-7.5909 -1.6883 -0.4896 C
-9.9889 -0.1947 -0.3218 O
-10.2795 2.5700 0.1393 C
-8.5096 -1.8624 -0.6298 H
-7.5171 -2.3313 0.2821 H
-7.0177 -1.8762 -1.3207 H
CH_3
O
$CH_2CHCH_2CH_2CCl + OPCl_3 + HCl$

Extraction of structural information from ChemDraw CDX files: easy, or an underestimated, difficult challenge?

Josef Eiblmaier, Hans Kraut, Sascha Hausberg, Peter Loew

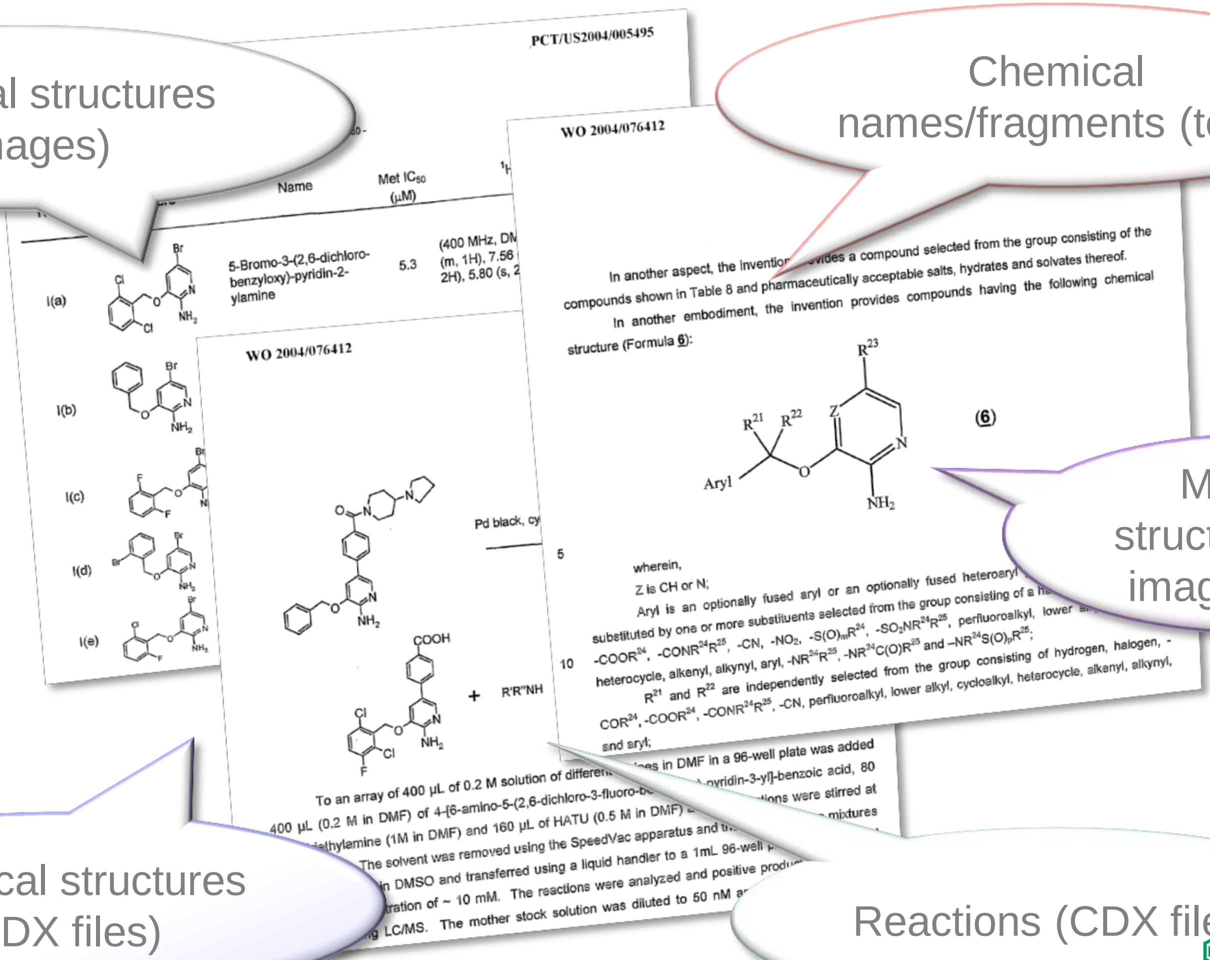
ICIC 2013 Vienna, October 13 – 16

Outline

- » ChemDraw files:
  Relevance and the Challenge
- » Approach
- » Projects
  - » InfoChem *ChemProspector*
  - » Wiley *Smart Article*
  - » Thieme *Science of Synthesis Update / Pharmaceutical Substances*
- » Conclusion / Outlook

© cora / PIXELIO, www.pixelio.de

Patents, Journal Articles and MRW's: a Buried Treasure?



