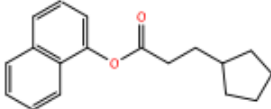
Max number of enumerated products: 8



Look out for trash molecules

Row ID	SDF Enumerated molecule
Row0	
Row1	
Row2	

Row ID	SDF Enumerated molecule
Row3	
Row4	
Row5	

Row ID	SDF Enumerated molecule
Row6	
Row7	

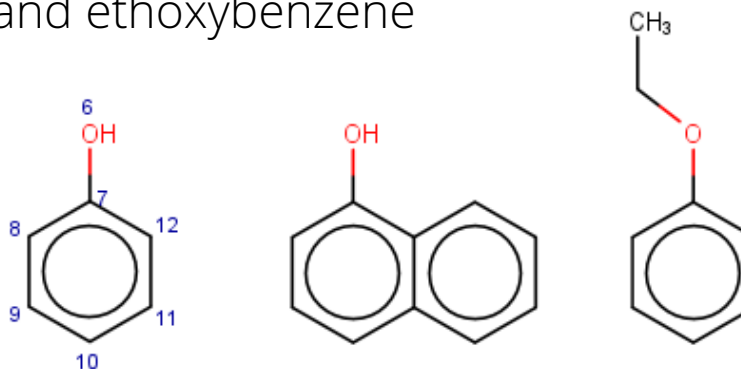
=> 2 trash products were generated



Define your substructures as precise as possible

From this set of products we can learn the most important thing for the correct definition of reactions: your drawing will be interpreted as a substructure!

- ◆ Phenol is the substructure of naphthol and ethoxybenzene



- ◆ Note that hydrogens are ignored!
They are just added by the drawing program by default
- ◆ Be as specific as possible and define the environment of an atom!
E.g. annotate, if something is supposed to be terminal. Even if the naphthyl could be accepted for this reaction, the ether is definitely not!



Define the Environment of atoms

The classical way to define the environment is using SMARTS. You find them in Marvin Sketch here:

- ◆ Click on the “periodic table” (1.)
- ◆ Go to “Advanced” (2.)
- ◆ Choose SMARTS (3.)
- ◆ Type in your definition (4.)
e.g. [O;D1] to define an alcohol

- ◆ Close and click on the atom we want to edit

Dialog - 0:1013 - MarvinSketch (Esterification)

File

Options

File Edit

Periodic Table of Chemical Elements

Periodic Table **Advanced** 2.

Description

Generic query atoms: A Q M X AH QH MH XH

Atom query properties: .H+ .v+ .X+ .R+ .r+ .rb+ .s+ .h+ .D+ .u .H- .v- .X- .R- .r- .rb- .s- .h- .D- .a/A

Periodic Table Groups: G1 G2 G3 G4 G5 G6 G7 G8 G9 G10 G11 G12 G13 G14 G15 G16 G17 G18

R-groups: R1 R2 R3 R4 R5 R6 R7 R8 R9 R10 R11 R12 R13 R14 R15 R16 R17 R18 R19 R20 R21 R22 R23 R24 R25 R26 R27 R28 R29 R30 R31 R32

Special nodes: Pol * Alkyl

Custom Property

Type: R-group Alias Pseudo **SMARTS** 3. Value

Value **[O;D1]** 4.

Close

OK Apply Cancel ?

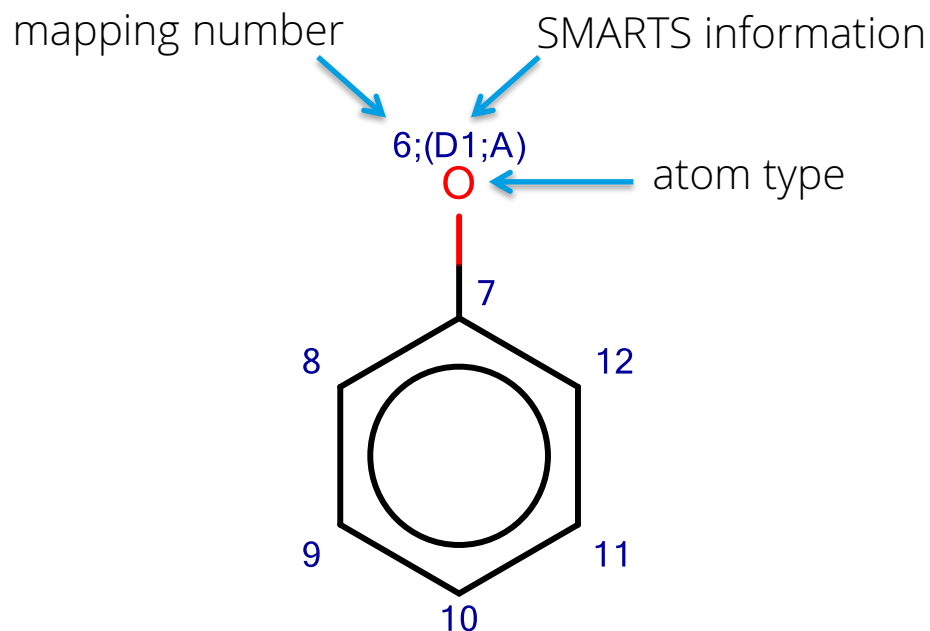
- ◆ read on to hear more about SMARTS



Phenol with defined Environment

The code `[O;D1:6]` defines an oxygen with only one heavy atom as neighbor (D1) and a mapping number of 6.

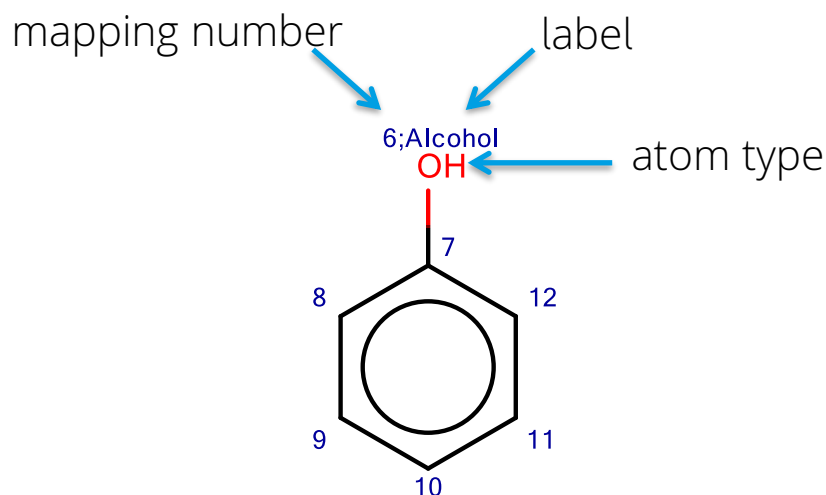
It will look like this:



SMARTS vs Labels

Alternatively to SMARTS you can use “labels”, which are simply pre-defined SMARTS-pattern that make the drawing easily human-readable. You find this option again in the advanced editor mode.

The molecule will look like this:

