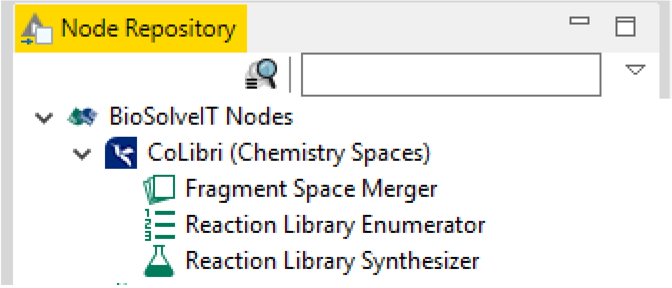
How to add your reactions to generate a Chemistry Space in KNIME

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 the 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.



Node Repository

BioSolveIT Nodes

CoLibri (Chemistry Spaces)

Fragment Space Merger

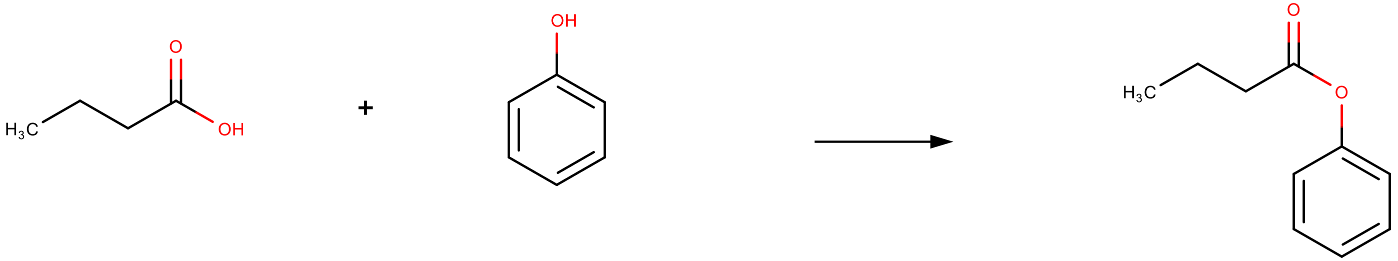
Reaction Library Enumerator

Reaction Library Synthesizer

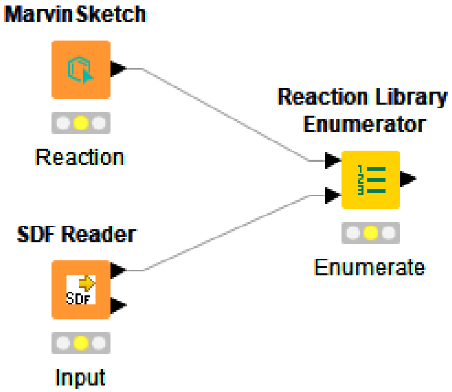
In the second part four examples are shown which are real parts of the CoLibri Space. These shall show you more cases of the reaction adoption to the tools.

PART 1 General Explanation using Esterification Reaction

In the very first step we want to explain how to define a reaction for CoLibri in the KNIME system. We showcase a simple esterification, where a carboxylic acid and an aromatic alcohol form an ester. One would “normally” draw it like this:

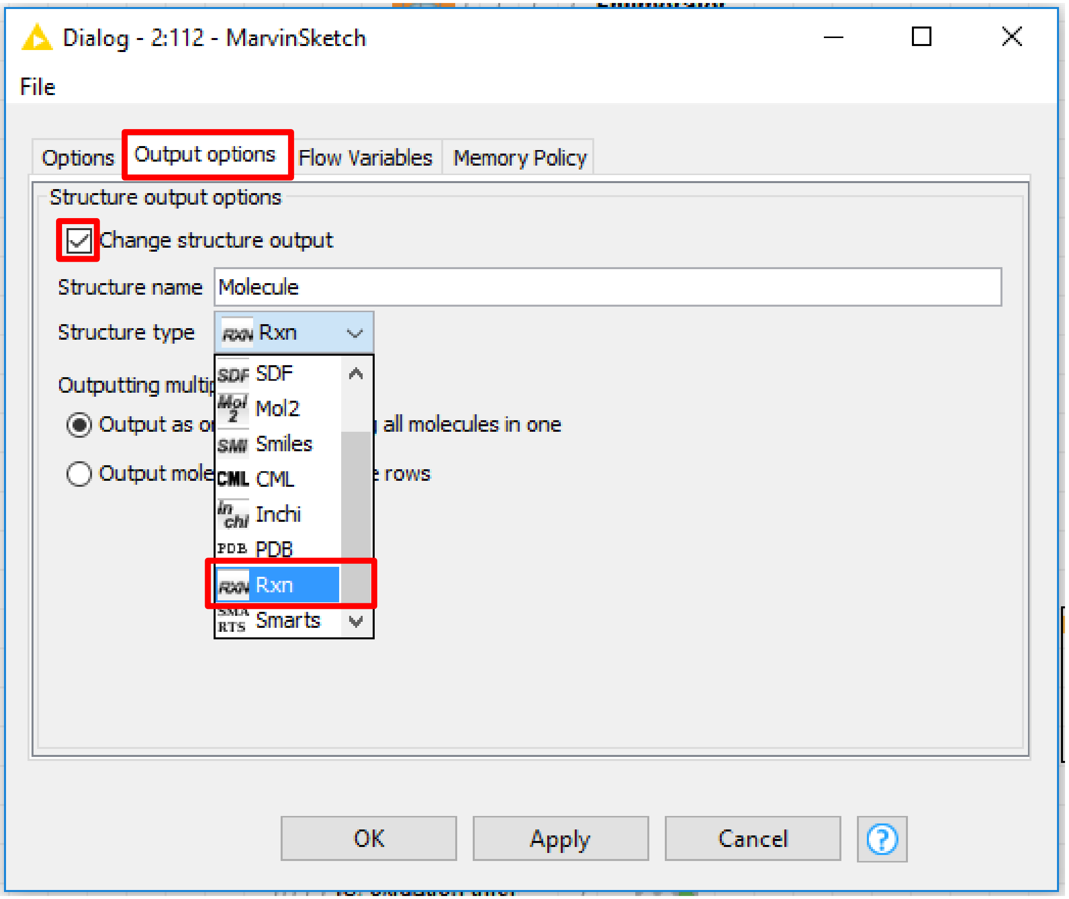


We recommend to use a recent update of the Marvin Sketch node to draw the reaction. Furthermore, we also draw some input molecules (reagents) and connect both to the reaction library enumerator node.



certainly the input port could also be connected to an *SDF-reader* node!

For this to work requires **a few critical changes**!
First, by default, Marvin uses its own format, but for the Reaction we need the reaction in RXN format! The reagent Input can be either SDF or smiles.


Dialog - 2:112 - MarvinSketch

File

Options | Output options | Flow Variables | Memory Policy

Structure output options

☑ Change structure output

Structure name: Molecule

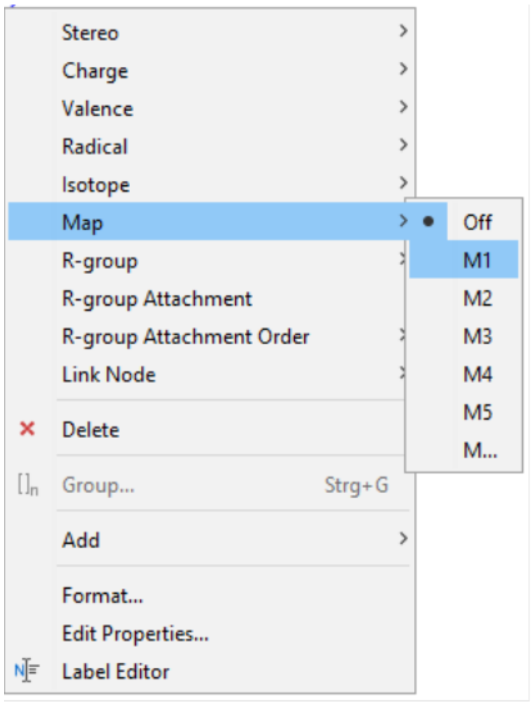
Structure type: Rxn

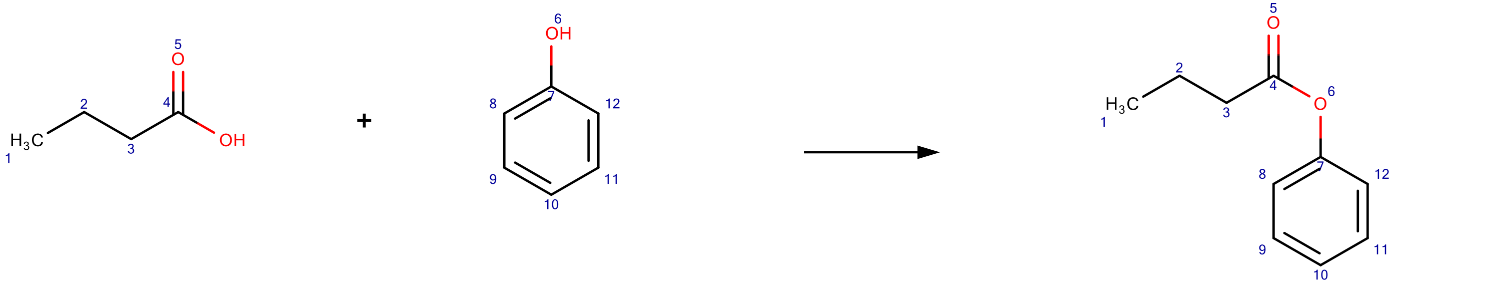
To make everything work properly go to the ”output options“ tab, check the “change structure output“ box and chose from the dropdown menu.

Note that this is not the default in Marvin

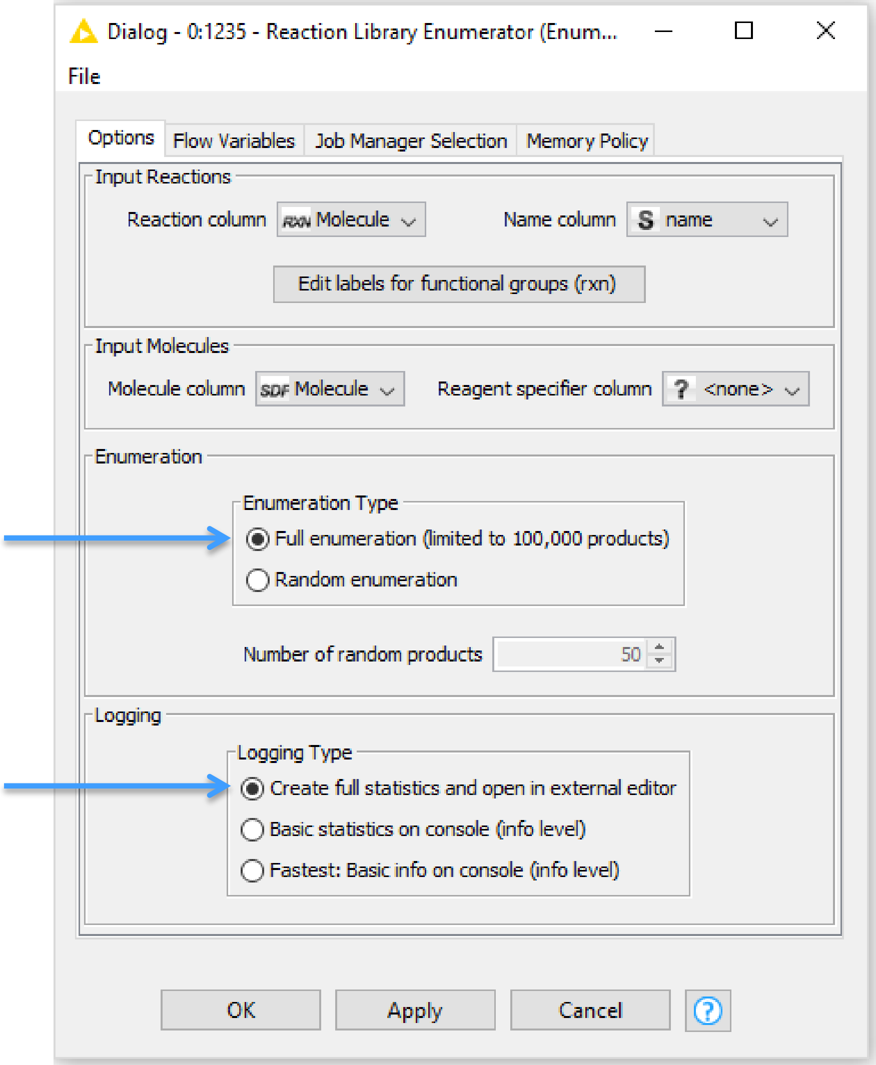
Second, the program needs a correspondence between the atoms on the reagent side and the related atoms on the product side. We call this “mapping atoms”. Simply right-click on an atom, then select ‘Map’ and a new number. Then do the same with the corresponding atom on the product side.