PCT/US2004/005495

		Name	Met IC50 (µM)	1H
I(a)		5-Bromo-3-(2,6-dichloro-benzyloxy)-pyridin-2-ylamine	5.3	(400 MHz, DM (m, 1H), 7.56 2H), 5.80 (s, 2
I(b)				
I(c)				
I(d)				
I(e)				

WO 2004/076412

Pd black, cy

R'R"NH

To an array of 400 µL of 0.2 M solution of differen… …nes in DMF in a 96-well plate was added 400 µL (0.2 M in DMF) of 4-[6-amino-5-(2,6-dichloro-3-fluoro-be… …-pyridin-3-yl]-benzoic acid, 80 …iethylamine (1M in DMF) and 160 µL of HATU (0.5 M in DMF) … …tions were stirred at … The solvent was removed using the SpeedVac apparatus and th… …mixtures … in DMSO and transferred using a liquid handler to a 1mL 96-well p… …tration of ~ 10 mM. The reactions were analyzed and positive produ… …g LC/MS. The mother stock solution was diluted to 50 nM a…

WO 2004/076412

In another aspect, the inventio… …vides a compound selected from the group consisting of the compounds shown in Table 8 and pharmaceutically acceptable salts, hydrates and solvates thereof.

In another embodiment, the invention provides compounds having the following chemical structure (Formula 6):

(6)

wherein,

Z is CH or N;

Aryl is an optionally fused aryl or an optionally fused heteroaryl … substituted by one or more substituents selected from the group consisting of a ha… perfluoroalkyl, lower al… $-COOR^{24}$, $-CONR^{24}R^{25}$, $-CN$, $-NO_2$, $-S(O)_mR^{24}$, $-SO_2NR^{24}R^{25}$, perfluoroalkyl, lower al… $-COOR^{24}$, $-CONR^{24}R^{25}$, $-CN$, $-NO_2$, $-S(O)_mR^{24}$, $-SO_2NR^{24}R^{25}$, heterocycle, alkenyl, alkynyl, aryl, $-NR^{24}R^{25}$, $-NR^{24}C(O)R^{25}$ and $-NR^{24}S(O)_pR^{25}$;

R^{21} and R^{22} are independently selected from the group consisting of hydrogen, halogen, -COR^{24}, $-COOR^{24}$, $-CONR^{24}R^{25}$, -CN, perfluoroalkyl, lower alkyl, cycloalkyl, heterocycle, alkenyl, alkynyl, and aryl;

Chemical structures (images)

Chemical names/fragments (text)

Markush structures (text, images, CDX)

Chemical structures (CDX files)

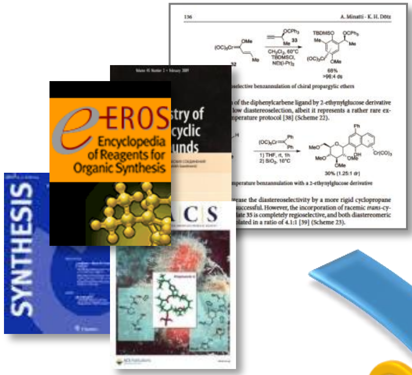
Reactions (CDX files)

Manuscript → Article → Database …

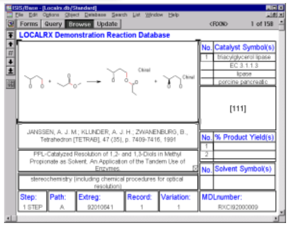
Manuscript submission

Thieme
Springer
science+business media
ACS
RSC Advancing the Chemical Sciences
WILEY Publishers Since 1807
ELSEVIER