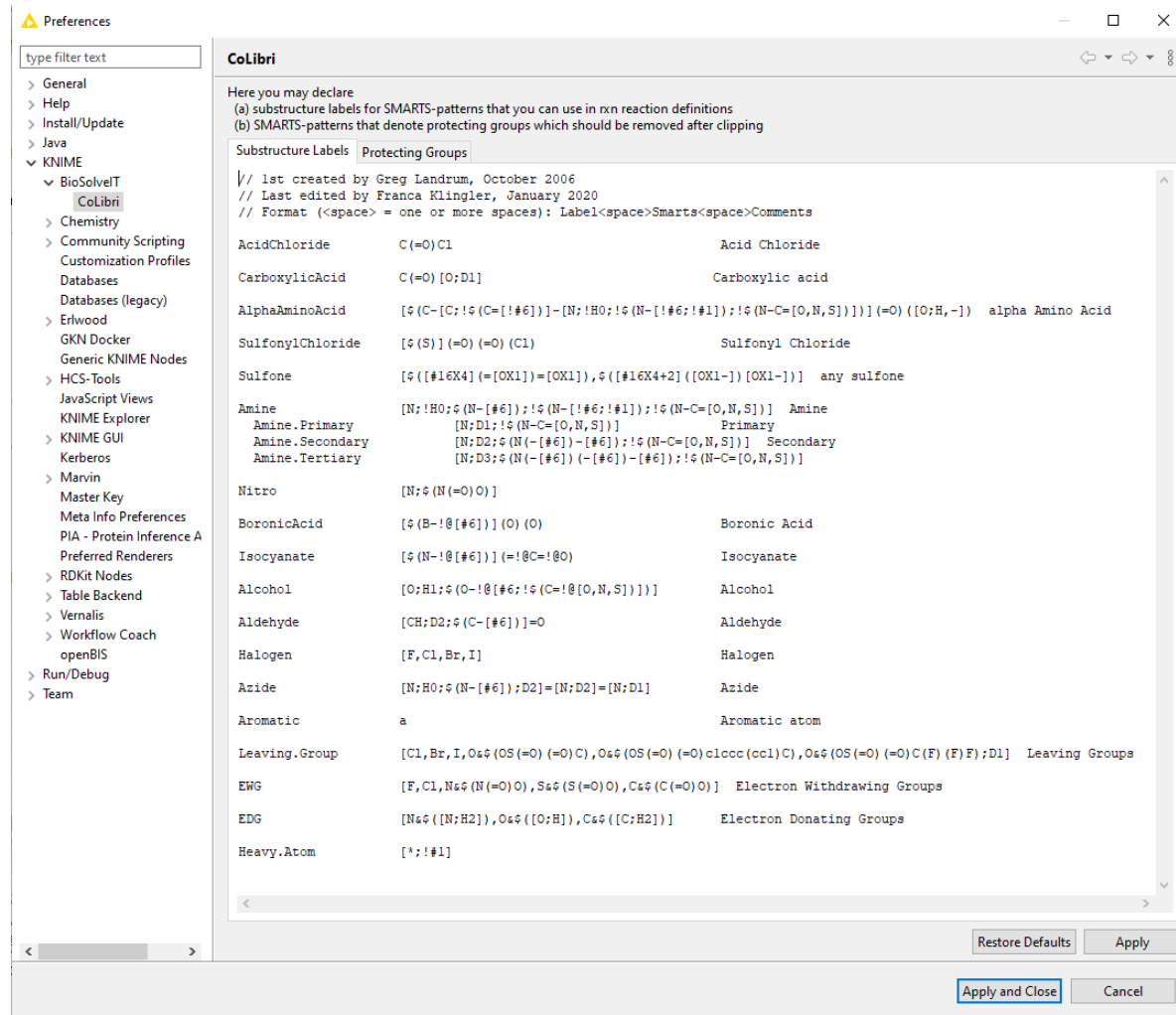
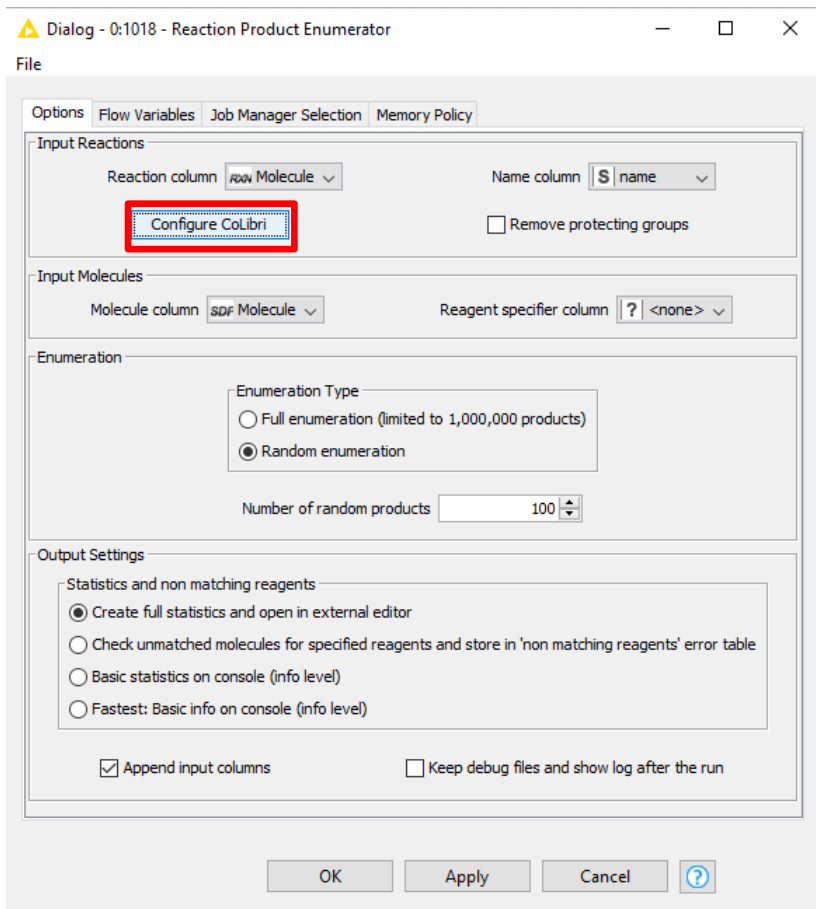


The screenshot shows a dialog box titled 'Periodic Table of Chemical Elements'. The 'Advanced' tab is selected and highlighted with a red box. The 'Description' field is empty. Below it are sections for 'Generic query atoms', 'Atom query properties', 'Periodic Table Groups', 'Special nodes', 'R-groups', and 'Custom Property'. The 'Custom Property' section has 'Type:' set to 'Value' (highlighted with a red box) and 'Value:' set to 'Alcohol' (also highlighted with a red box). A 'Close' button is at the bottom right.

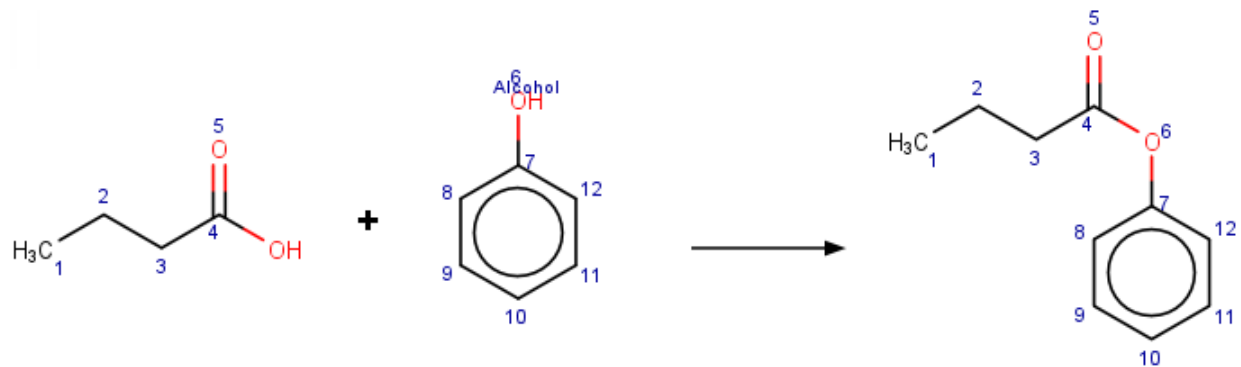


Where to find pre-defined Labels

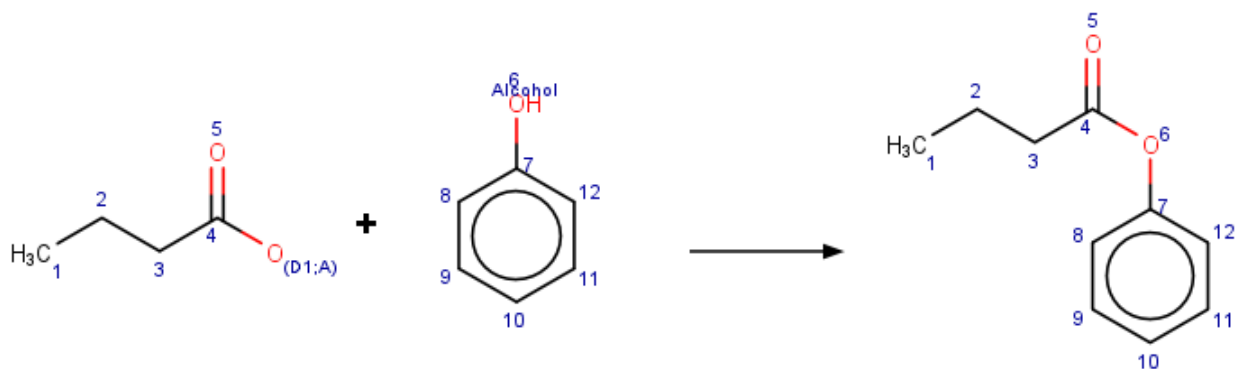
All pre-defined labels can be found and edited in the Enumerator.