Add mapping numbers only to those atoms which occur on both sides! Atoms without mapping number on the reagent side will be removed, e.g. leaving groups.





Let's try to run the reaction now. During the editing of a reaction, it is always useful to let the *Enumerator* generate some products and look at the statistics. Therefore the node should be configured like that:



Dialog - 0:1235 - Reaction Library Enumerator (Enum...

File

Options | Flow Variables | Job Manager Selection | Memory Policy

Input Reactions

Reaction column: RXN Molecule

Name column: S name

Edit labels for functional groups (rxn)

Input Molecules

Molecule column: SDF Molecule

Reagent specifier column: ? <none>

Enumeration

Enumeration Type

● Full enumeration (limited to 100,000 products)

○ Random enumeration

Number of random products: 50

Logging

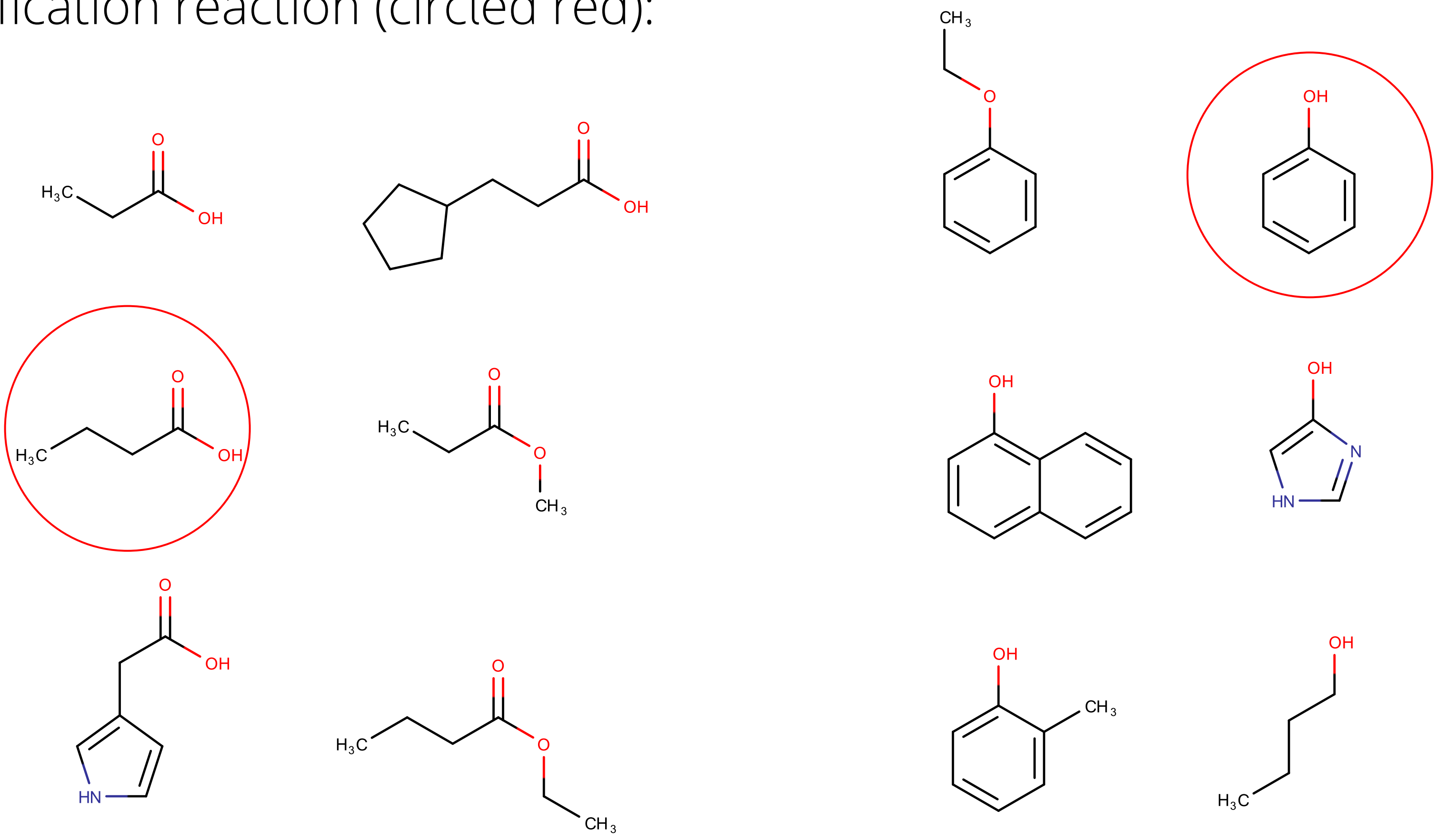
Logging Type

● Create full statistics and open in external editor

○ Basic statistics on console (info level)

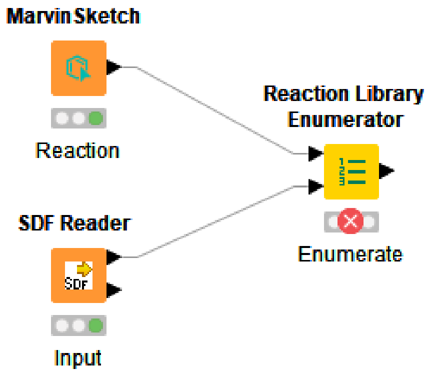
○ Fastest: Basic info on console (info level)

In the input set we have drawn the following molecules. We recommend you to also sketch some molecules that should match and some that should not! In the set are the two reagents relevant for the sketched esterification reaction (circled red):

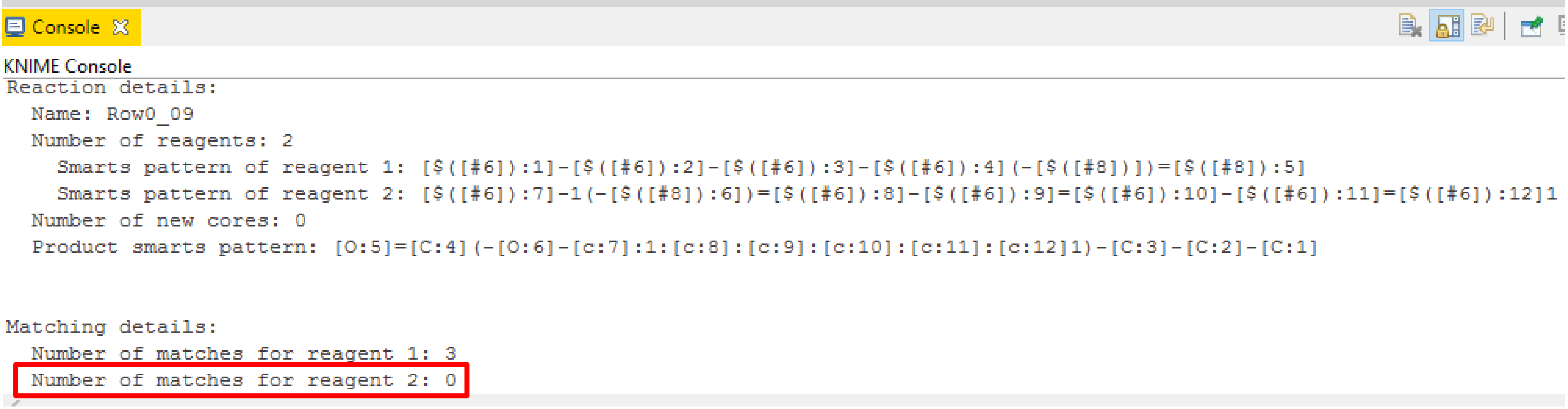


So, let’s try running the reaction!

Unfortunately the Synthesizer node turns red:



When we look at the Console in KNIME or into the log file, we see that no matches were found for reagent 2!



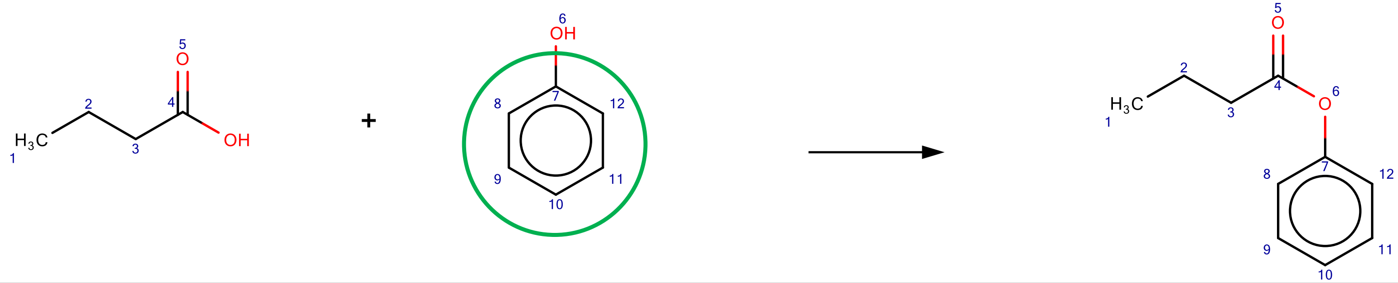
KNIME Console

```
Reaction details:
  Name: Row0_09
  Number of reagents: 2
    Smarts pattern of reagent 1: [$([#6]):1]-[$([#6]):2]-[$([#6]):3]-[$([#6]):4](-[$([#8])])=[$([#8]):5]
    Smarts pattern of reagent 2: [$([#6]):7]-1(-[$([#8]):6])=[$([#6]):8]-[$([#6]):9]=[$([#6]):10]-[$([#6]):11]=[$([#6]):12]1
  Number of new cores: 0
  Product smarts pattern: [O:5]=[C:4](-[O:6]-[c:7]:1:[c:8]:[c:9]:[c:10]:[c:11]:[c:12]1)-[C:3]-[C:2]-[C:1]


Matching details:
  Number of matches for reagent 1: 3
  Number of matches for reagent 2: 0
```

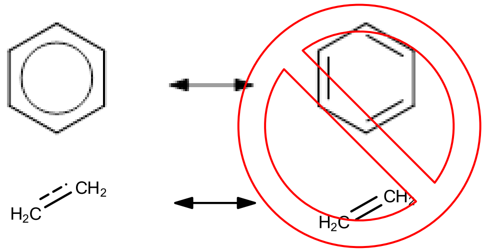
This seems strange, because we have drawn the matching reagent. But it is a question of the details in the drawing...