Publishing



Manual Indexing


ISIS/Base - [Localrx.db/Standard]
File Edit Options Object Database Search List Window Help
Forms Query Browse Update
<RXN>
1 of 158
Chiral
Chiral

LOCALRX Demonstration Reaction Database

No.	Catalyst Symbol(s)
1	triacylglycerol lipase
	EC 3.1.1.3
	lipase
	porcine pancreatic

[111]

JANSSEN, A. J. M.; KLUNDER, A. J. H.; ZWANENBURG, B., Tetrahedron [TETRAB], 47 (35), p. 7409-7416, 1991

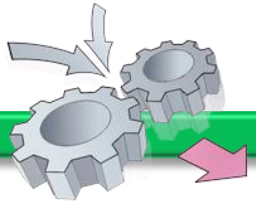
PPL-Catalyzed Resolution of 1,2- and 1,3-Diols in Methyl Propionate as Solvent. An Application of the Tandem Use of Enzymes

stereochemistry (including chemical procedures for optical resolution)

No.	% Product Yield(s)
1	
2	

No.	Solvent Symbol(s)

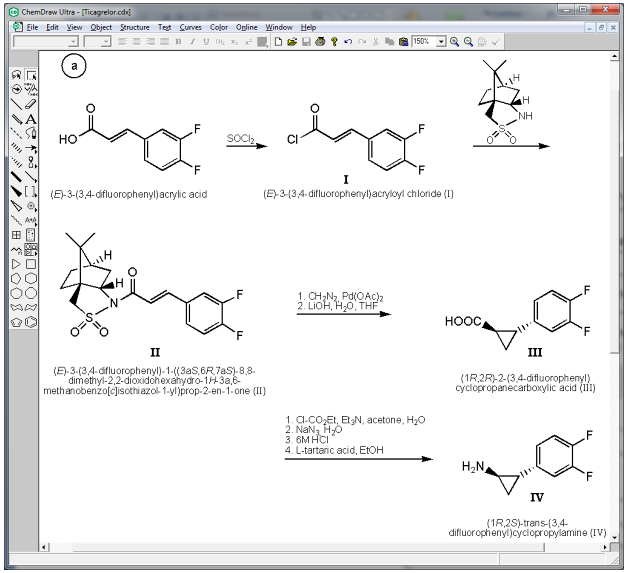
Step:	Path:	Extreg:	Record:	Variation:	MDLnumber:
1 STEP	A	92010641	1	1	RXCI92000009



Database production e.g. SciFinder, Reaxys, SPRESI eEROS, ...

CDX Scheme vs. Database Record

ChemDraw file	Database
Purpose: presentation / publishing no search	Purpose: search / retrieval
Unstructured	Structured
Structures: no strict rules	Structures: strict rules
General rules: none	Database rules: strict



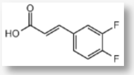
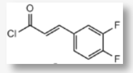
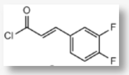
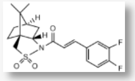
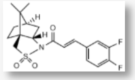
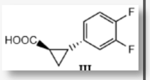
(*E*)-3-(3,4-difluorophenyl)acrylic acid

(*E*)-3-(3,4-difluorophenyl)acryloyl chloride (I)

(*E*)-3-(3,4-difluorophenyl)-1-((3a*S*,6*R*,7a*S*)-8,8-dimethyl-2,2-dioxidohexahydro-1*H*-3a,6-methanobenzo[*c*]isothiazol-1-yl)prop-2-en-1-one (II)