Update Esterification Reaction



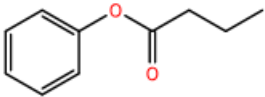
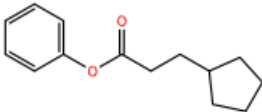
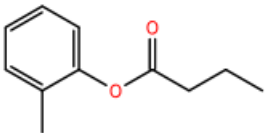
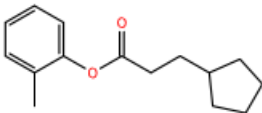
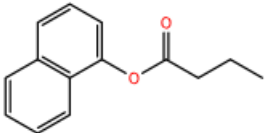
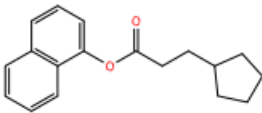
The aromatic alcohol is precisely defined but the hydroxyl group of the carboxylic acid is not. The label CarboxylicAcid exists but as we cut off a bond inside this group it is better to use the SMARTS `[O;D1]`:



Rerun “Reaction Product Enumerator”

The list of enumerated molecules does not contain trash molecules with senseless valences at the oxygen anymore.

But still other building blocks than phenol match the aromatic ring (e.g. naphthol and cresol). Let's look at the input and decide which building blocks to allow for the individual reagents.

Row ID	SDF Enumerated molecule	Row ID	SDF Enumerated molecule
Row0		Row3	
Row1		Row4	
Row2		Row5	

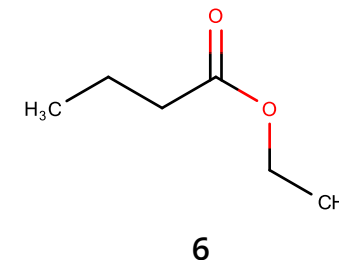
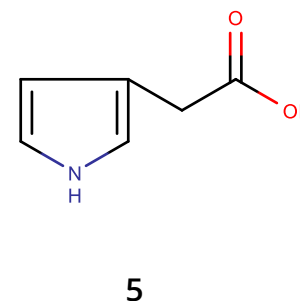
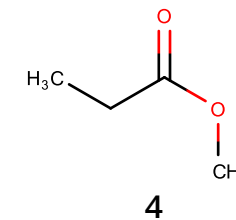
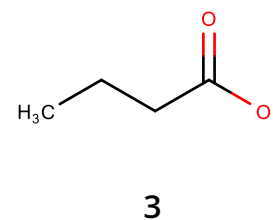
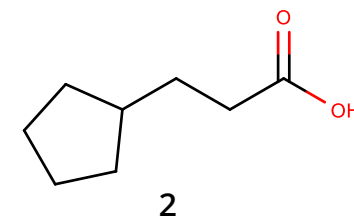
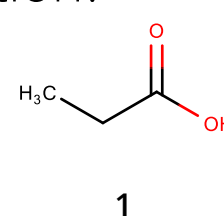
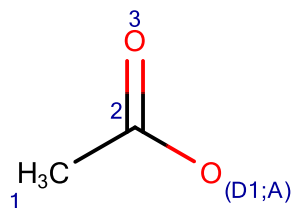


Building Block 1

From a chemists point of view free carboxylic acids could work for reagent 1. Which reagents match/ do not match on our definition?

- ◆ Reagent 2 & 3 match
- ◆ Reagent 1 & 5 does not match
- ◆ Reagent 4 & 6 does not match

A butyric acid is drawn which is a substructure only of reagents 2 & 3. We need to reduce our definition to the minimum substructure, which is acetic acid:



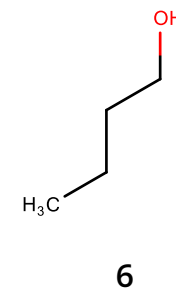
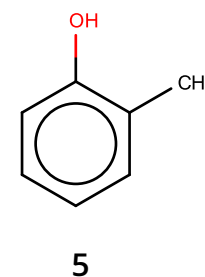
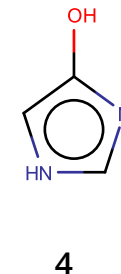
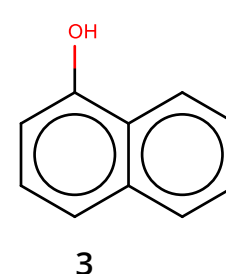
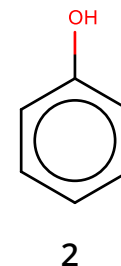
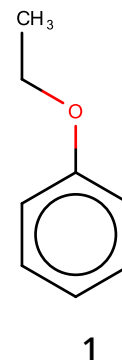
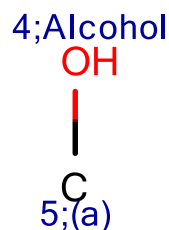
Building Block 2

Decide which of the building blocks should run the reaction. If we go for aromatic alcohol only 2 – 5 are desired but not 1 & 6.

Which reagents match/ do not match on our definition?

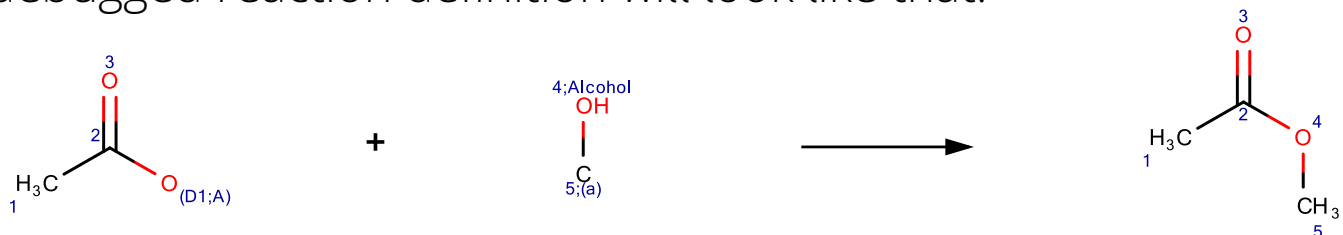
- ◆ Reagent 2, 3 & 5 match
- ◆ Reagent 1, 4 & 6 does not match

A phenol ring is drawn which is a substructure of reagents 2, 3 and 5 but doesn't match on 4. Again we need to reduce our definition to the minimum substructure. This can be done by using the SMARTS [c] (aromatic carbon marked as 'a') and the label 'Alcohol':



Debugged Reaction Definition

The debugged reaction definition will look like that:



Rerunning the “Reaction Product Enumerator” will lead to 4 matching building blocks for every reagent and 16 different products:

Row ID	SDF Enumerated molecule
Row0	
Row1	
Row2	

Row ID	SDF Enumerated molecule
Row3	
Row4	
Row5	

Row ID	SDF Enumerated molecule
Row6	
Row7	
Row8	

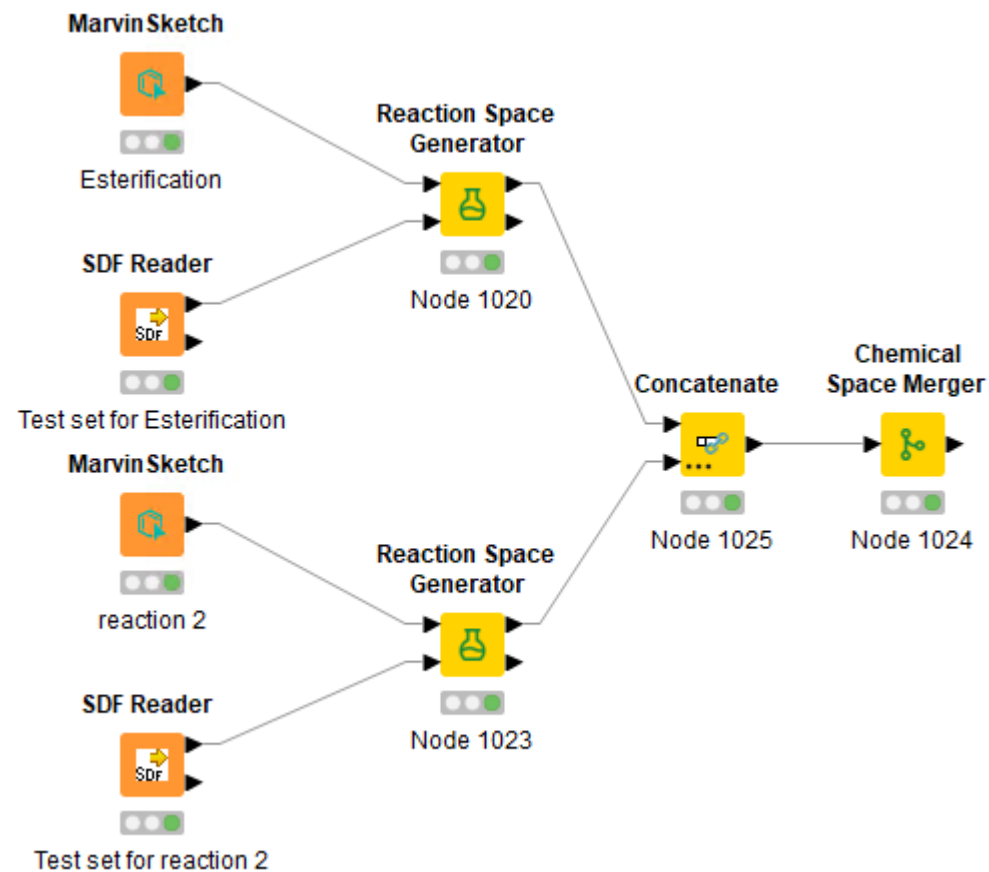
Row ID	SDF Enumerated molecule
Row9	
Row10	
Row11	

Row ID	SDF Enumerated molecule
Row12	
Row13	
Row14	
Row15	



Build a Fragment Space

Now that we have debugged the reaction setup we like to build a fragment space. Therefore, replace the “Reaction Product Enumerator” by the “Reaction Space Generator”. Combine several spaces using the “Concatenate” node followed by a “Chemical Space Merger” node to generate a searchable space.



take a look into our BioSphere workflow



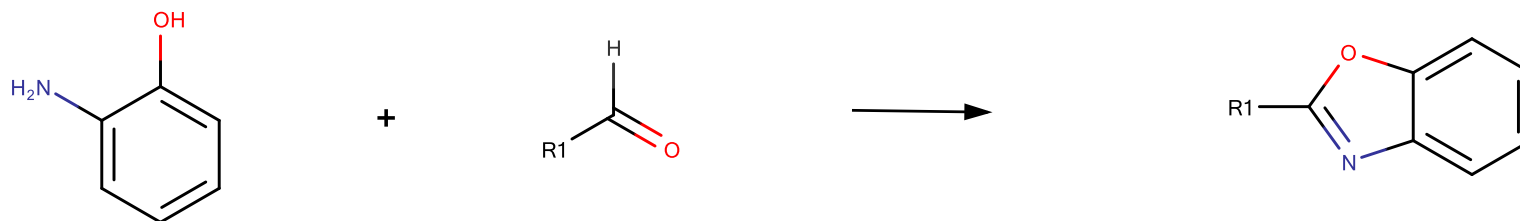
Part 2: Example Reaction from KnowledgeSpace

- ◆ Example 1: Ringclosing – Benzoxazole formation
- ◆ Example 2: Multicomponent – Scicinski reaction
- ◆ Example 3: Substitution – hetero aromatic substitution
- ◆ Example 4: Coupling – Di Mauro reaction

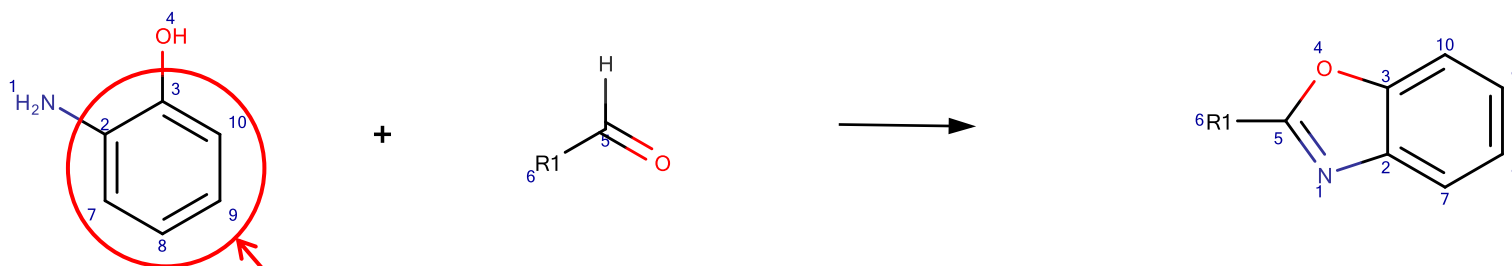


Example 1 - Ringclosing

As first example we use something similar to a benzoxazole ring closure, where an aminophenol reacts with an aldehyde:



Let's make it suitable for CoLibri and add the mapping atoms:



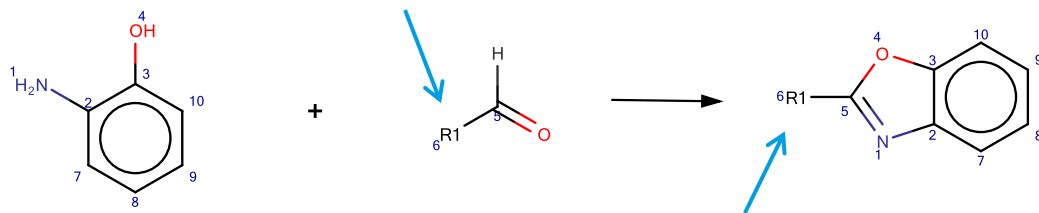
Please use the aromatic form on educt and product side. Don't draw alternating single and double bonds!



Example 1 – R Groups

CoLibri does not understand R groups. Replace it with the atom that should be in this position. Think in substructures! In this example R1 can be carbon or nitrogen_

- ◆ open the Advanced settings
- ◆ click on SMARTS (1)
- ◆ type [C,N] in the “value” field (2)
- ◆ Close and replace R1 on educt side



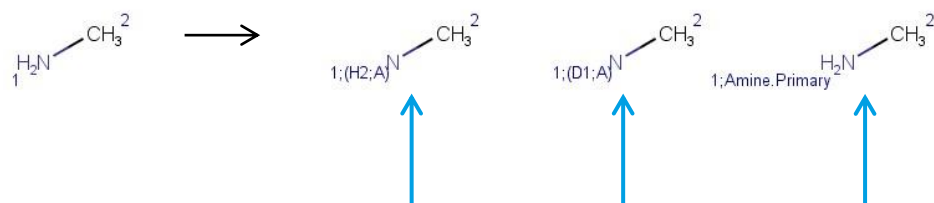
- ◆ Open Advanced settings
- ◆ Choose Generic query atom “A” (3)
- ◆ Close and replace R1 on product side

The screenshot shows the 'Periodic Table of Chemical Elements' dialog box with the 'Advanced' tab selected. The 'Description' field is empty. Under 'Generic query atoms', the 'A' button is highlighted with a red box and labeled '3.'. Under 'Atom query properties', the 'SMARTS' button is highlighted with a red box and labeled '1.'. The 'Value' field is set to '[C,N]' and is highlighted with a red box and labeled '2.'. A 'Close' button is visible at the bottom right.