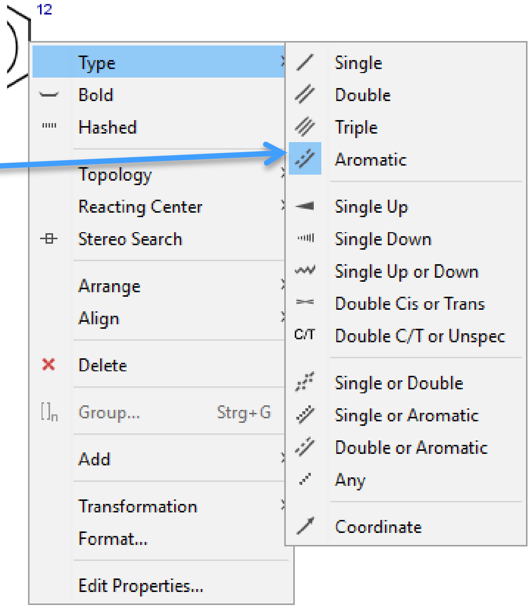
When using the RXN format for the reaction, Marvin processes compounds with alternating bonds exactly like they were sketched. The SDF reagent input is interpreted to have aromatic bonds. So **third**, we have to change our reaction drawing as follows:



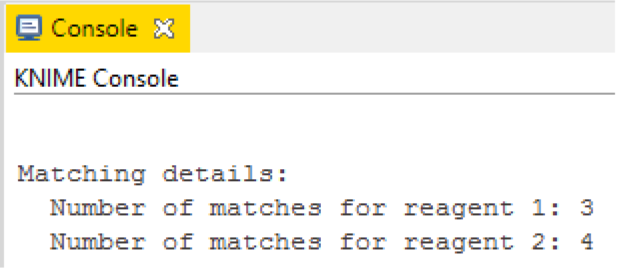
Note: Please use the aromatic form whenever possible and appropriate. It is available in the templates or by right-clicking on the bond and selecting aromatic.

Don't draw alternating single and double bonds!



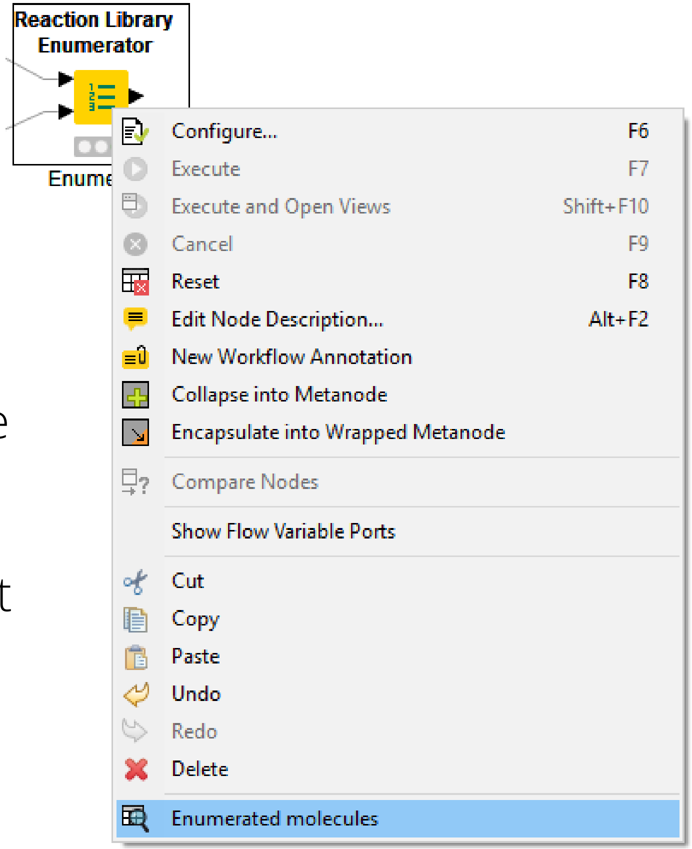


Now if we run the reaction it works fine!
But we get 8 products instead of the expected 1. So have many more matches than anticipated:



```
Matching details:
  Number of matches for reagent 1: 3
  Number of matches for reagent 2: 4
```

Let’s take a look at the enumerated products by right-clicking on the Synthesizer and choosing ‘Enumerated molecules’ to see what happened.

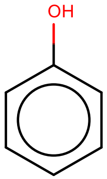


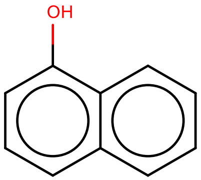
Row ID	SDF Enumerated molecule
Row0	
Row1	
Row2 ☹	
Row3	
Row4	
Row5	
Row6	
Row7	

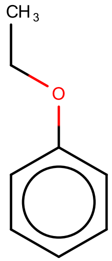
In Row3 we find the desired product. But we see as well trash products like in Row2…

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!

Let's look at reagent two:
Phenol is not only the substructure of Napthol but as well of the Ethoxybenzene.







This leads us to an important conclusion:

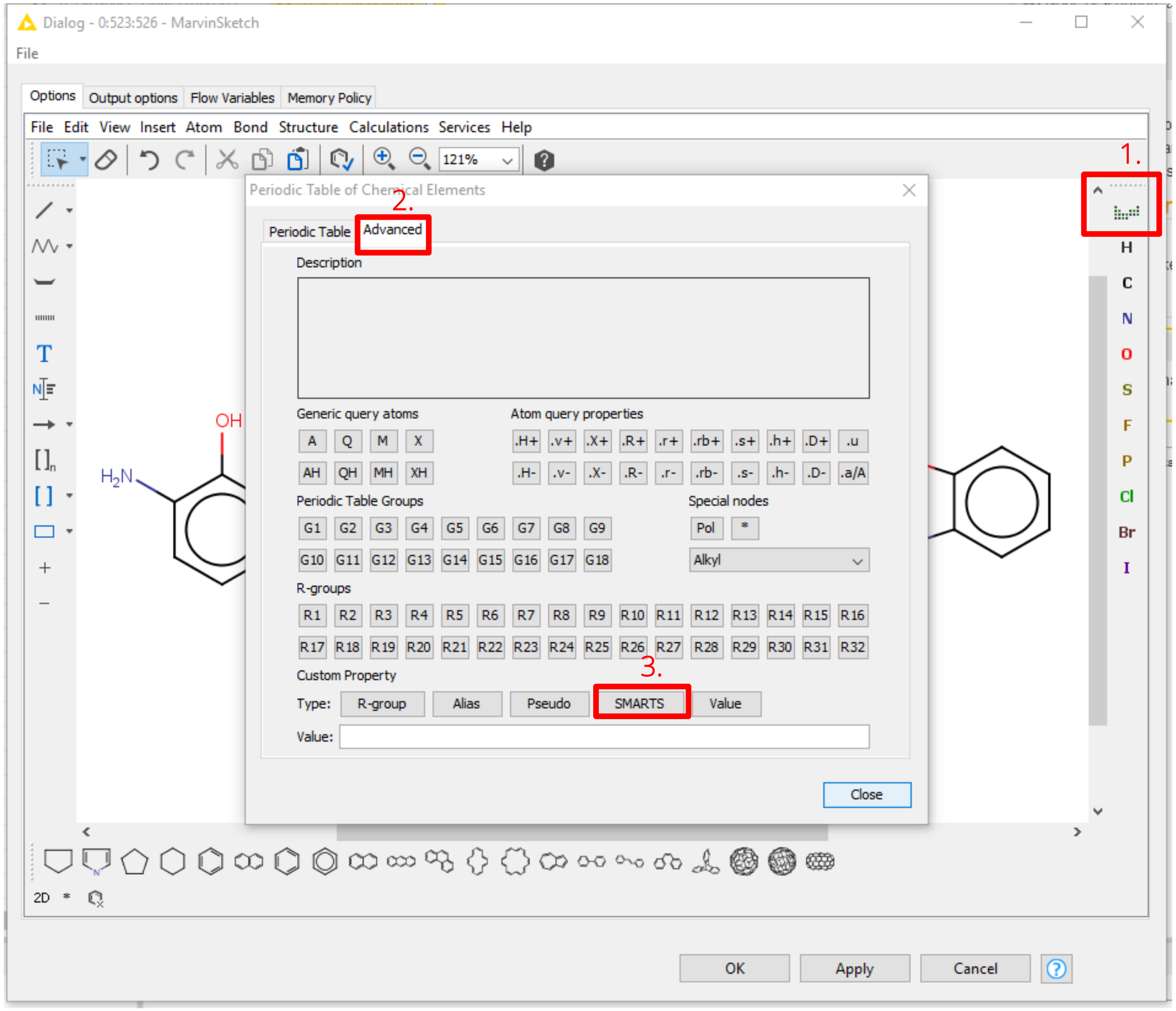
Be as specific as possible!

E.g. annotate, if something is supposed to be terminal. Even if the Napthyl could be acceptable for this reaction the Ether is definitely not! Also:

Note that Hydrogens are ignored!

They are just added by the drawing program by default.
You can easily define the environment of an atom in two different ways...

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


Dialog - 0:523:526 - MarvinSketch
File
Options | Output options | Flow Variables | Memory Policy
File Edit View Insert Atom Bond Structure Calculations Services Help
121%
1.
Periodic Table of Chemical Elements
2.
Periodic Table | Advanced
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
Special nodes
Pol *
Alkyl
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
3.
Custom Property
Type: R-group Alias Pseudo SMARTS Value
Value:
Close
H C N O S F P Cl Br I
OH
H_2N
2D
OK Apply Cancel

1.Click on the periodic system symbol

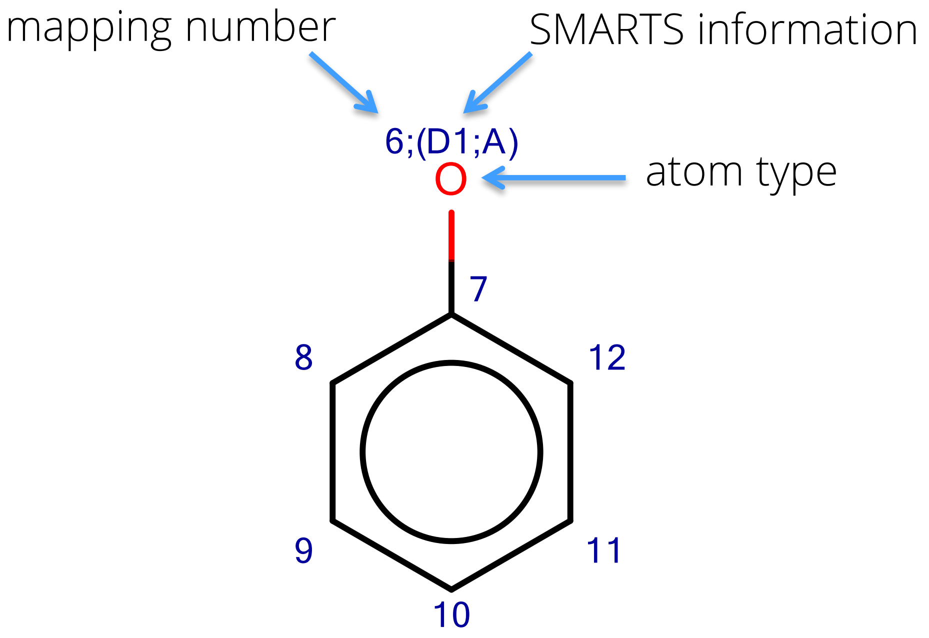
2. Click on the “Advanced” tab and

3. chose SMARTS tab to fill in your definition. If you are not sure about your SMARTS please check

http://daylight.com/dayhtml_tutorials/languages/smarts/index.html

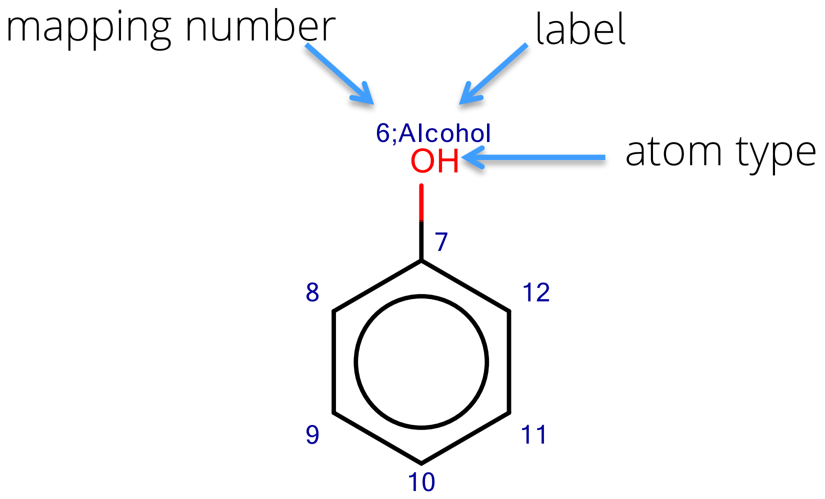
for those of you who have a SMARTS-phobia, still read on …