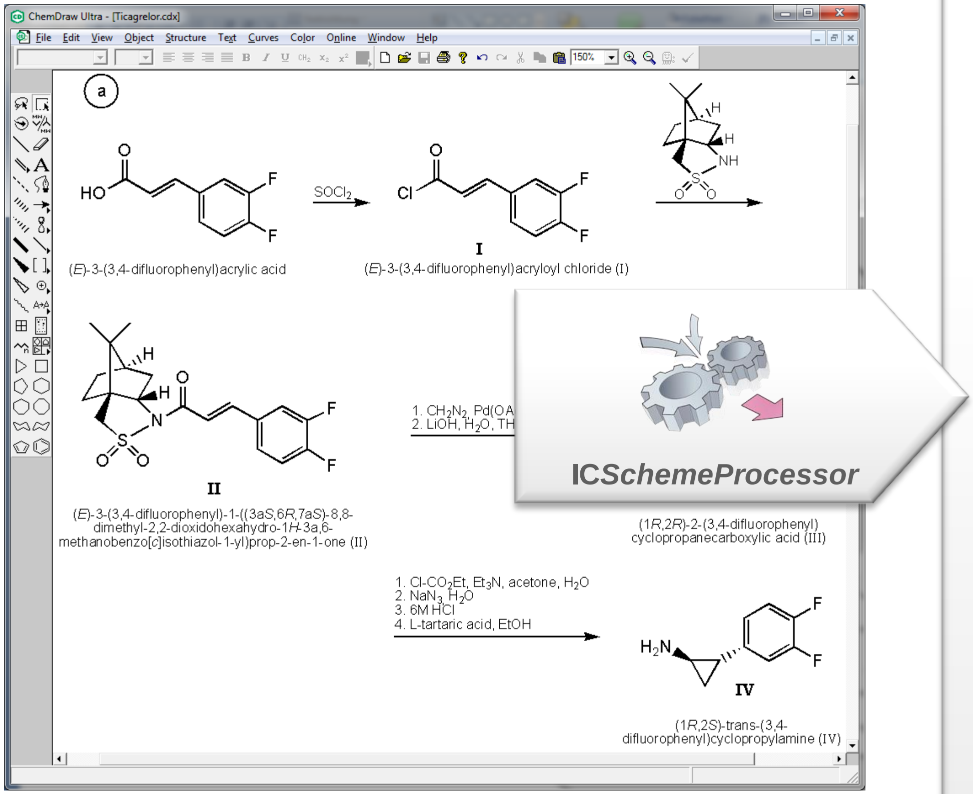
(1*R*,2*R*)-2-(3,4-difluorophenyl) cyclopropanecarboxylic acid (III)

(1*R*,2*S*)-trans-(3,4-difluorophenyl)cyclopropylamine (IV)

Reactant	Product	Reagent	Solvent	Catalyst
		$SOCl_2$		
		LiOH	H_2O, THF	$Pd(OAc)_2$
		Cl-Co2Et, Et_3N	Acetone, H_2O	

Source: Thieme *Pharmaceutical Substances*, Ticagrelor (in production)

CDX Scheme Processing, what does that mean?


ChemDraw Ultra - [Ticagrelor.cdx]
File Edit View Object Structure Text Curves Color Online Window Help
150%
a
HO
O
F
F
SOCl$_2$
Cl
I
H
NH
O S O
II
N
1. CH_2N_2, Pd(OA
2. LiOH, H_2O, TH
ICSchemeProcessor
H_2N
IV

(*E*)-3-(3,4-difluorophenyl)acrylic acid

(*E*)-3-(3,4-difluorophenyl)acryloyl chloride (I)

(*E*)-3-(3,4-difluorophenyl)-1-((3a*S*,6*R*,7a*S*)-8,8-dimethyl-2,2-dioxidohexahydro-1*H*-3a,6-methanobenzo[*c*]isothiazol-1-yl)prop-2-en-1-one (II)

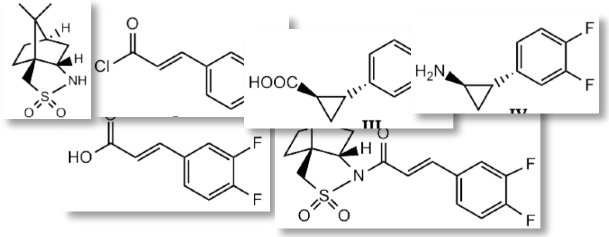
(1*R*,2*R*)-2-(3,4-difluorophenyl) cyclopropanecarboxylic acid (III)

1. Cl-CO_2Et, Et_3N, acetone, H_2O
2. NaN_3, H_2O
3. 6M HCl
4. L-tartaric acid, EtOH

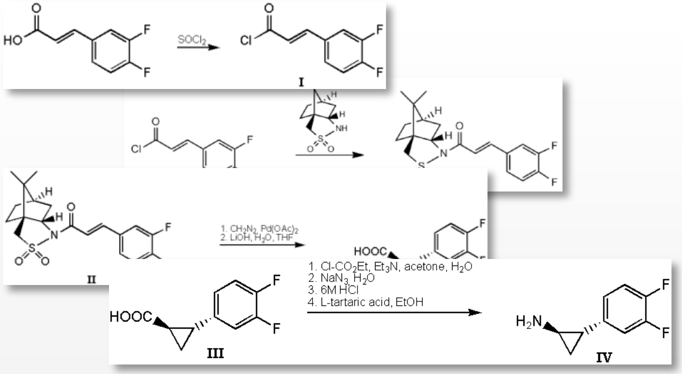
(1*R*,2*S*)-trans-(3,4-difluorophenyl)cyclopropylamine (IV)