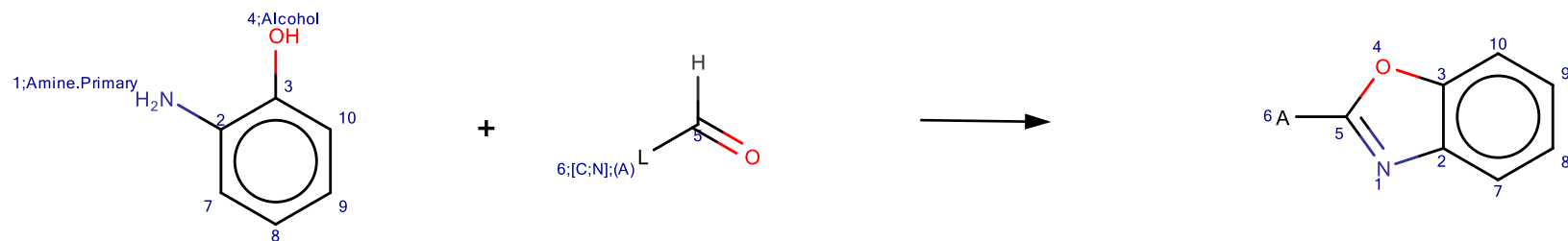
Example 1 – Functional Groups

The amine definition need to be more precise. Remember: Hydrogens are ignored! You can either use the SMARTS or a label:



These three definitions mean similar things, but the label 'Amine.Primary' contains the most precise definition, i.e. $[\text{N};\text{D}1;! \$(\text{N}-\text{C}=[\text{O},\text{N},\text{S}])]$

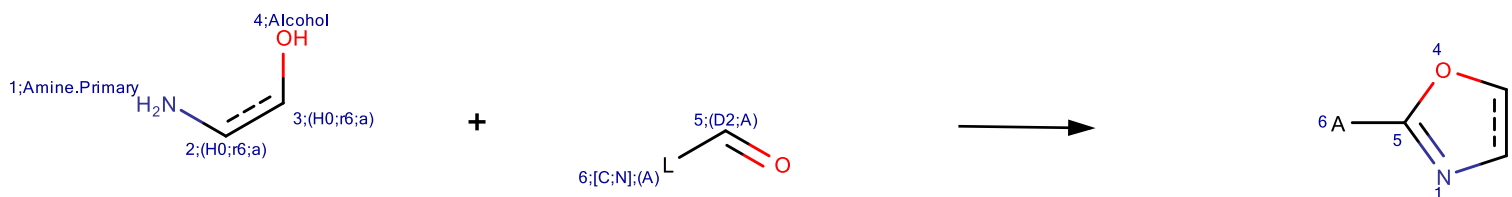
The oxygen need to be defined as alcohol (and not the substructure of an ether). Again, the easiest way is to use the label 'Alcohol' leading to:



Example 1 – Reduce to Essential Minimum

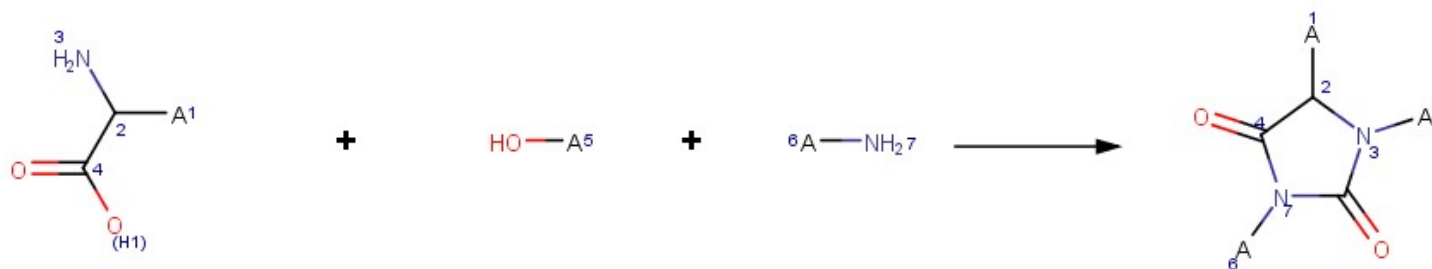
In the last step, we remove any needless atoms to define the reaction. Keeping the substructure to the essential minimum reduces the risk of errors during the subsequent processing.

- ◆ The aromatic ring can be reduced to the connecting carbons with an implicit SMARTS definition: `[cH0;r6]` (which means an aromatic carbon without hydrogens in a six-membered ring).
- ◆ Avoid explicit hydrogens! Therefore, atom 5 should be edited to `[CD2]` (carbon connected to two heavy atoms only) containing the hydrogen in an implicit definition. Note that `[CH]` and `[CD2]` are in principle the same. Definition of heavy atoms is easier applicable to more cases.

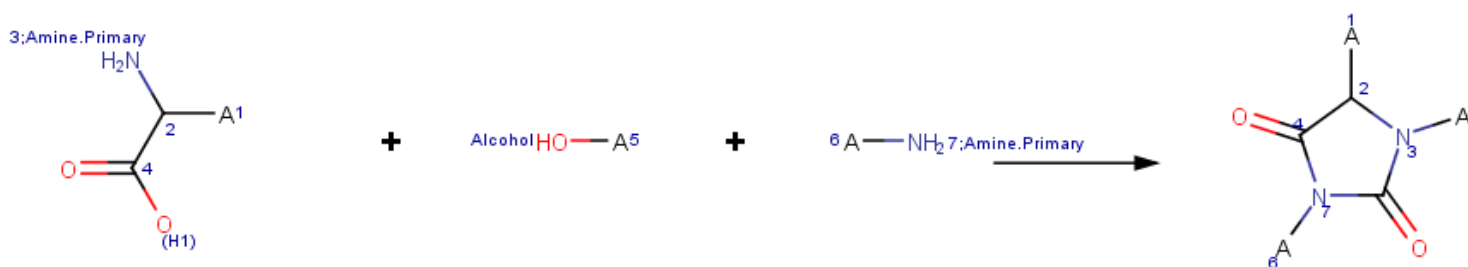


Example 2 – Multicomponent

In this reaction, an amino acid reacts with an alcohol and an amine to form a heterocycle. Reagents were drawn and atom mappings added to all atoms occurring on both sides of the reaction arrow:

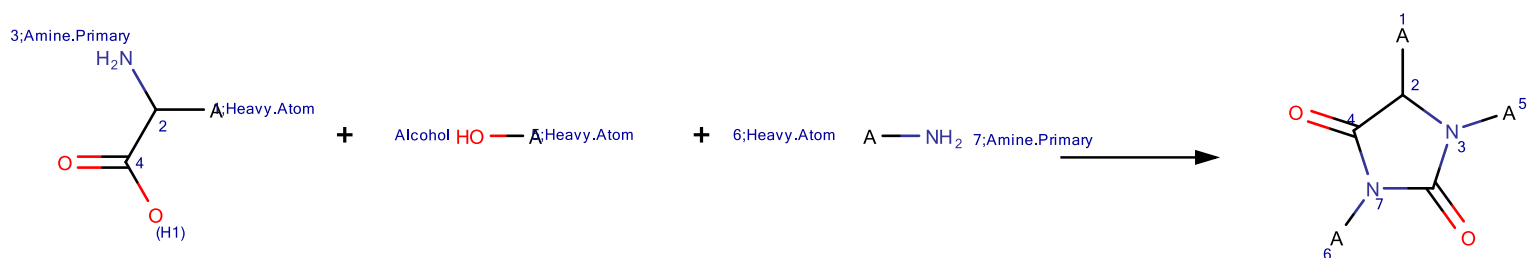


First, we define N3 and N7 as primary amines and the oxygen in reagent 2 as an alcohol:



Example 2 – Any Atom

Atom 1 should be any heavy atom. The "*" in a SMARTS covers also a hydrogen and need to be adjusted to [*;!#1]. We recommend to use the label "Heavy.Atom", which leads to a suitable reaction definition. The same expression is used for atoms 5 and 7:



Example 3 – Substitution

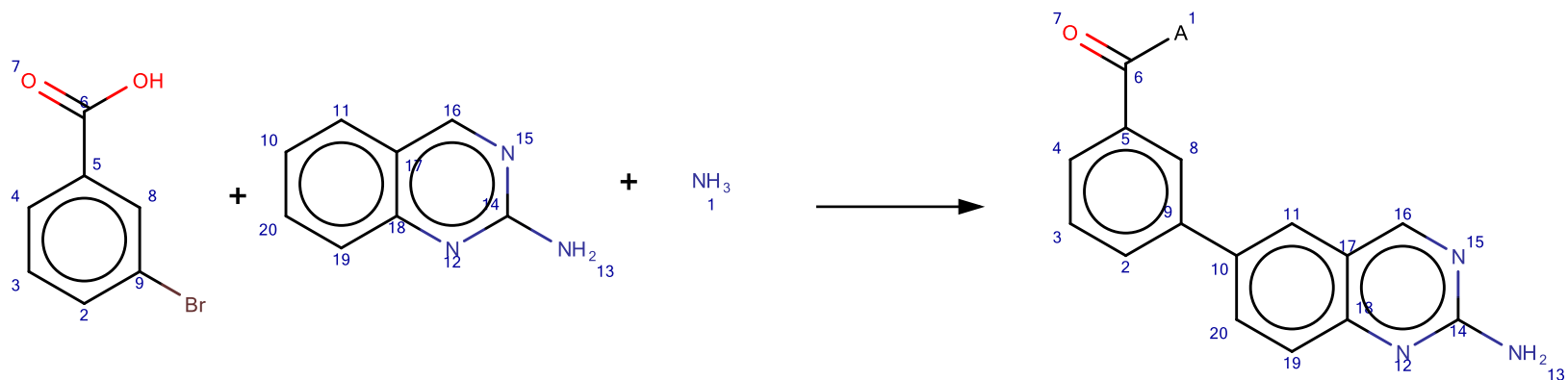
In a hetero aromatic substitution, different numbers of nitrogens can be in the aromatic ring. Thus, only one nitrogen is mandatory and in two more positions optional. For this we use lists by adding the SMARTS [c,n] (aromatic carbon or aromatic nitrogen). On the product side the corresponding positions are defined as "A".

Furthermore an amine (primary or secondary) is needed as second reagent, which can be defined with label "Amine".



Example 4 – Coupling

This coupling reaction can also be described as a two-stage reaction. The core fragment takes place in a carbon-carbon coupling first, followed by an amidation. The reagents are drawn below and all atoms are already mapped:

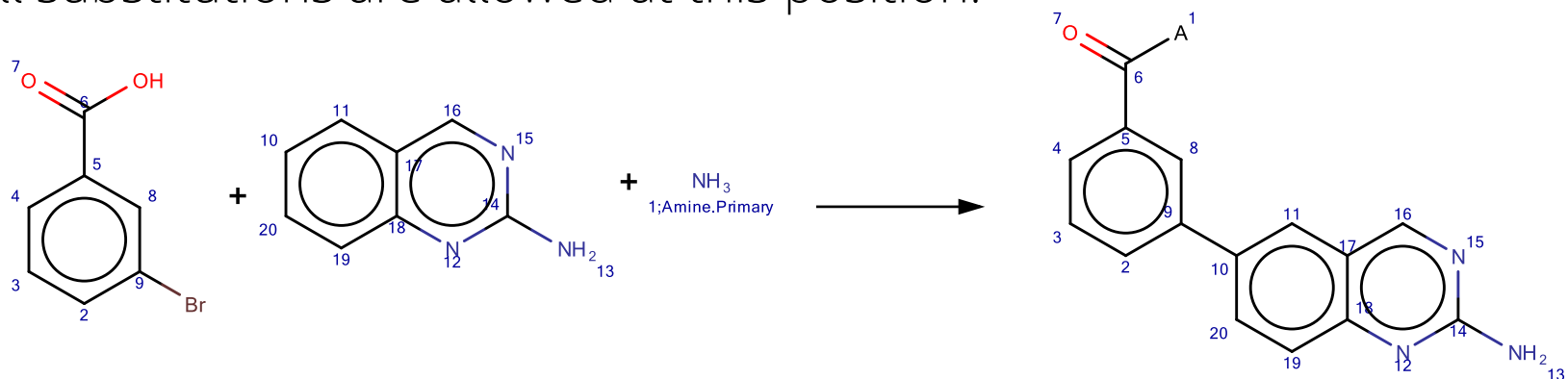


Think about how the reaction definition needs to be improved. There are 4 unprecise information.

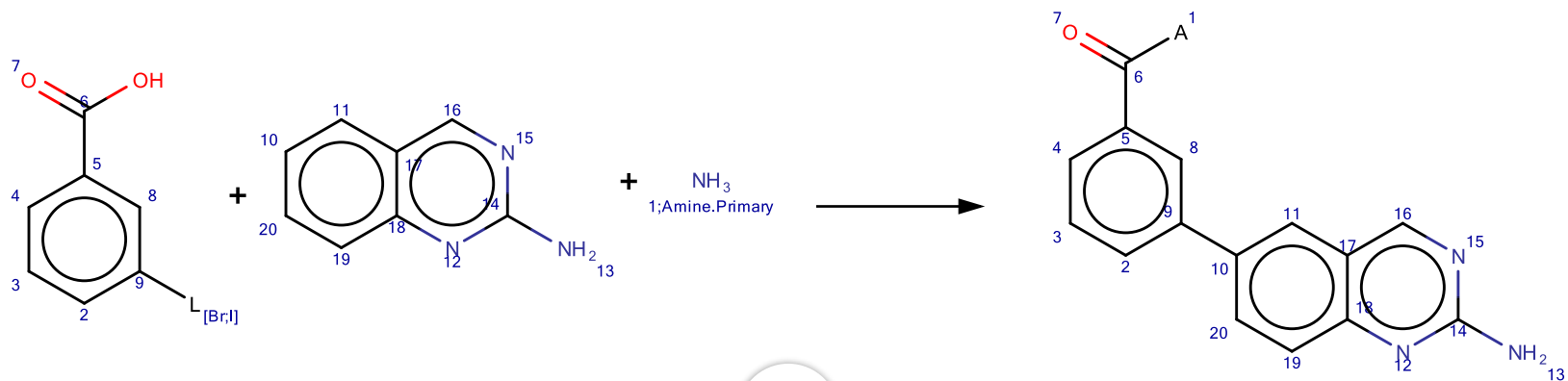


Example 4 - Leaving Groups

First, we add the Amine.Primary label to N1. Note that this label is not needed for N13. All substitutions are allowed at this position.

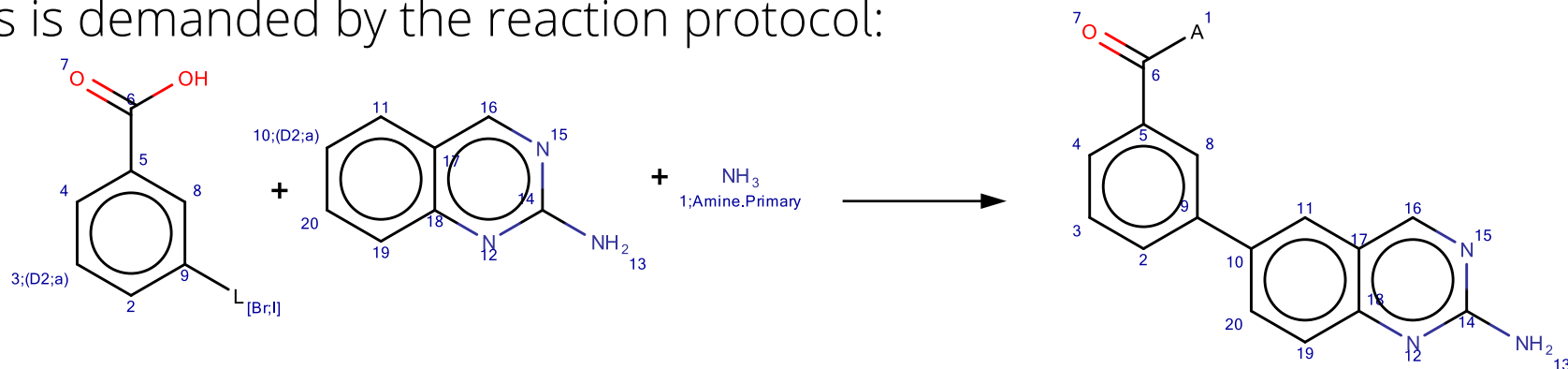


Second, we modify the leaving group connected to C9. According to the reaction description, it can be either Br or I which is described by the SMARTS [Br,I]:

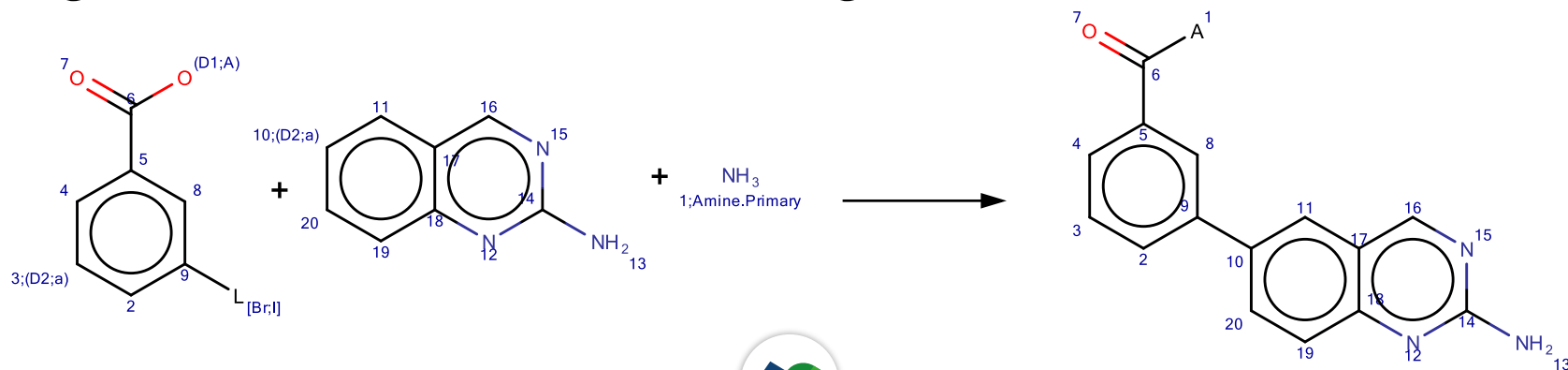


Example 4 – Unsubstituted Atoms

Third, c10 need to unsubstituted. The SMARTS `[c;D2]` ensures that the aromatic carbon is connected to two heavy atoms only. The same SMARTS was added to c3 as this is demanded by the reaction protocol:



Fourth, the carboxylic acid need to be defined to have a terminal hydroxy group, excluding the substructure of an ester using the SMARTS `[O;D1]`:



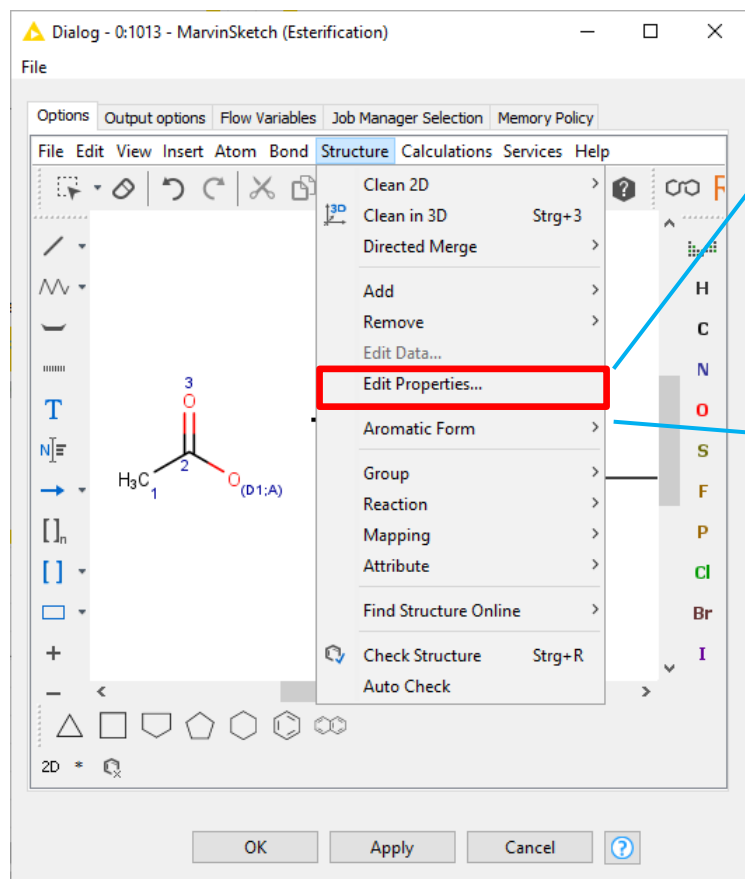
Part 3: Tips and Tricks

- ◆ Reaction names
- ◆ Compound ID Generation
- ◆ Reagent specifier
- ◆ Remove protecting groups

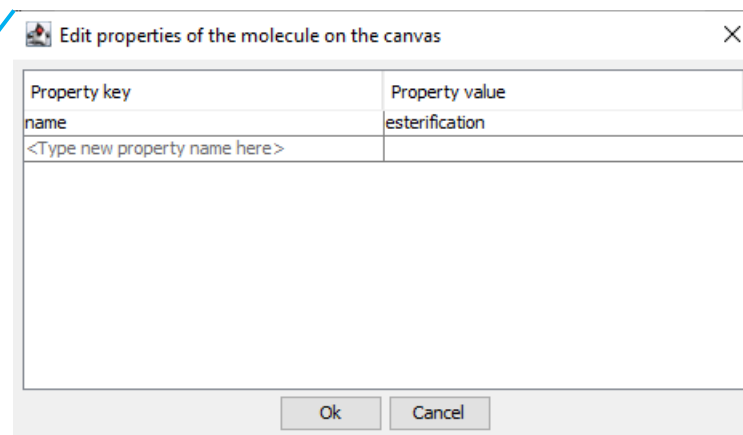


Reaction Name – Marvin Sketch

We recommend to pass the reaction names as properties. This enables CoLibri to generate a product ID including the reaction name. Choose Name column in the “Reaction Product Enumerator”/“Reaction Space Generator” nodes.

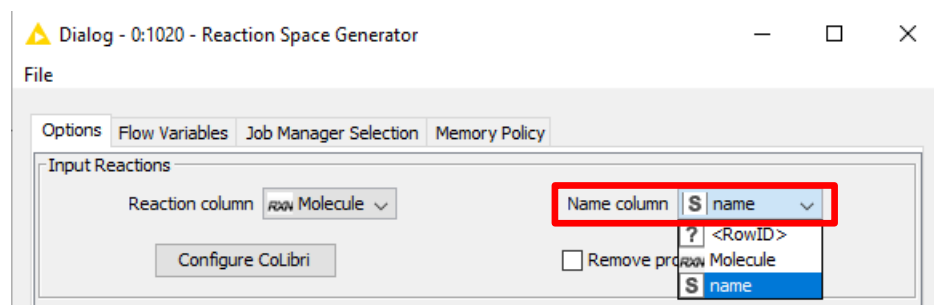


The screenshot shows the MarvinSketch application window titled "Dialog - 0:1013 - MarvinSketch (Esterification)". The "Structure" menu is open, and the "Edit Properties..." option is highlighted with a red rectangle. The main canvas displays a chemical structure of an ester with atoms numbered 1, 2, and 3. The "Edit Properties..." option is also highlighted with a red box in the "Edit properties of the molecule on the canvas" dialog box shown in the next image.



The dialog box titled "Edit properties of the molecule on the canvas" contains a table with two columns: "Property key" and "Property value". The first row shows "name" as the property key and "esterification" as the property value. Below the table is a text input field with the placeholder text "<Type new property name here>". The "Ok" and "Cancel" buttons are at the bottom.

Property key	Property value
name	esterification
<Type new property name here>	



The screenshot shows the "Reaction Space Generator" dialog box titled "Dialog - 0:1020 - Reaction Space Generator". The "Input Reactions" section has a "Name column" dropdown menu highlighted with a red rectangle. The dropdown menu is open, showing options: "S name", "? <RowID>", "Molecule", and "S name". The "S name" option is selected.