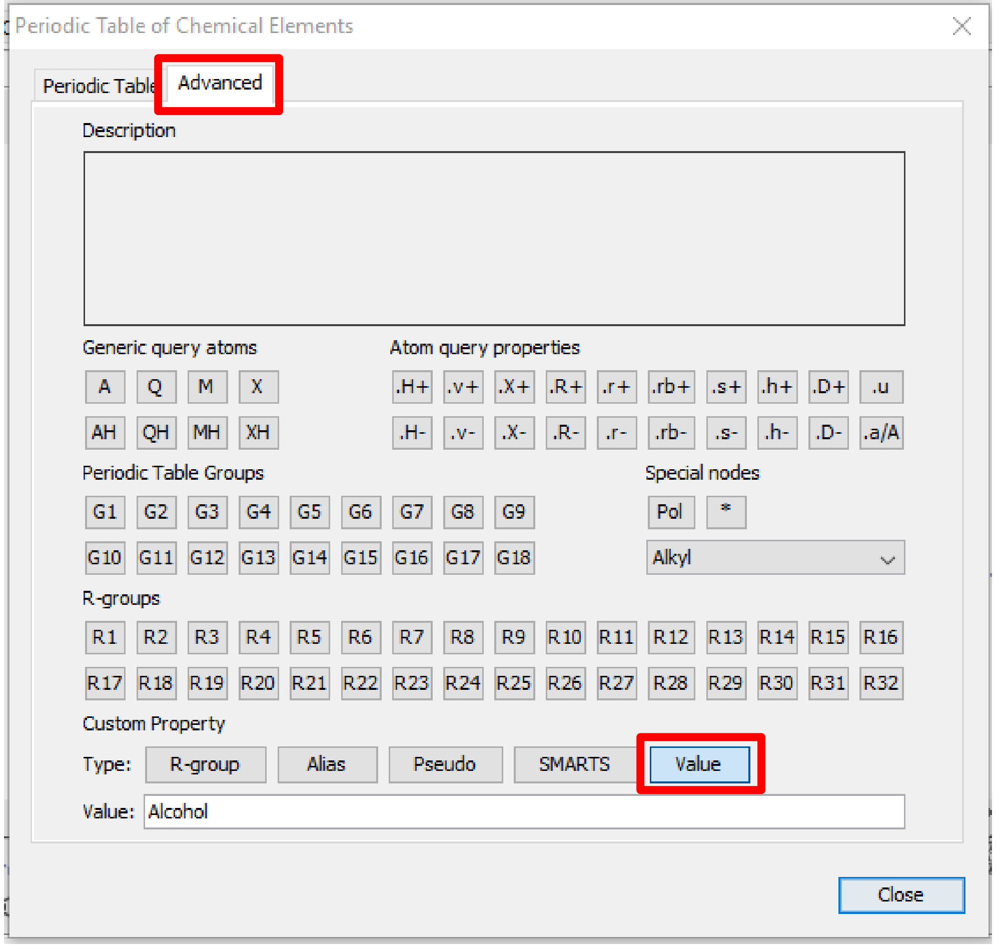
For the Phenol we would type [O;D1] and click on the atom we want to edit. D1 says that only one neighbor atom is allowed to be a heavy atom. The drawing looks now like that:



Alternatively to SMARTS you can use **"labels"**. Labels 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 with the label:




Periodic Table of Chemical Elements

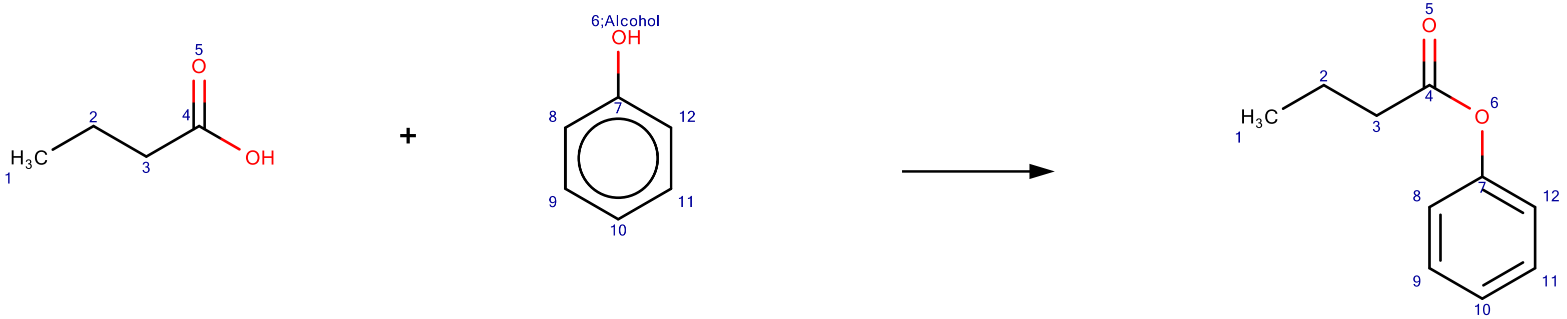
Periodic Table
Advanced
Description
Generic query atoms
A Q M X
AH QH MH XH
Atom query properties
.H+ .v+ .X+ .R+ .r+ .rb+ .s+ .h+ .D+ .u
.H- .v- .X- .R- .r- .rb- .s- .h- .D- .a/A
Periodic Table Groups
G1 G2 G3 G4 G5 G6 G7 G8 G9
G10 G11 G12 G13 G14 G15 G16 G17 G18
Special nodes
Pol *
Alkyl
R-groups
R1 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
Custom Property
Type: R-group Alias Pseudo SMARTS Value
Close

Value: Alcohol

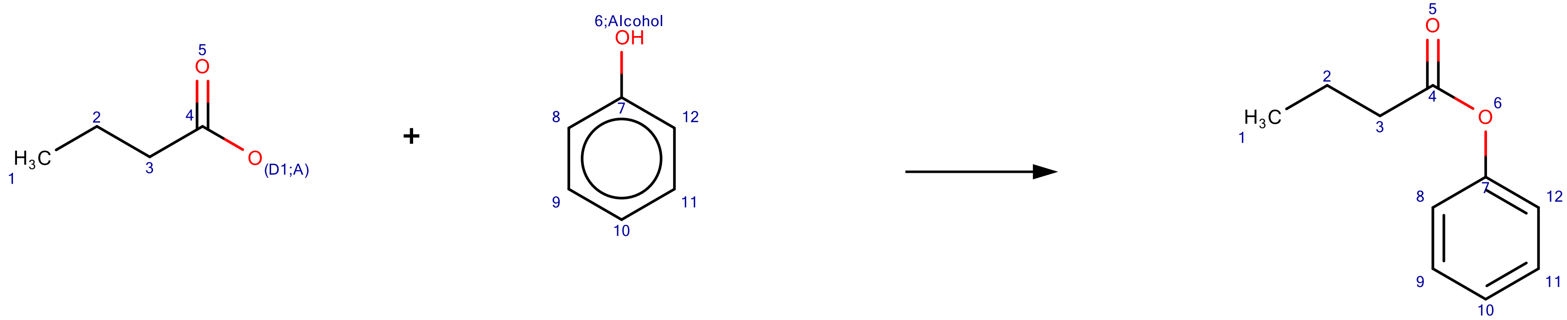
All pre-defined labels can be found on the next slides together with the SMARTS that they represent. They can be found an edited in the *Enumerator* as well.

LABEL	The SMARTS behind it
AcidChloride	C(=O)Cl
CarboxylicAcid	C(=O)[O;H,-]
AlphaAminoAcid	[$(C-[C;!$(C=[!#6])]-[N;!H0;!$(N-[!#6;!#1]);!$(N-C=[O,N,S])])](=O)([O;H,-])
SulfonylChloride	[$(S)](=O)(=O)(Cl)
Sulfone	[$([#16X4](=[OX1])=[OX1]),$([#16X4+2]([OX1-])[OX1-])]
Amine	[N;!H0;$(N-[#6]);!$(N-[!#6;!#1]);!$(N-C=[O,N,S])]
Amine.Primary	[N;D1;!$(N-C=[O,N,S])]
Amine.Secondary	[N;D2;$(N(-[#6])-[#6]);!$(N-[!#6;!#1]);!$(N-C=[O,N,S])]
Amine.Tertiary	[NX3;H0;D3;!+1]
Nitro	[N;$(N(=O)O)]
BoronicAcid	[$(B-!@[#6])](O)(O)
Isocyanate	[$(N-!@[#6])](=!@C=!@O)
Alcohol	[O;H1;$(O-!@[#6;!$(C=!@[O,N,S])])]
Aldehyde	[CH;D2;$(C-[#6])]=O
Halogen	[F,Cl,Br,I]
Azide	[N;H0;$(N-[#6]);D2]=[N;D2]=[N;D1]
Aromatic	a
Leaving.Group	[Cl,Br,I,O&$(OS(=O)(=O)C),O&$(OS(=O)(=O)c1ccc(cc1)C),O&$(OS(=O)(=O)C(F)(F)F)]
EWG	[F,Cl,N&$(N(=O)O),S&$(S(=O)O),C&$(C(=O)O)]
EDG	[N&$([N;H2]),O&$([O;H]),C&$([C;H2])]
Heavy.Atom	[*;!#1]

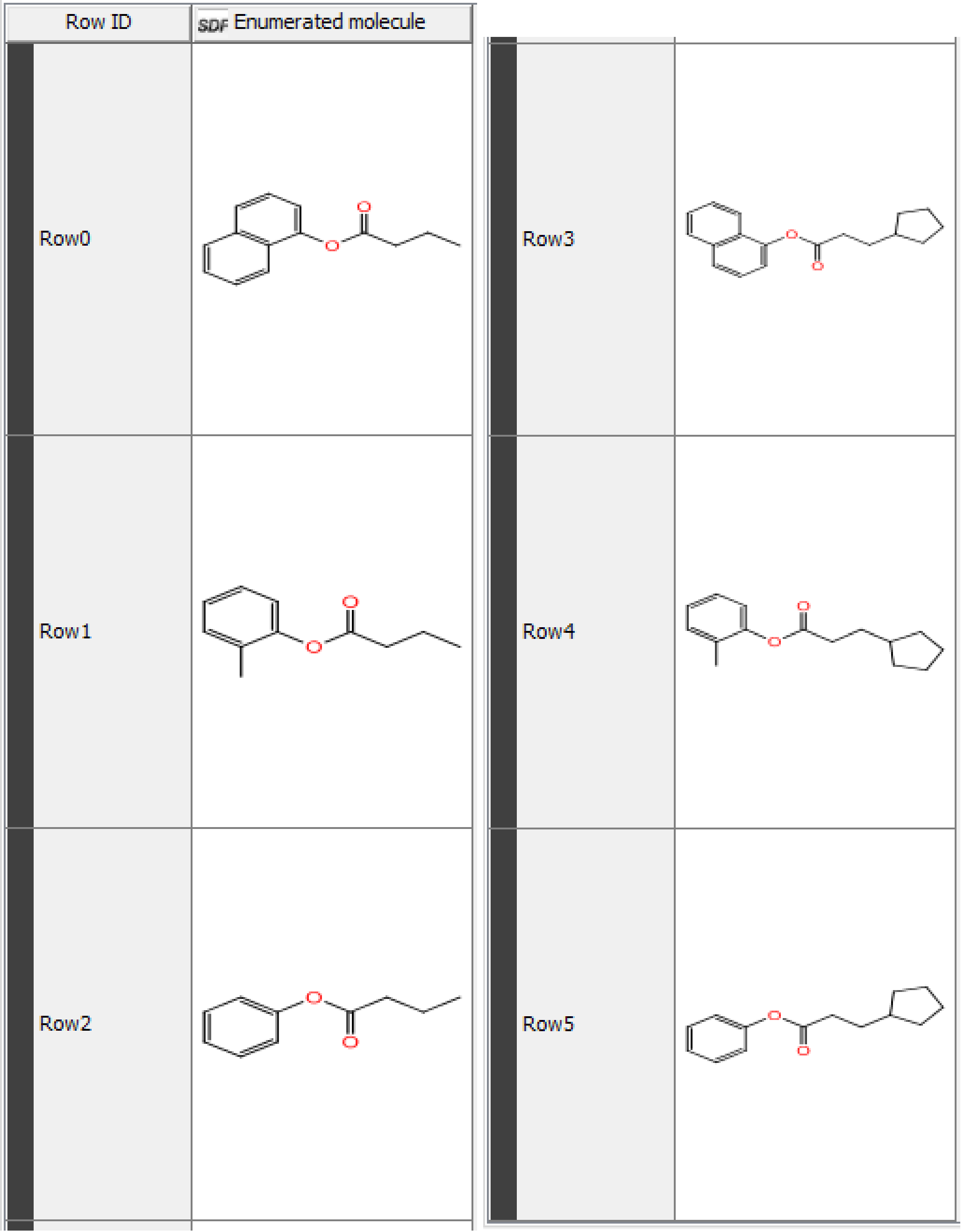
Now let’s look again at the whole reaction:



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

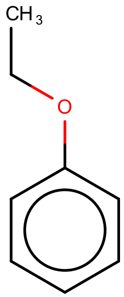


Now run the reaction again!

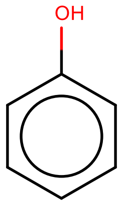


Row ID	SDF Enumerated molecule
Row0	
Row1	
Row2	
Row3	
Row4	
Row5	

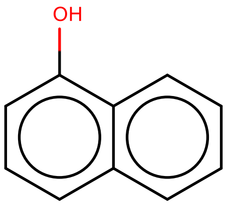
These are the resulting molecules.
Now we have no senseless valences at the oxygen anymore.
But we have still Naphtol and Cresol matching the aromatic ring.
Let’s look at the input and decide what we want to allow for the individual reagents.



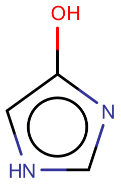
1



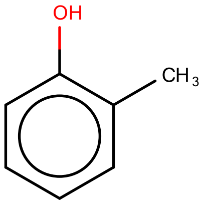
2



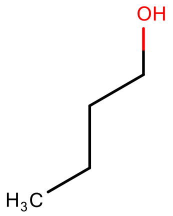
3



4



5

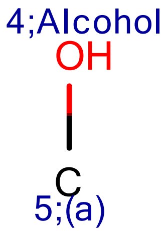


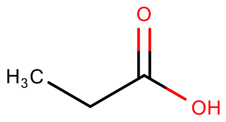
6

Look at the input molecules for reagent 2 and think about which structures you want to allow for the reaction.

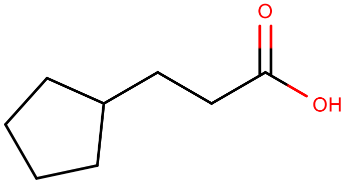
As we want an aromatic alcohol as reagent, let's say 2 – 5 are desired, but not 1 and 6. Therefore we have to reduce the atoms in the reaction definition. The common substructure molecules 2 – 5 have are an aromatic carbon with a primary alcohol. Therefore we reduce the definition of reagent 2 this by using the SMARTS [c] and the label 'Alcohol'.

An aromatic Carbon is by Marvin marked with (a) in contrast to (A) for aliphatic.

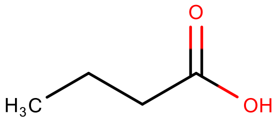




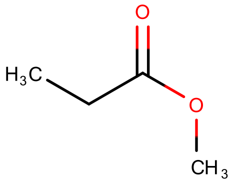
1



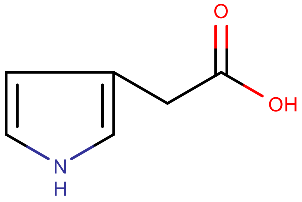
2



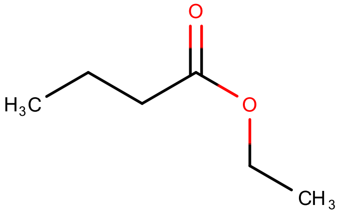
3



4

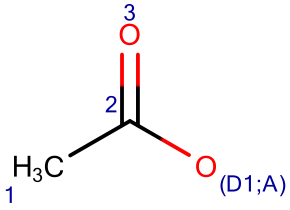


5

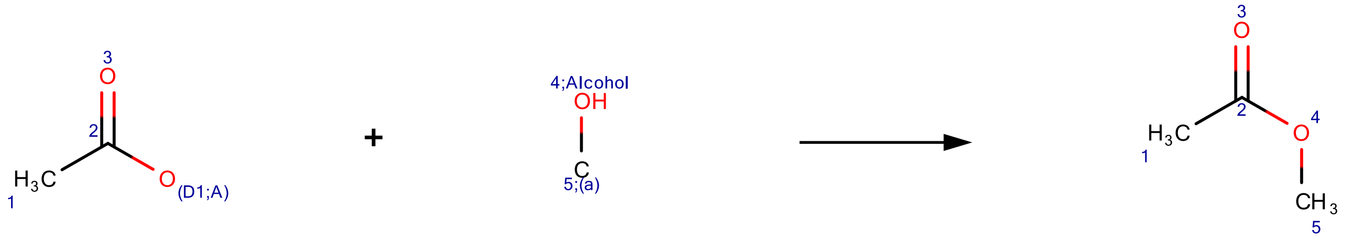


6

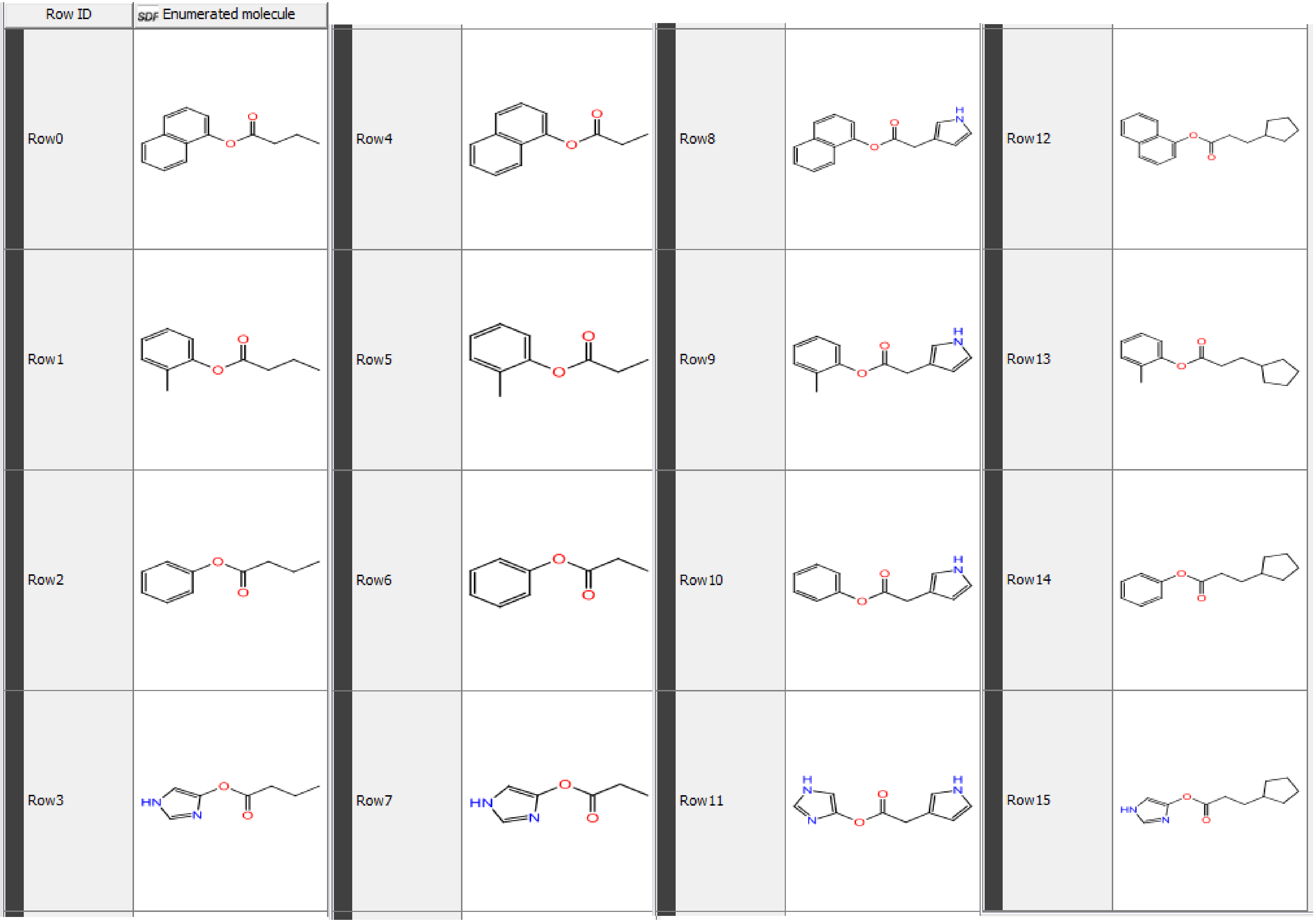
Now looking at reagent 1, from a chemists point of view free acids could work for reagent 1 (all except 4 and 6).
With the actual definition of the reaction only 2 and 3 are matching. Because a butyric acid is drawn, 1 and 5 do not match. Therefore we reduce the describing atoms to the minimum substructure, which is acetic acid



Altogether, this leads to the final definition of the reaction:

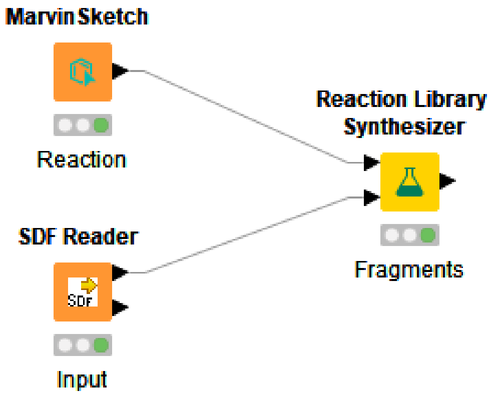


If we run this reaction we get 4 matching molecules for every reagent, like desired. This leads to 16 different products that are shown on the next slide.



Row ID	SDF Enumerated molecule
Row0	
Row1	
Row2	
Row3	
Row4	
Row5	
Row6	
Row7	
Row8	
Row9	
Row10	
Row11	
Row12	
Row13	
Row14	
Row15	

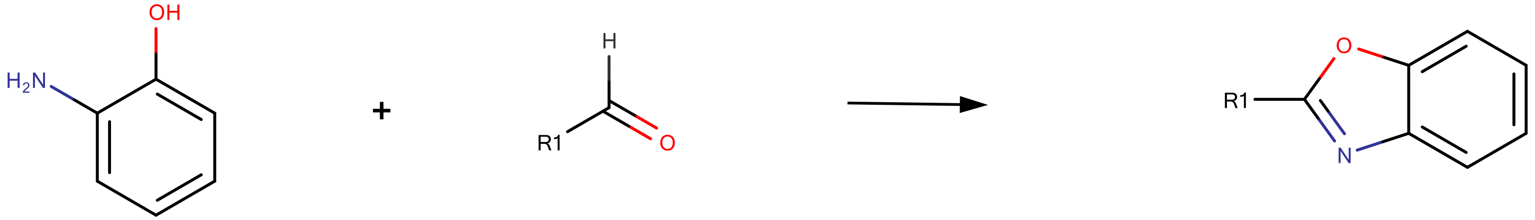
Last but not least: if you want to build a real Fragment Space, replace the *Enumerator* by a *Reaction Library Synthesizer*.