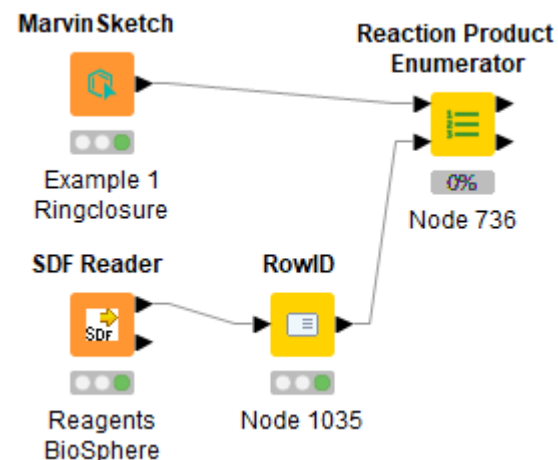
Compound ID Generation in CoLibri

Compounds of combinatorial libraries usually have an ID following this scheme:

Reaction-name_ID-building-block-1_ID-building-block2

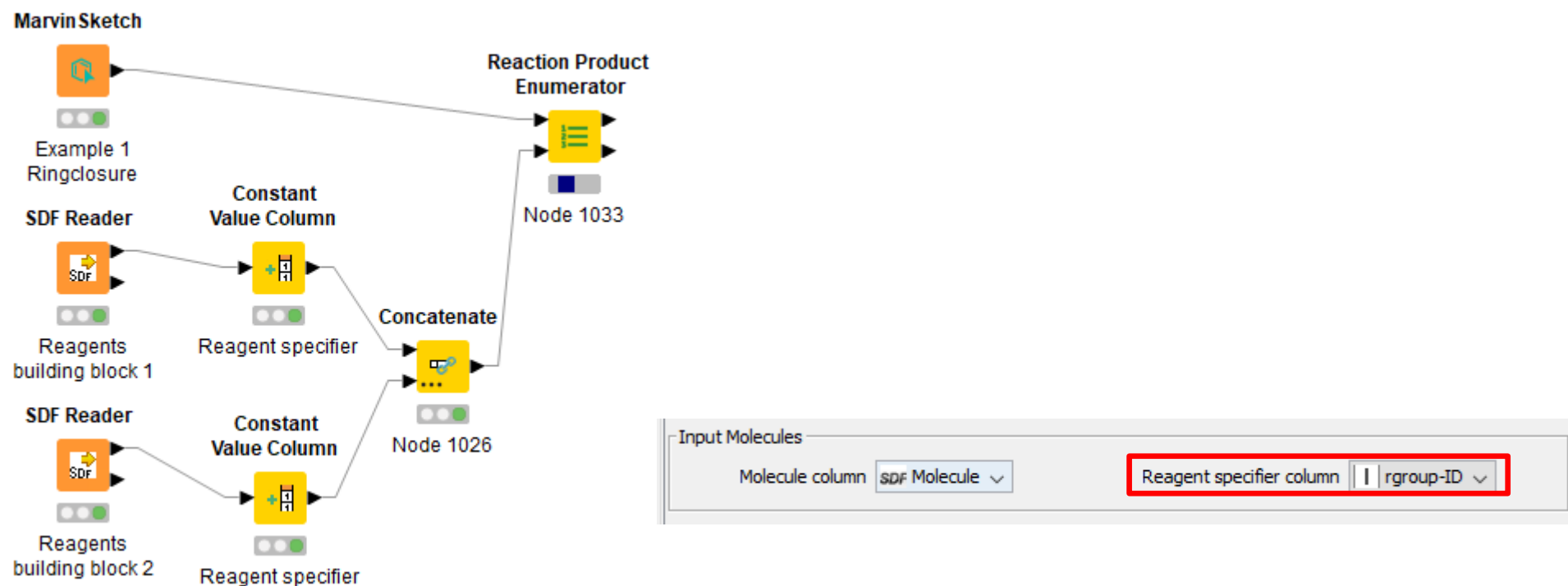
In order to generate this name within KNIME the compound ID need to be set as RowID. This can be done in 2 ways:

- ◆ SDF Reader: Enable “Use molecule name as rowID” in the node’s configuration dialog.
- ◆ RowID: Interconnect the RowID node and define the Molecule name as “New RowID column” in the configuration dialog.



Reagent Specifier

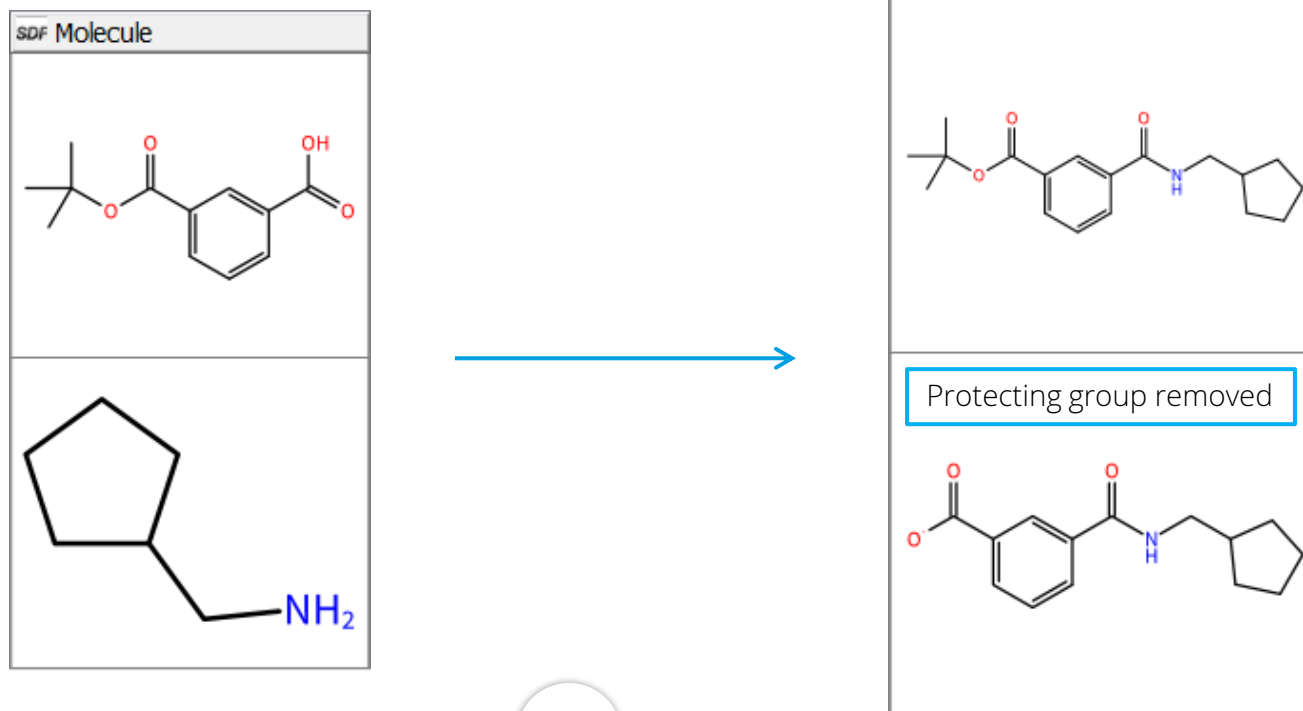
For some reactions it might be useful to specifically select building blocks for reagent 1 and building blocks for reagent 2. In this case the implementation of a Reagent specifier (number (integer)) hands over the information to CoLibri. Simply select the reagent specifier column in the “Reaction Product Enumerator”.



Remove Protecting Groups

Functional groups not participated in the reaction are sometimes protected. A list of protection groups can be found in the same way as the list of labels. If the “Remove Protecting Groups” is enabled in CoLibri, listed protecting group will be removed in the product molecule.

A carboxylic acid can be protected by a tert-butyl group. The protected or the deprotected form can be obtained in the product molecules



Recommendation summary

- ◆ Think in substructures! Identify the minimum substructure that precisely defines a suitable reagent! Use lists, SMARTS and labels to be as specific as needed!
- ◆ An “R” for a side chain is not necessary – in fact, it is not allowed!
- ◆ Use mapping numbers for all atoms which occur on both sides!
- ◆ No atom lists on the product side. Use any atom “A” instead!
- ◆ Do not draw localized bonds when you mean aromaticity!
- ◆ Do not draw explicit hydrogens. If needed define the number of heavy atom neighbors!