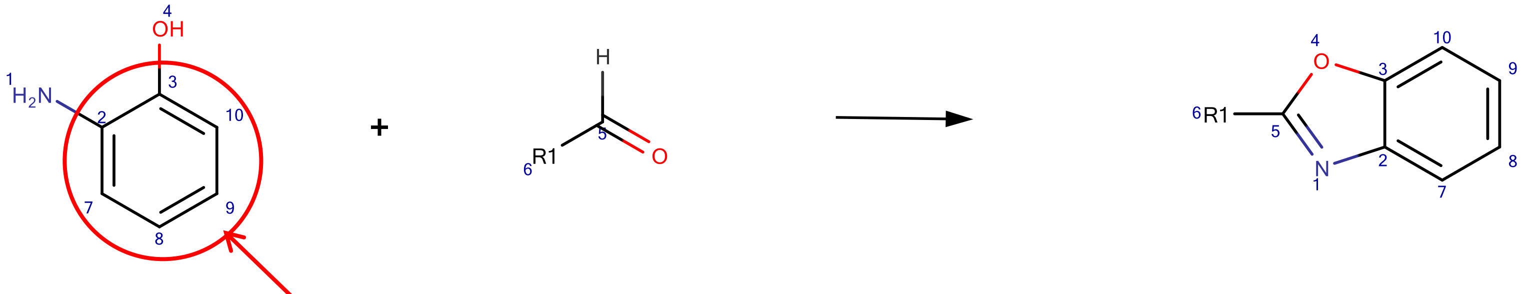
Use a *Concatenate* node to combine several reactions followed by the *Fragment Space Merger* to generate a searchable space.

PART 2: Example 1 – Ringclosing: (Benzoxazole-Reaction)

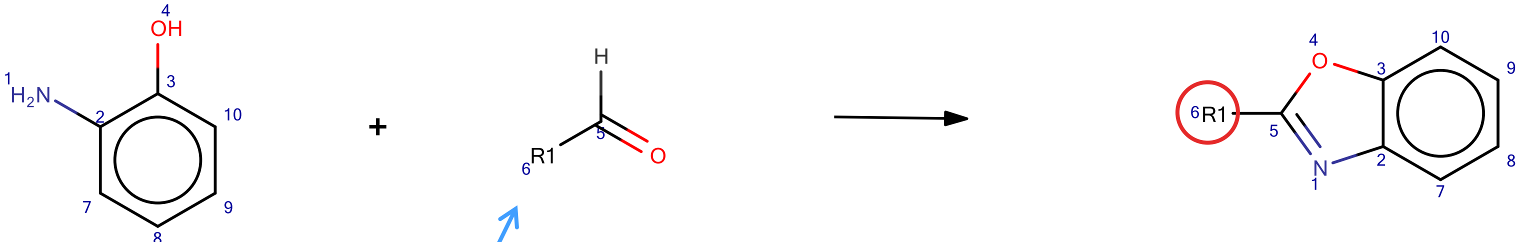
As first example we use something similar to Benzoxazole ring closure, where an aminophenol reacts with an aldehyde:



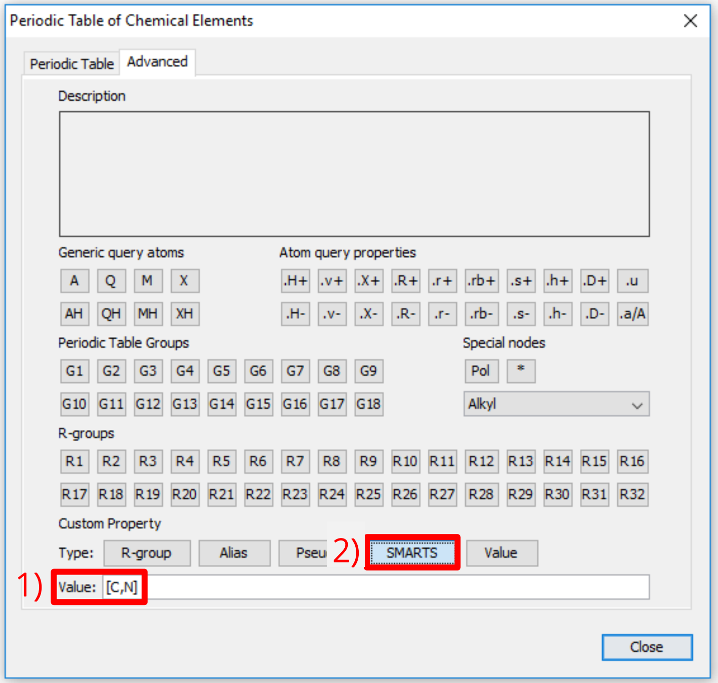
Let’s make it suitable for the computer. First we add the mapping atoms:



Please use the aromatic form. Don’t draw alternating single and double bonds!



The software does not know R groups so replace it with the atom that should be in this position. **Think in substructures!**
In this example we want R1 to be Carbon or Nitrogen. So we type [C,N] in the "value" field, click on 'SMARTS' and afterwards on R1 in the drawing.

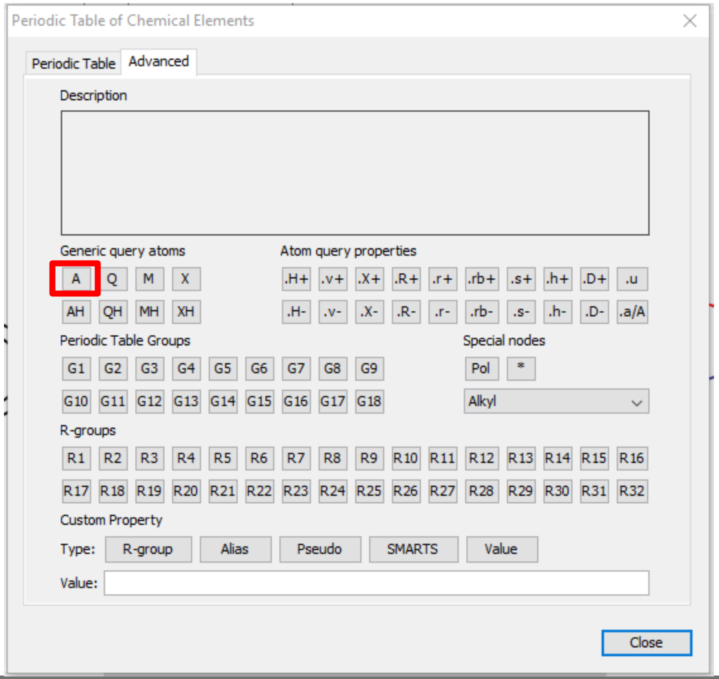


Periodic Table of Chemical Elements
Periodic Table
Advanced
Description
Generic query atoms
A Q M X
AH QH MH XH
Atom query properties
.H+ .v+ .X+ .R+ .r+ .rb+ .s+ .h+ .D+ .u
.H- .v- .X- .R- .r- .rb- .s- .h- .D- .a/A
Periodic Table Groups
G1 G2 G3 G4 G5 G6 G7 G8 G9
G10 G11 G12 G13 G14 G15 G16 G17 G18
Special nodes
Pol *
Alkyl
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
Custom Property
Type:
R-group
Alias
Pseu
SMARTS
Value
Close

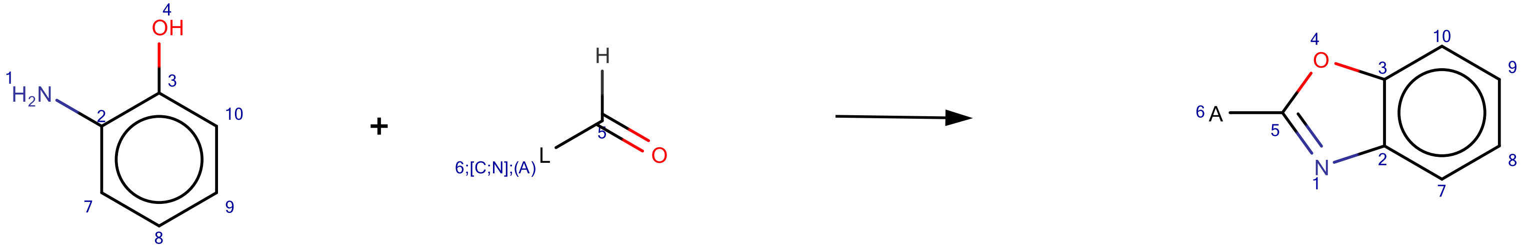
1) Value: [C,N]

2) SMARTS

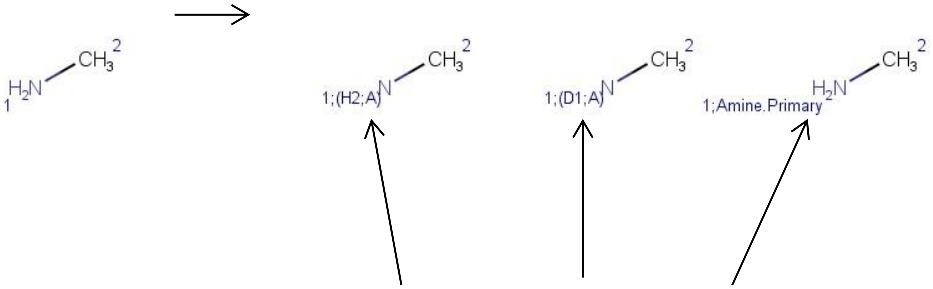
Atom lists or R-group identifiers are not allowed on the product side! Use the query atom type "A" for "any atom" instead.



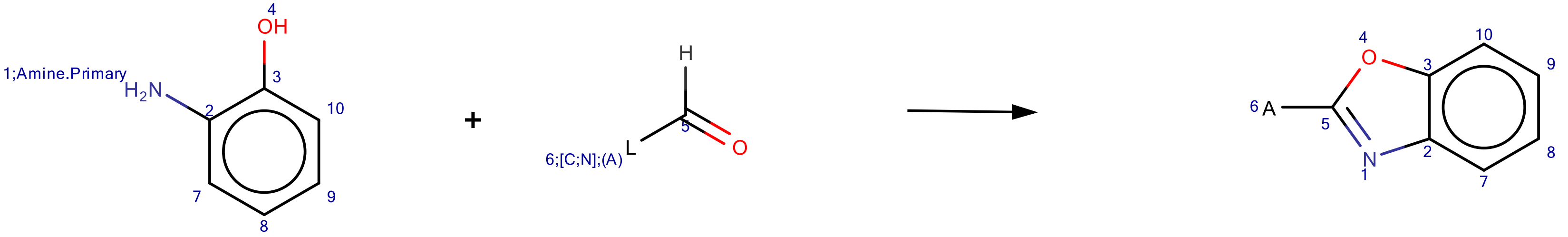
Periodic Table of Chemical Elements
Periodic Table
Advanced
Description
Generic query atoms
A Q M X
AH QH MH XH
Atom query properties
.H+ .v+ .X+ .R+ .r+ .rb+ .s+ .h+ .D+ .u
.H- .v- .X- .R- .r- .rb- .s- .h- .D- .a/A
Periodic Table Groups
G1 G2 G3 G4 G5 G6 G7 G8 G9
G10 G11 G12 G13 G14 G15 G16 G17 G18
Special nodes
Pol *
Alkyl
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
Custom Property
Type:
R-group
Alias
Pseudo
SMARTS
Value
Value:
Close



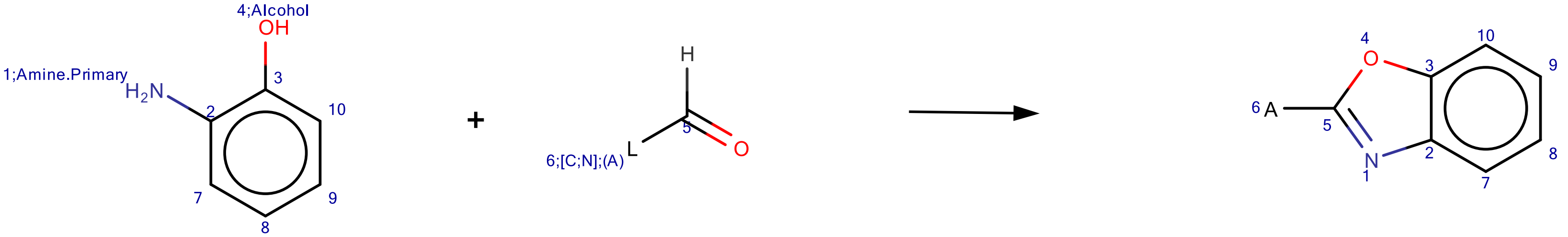
The reaction is nearly right now, but the amine needs to be more specific. Again: Hydrogens are ignored. You can either use the SMARTS, alternatively you can use a label.



These three mean similar things, but the label 'Amine.Primary' contains the most precise definition, i.e. [N;D1;!$(N-C=[O,N,S])]
All labels can be found in the on page14.

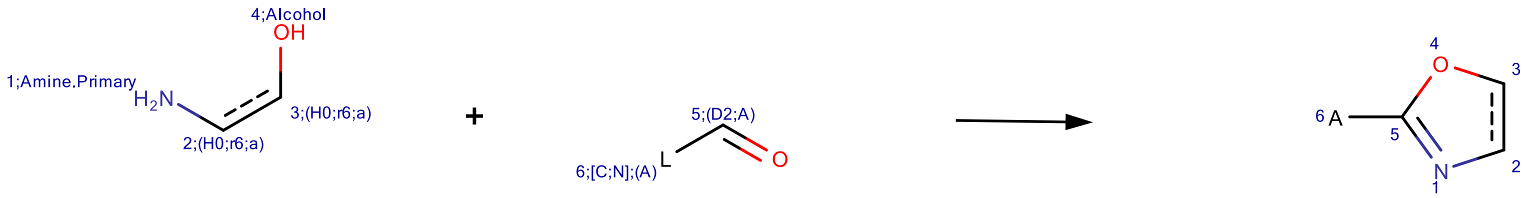


Next we mark the oxygen to be an alcohol (and not the substructure of an ether). Again the easiest way is to use the label 'Alcohol' leading to:



Again remember: Atom lists, Labels and SMARTS are not needed – and in fact not allowed – on the product side! The mapping atom is sufficient to transfer the information.

In the last step we remove any atoms not needed to define the reaction. Keeping the substructures to the essential minimum reduces the risk of errors during the subsequent processing. E.g. the aromatic ring is not important for the reaction. Hence we reduce it to the connecting carbons with an implicit definition. This is again done in the advanced editor by typing the SMARTS [cH0;r6] (which means an aromatic carbon without hydrogens in a six-membered ring. This leads us to the final reaction definition:

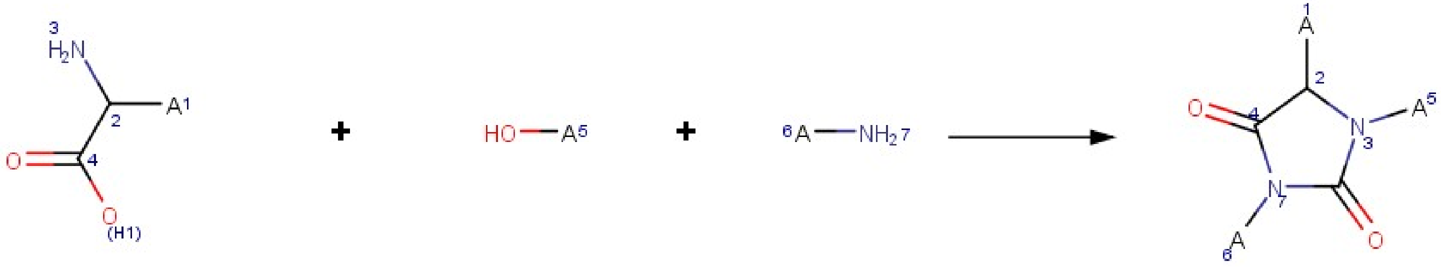


Explicit Hydrogens should be avoided! This could be again a reason for wrong matching. Therefore atom 5 was edited to [CD2] (carbon connected to two heavy atoms only) containing the hydrogen in an implicit definition.

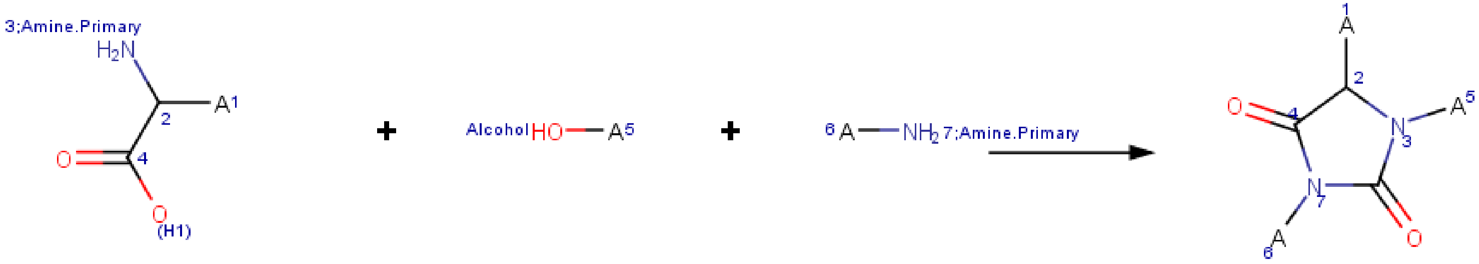
NOTE: you can either type [CH] or [CD2] which is in principle the same. Definition of heavy atoms is easier applicable to more cases.

Example 2 - Multicomponent: (Scicinski-Reaction)

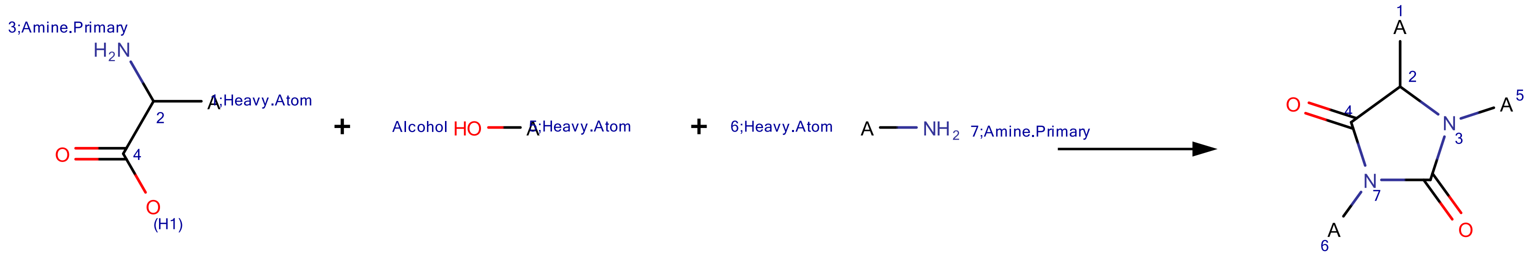
In this reaction we want an amino acid to react with an alcohol and amine to build a heterocycle. Reagents were drawn and atom mappings added to all atoms occurring on both sides of the reaction arrow:



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



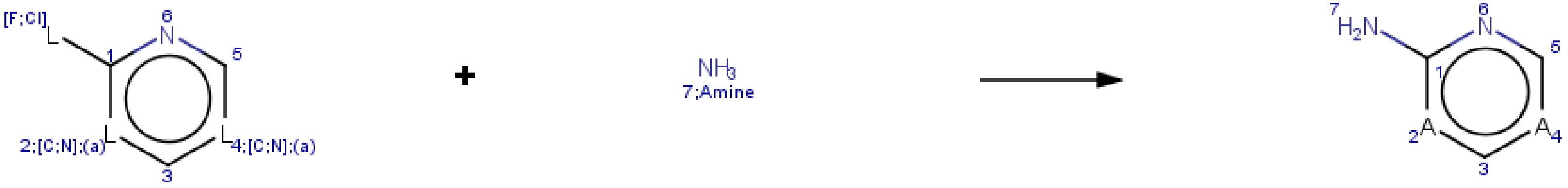
Atom 1 shall be allowed to be any heavy atom. The “*” in a SMARTS covers also a Hydrogen, therefore the correct SMARTS must be [*;!#1]. Alternatively you could use the label ‘Heavy.Atom’, which leads us to a suitable reaction definition. Same expression is used for Atoms 5 and 7.


3;Amine.Primary
H_2N
A 1;Heavy.Atom
2
O
4
O
(H1)
+
Alcohol HO — A 5;Heavy.Atom
+
6;Heavy.Atom A — NH_2 7;Amine.Primary
1
A
2
O
4
N
3
A
5
N
7
A
6
O

Example 3 -Substitution:

(hetero-aromatic-Substitution)

In a hetero-aromatic substitution different numbers of Nitrogens can be in the aromatic ring. Therefore we make only one Nitrogen mandatory (to be surely having a hetero-aromatic substitution) and in two more positons 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 A’s.

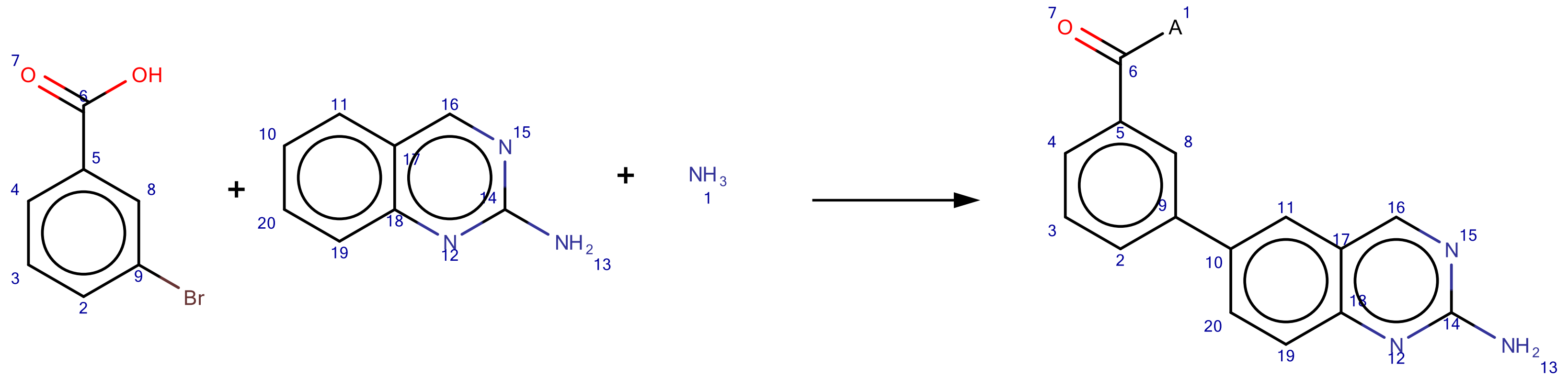


Furthermore we need an amine (primary or secondary). For this we add N7 as second reagent and use the label ‘Amine’.

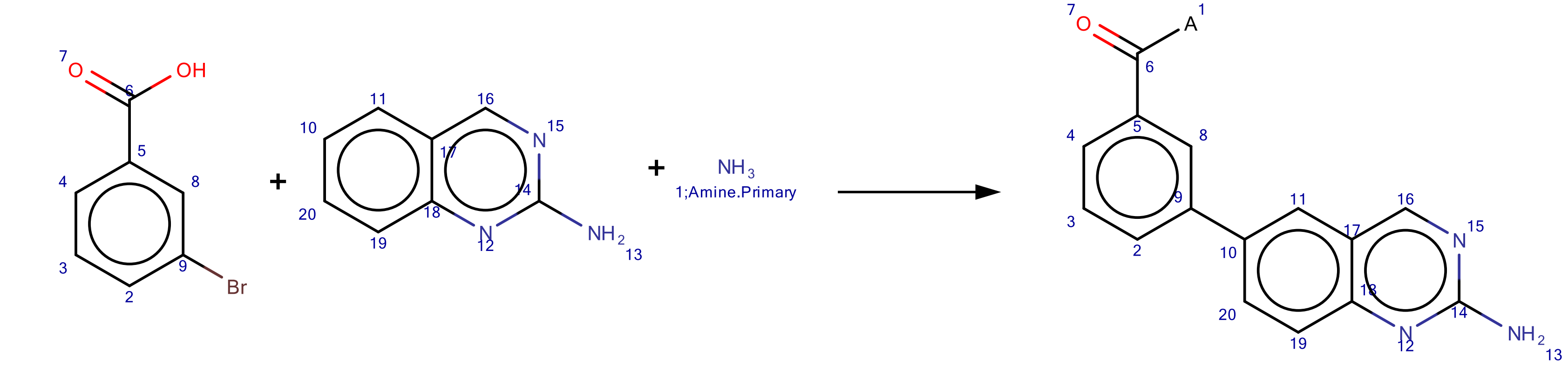
Example 4 - Coupling:

(Di Mauro-Reaction)

This reaction is not only an example for a coupling reaction, but as well for a two-stage reaction. In the first step a carbon-carbon coupling takes place, followed by an amidation. The reagents are drawn below and all atoms already mapped:

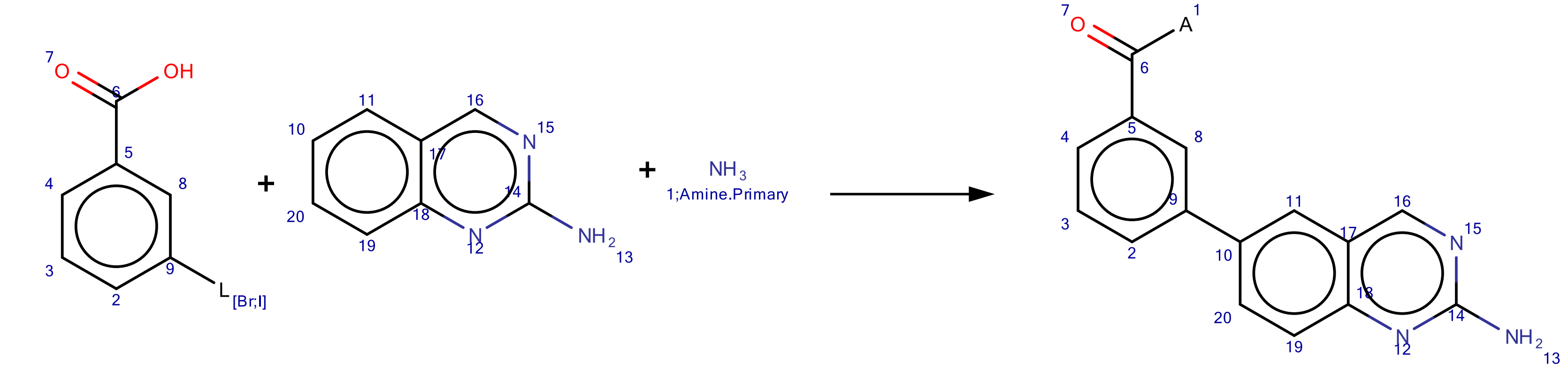


The reaction definition needs 4 more bits of information to be precise.

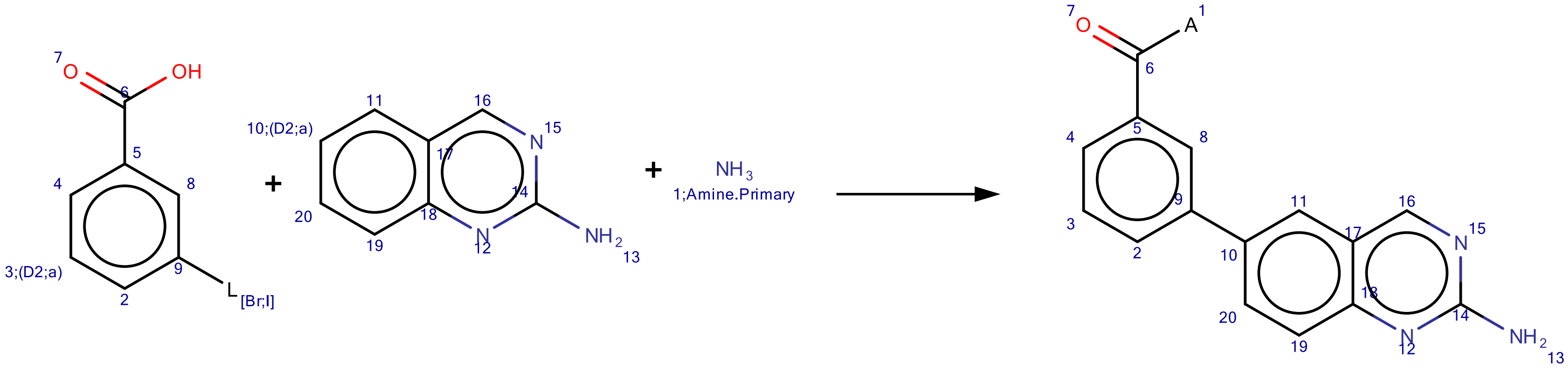


First, we add the Amine.Primary label to N1. Note that this label is not needed for N13, all substitutions are allowed there.

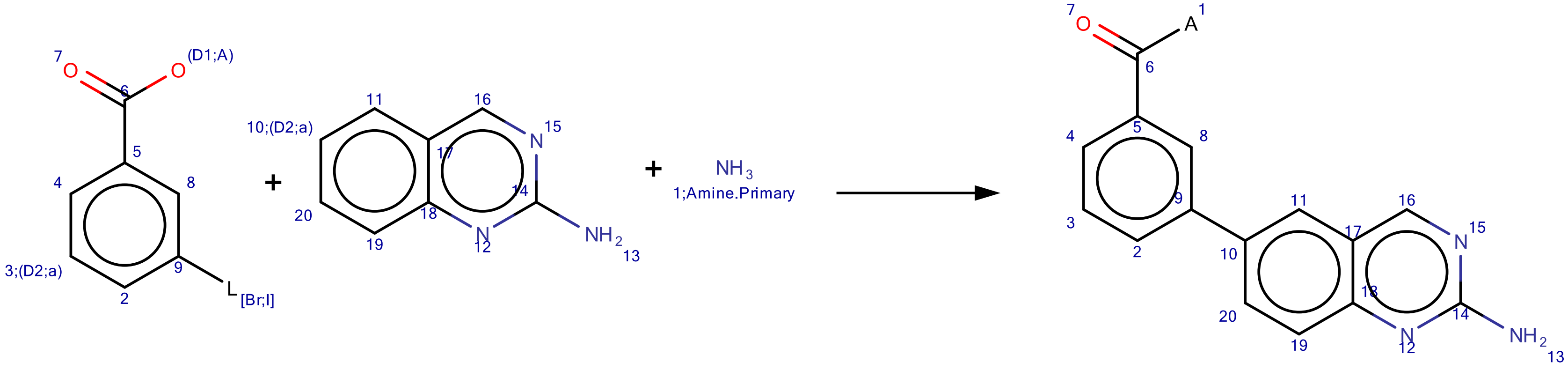
Second, we modify the leaving group connected to c_9. According to the reaction description, it can be either Br or I, therefore we add the SMARTS [Br,I] here



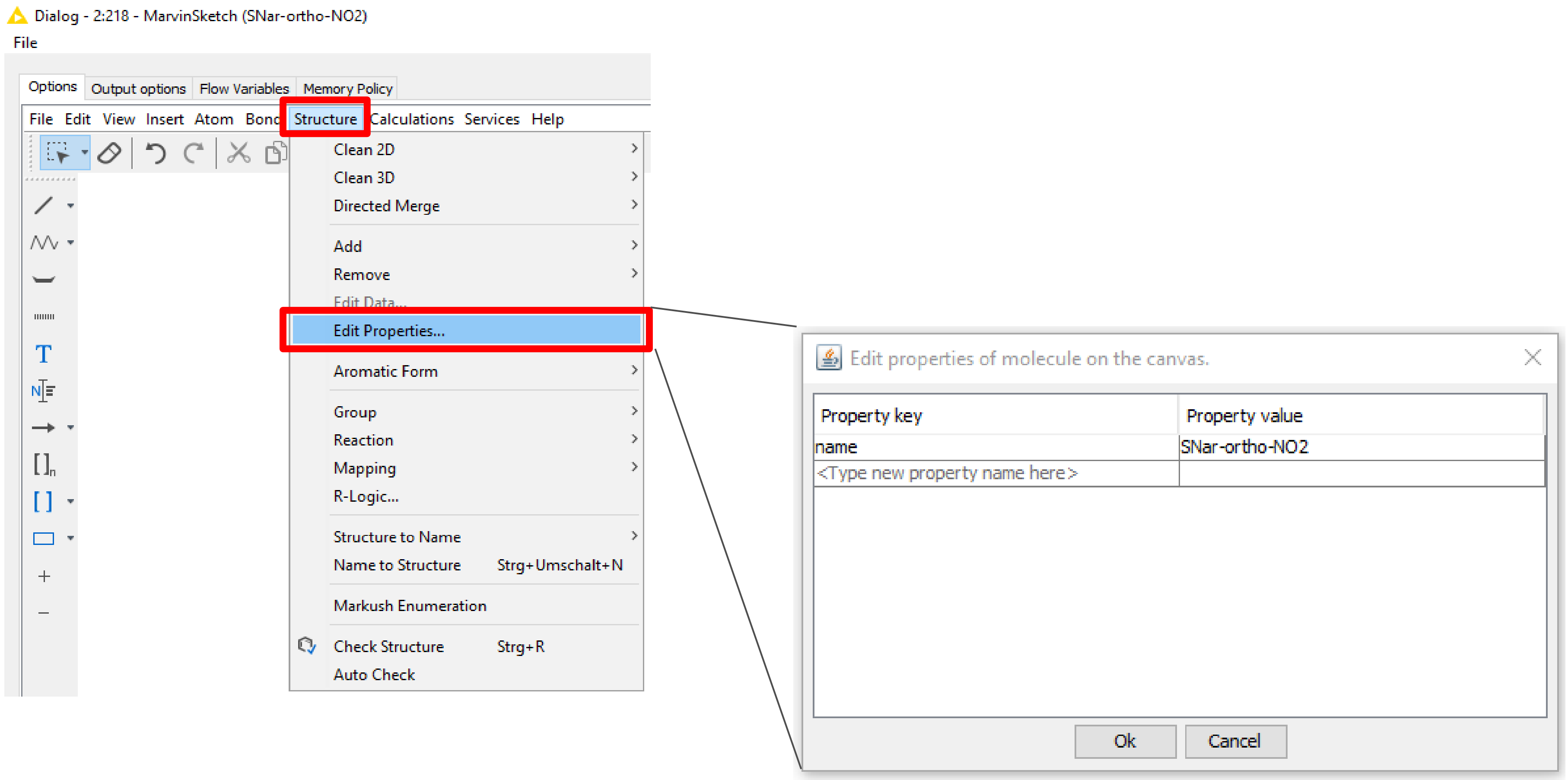
Third, we have to make sure, that c10 is unsubstituted, otherwise the coupling at this position would be impossible. The SMARTS [cD2] ensures that the aromatic carbon is connected to two other heavy atoms only:



The same SMARTS was added to c3 as this is also demanded by the reaction protocol. The fourth and last thing to do is to define the OH of the carboxylic acid to be terminal and not the substructure of an ester. This is done using the SMARTS [OD1] for the hydroxyl O and leads to the final definition:



We recommend to pass the reactions name as a property. This allows you later to see which reaction a virtual product came from. Add the name as a property:



Property key	Property value
name	SNar-ortho-NO2
<Type new property name here>	

In the popup window double-click in the fields add “name” as property key and fill a name in the value field.
Attention: avoid anything that is not on an standard english keyboard, e.g. ä, ö, ü…

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 neighbours!