Source: Thieme *Pharmaceutical Substances*, Ticagrelor (in production)

Chemical structures (SD files)



Reactions (RD files)



Conditions (RD files)

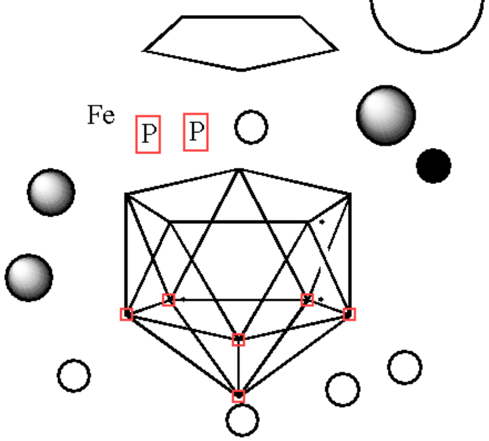
Reagent	Solvent	Catalyst
$SOCl_2$		
LiOH	H_2O, THF	$Pd(OAc)_2$
Cl-Co2Et, Et_3N	Acetone, H_2O	

But: CDX files, often an optical illusion! Authors are very inventive for a ‚perfect‘ layout! Appearences are deceiving!

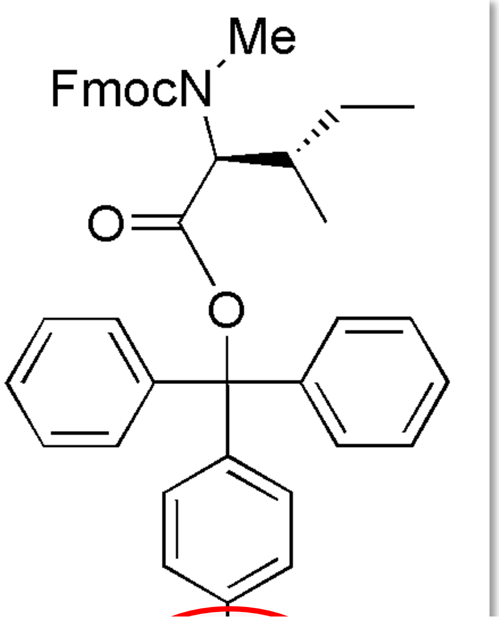
» **Usage of graphical symbols**

- Polymer supports
- Heteroatoms

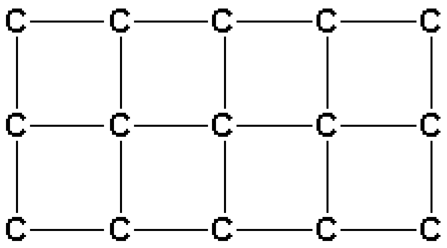
[1-Cp-*closo*-1,2,3,5-$FeP_2CB_8H_9$]



$C_6H_{14}B_8FeP_2$



C Grid:

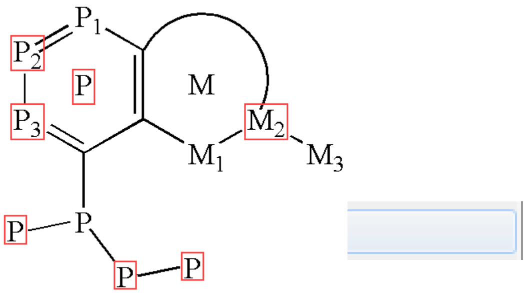


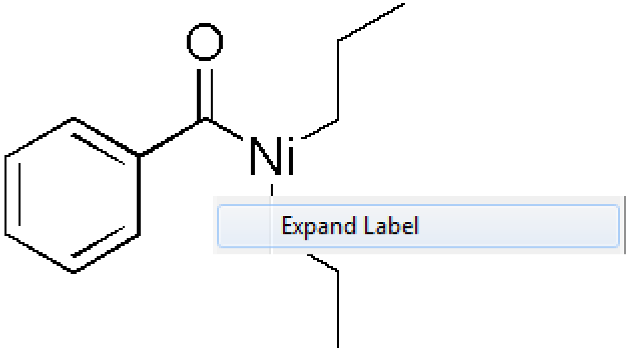
r chloride SynPhase
ally available trityl a
ol/lantern) with Ac
lethylamine (DIEA)/CH_2Cl_2 (Scheme 2).[12]
of **13** from the polymer-supported **19** with
Cl_2, the loading amount of **13** was deter-
33 mmol per lantern. After removal of the
ith 20% piperidine/DMF acylation with

Optical illusions 2

» Unresolvable labels

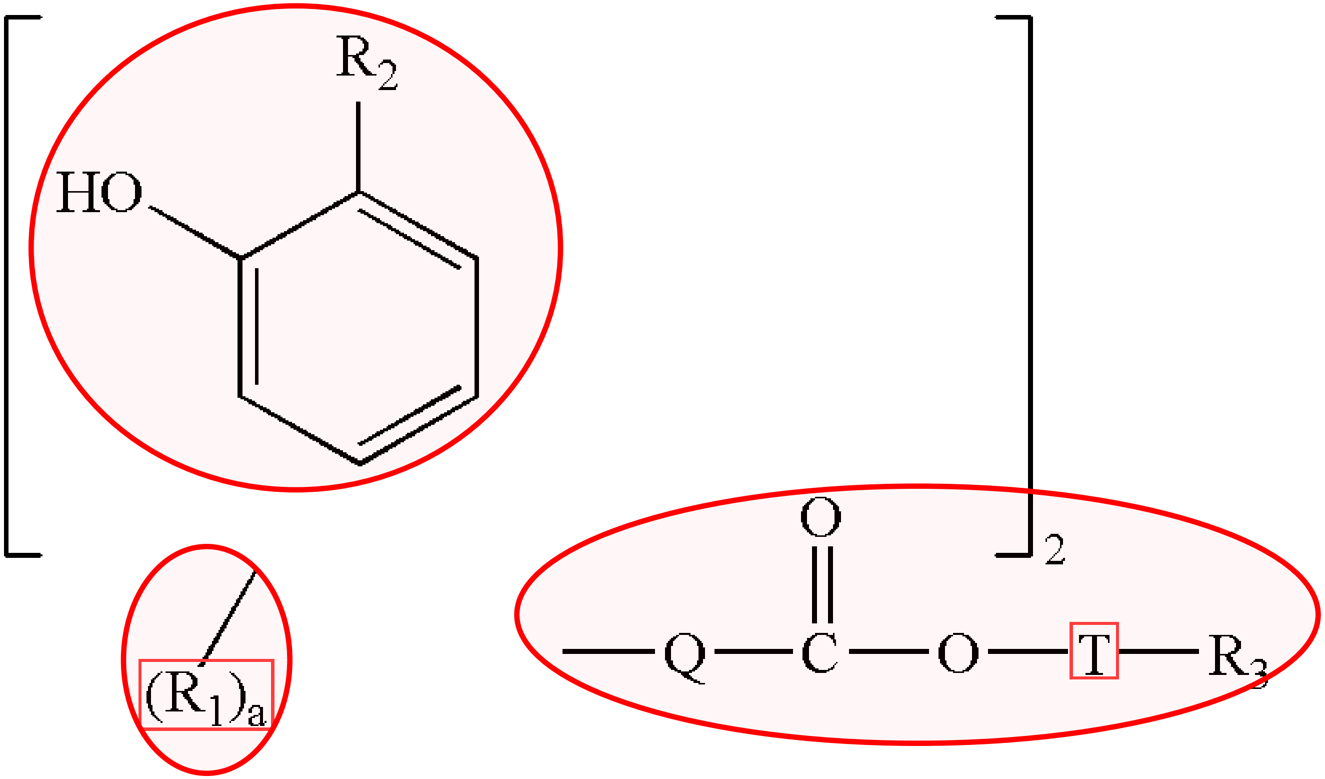
- Labels not defined
- Element symbols used as R-group labels
- Ambiguous fragment labels (e.g. molecular formula)





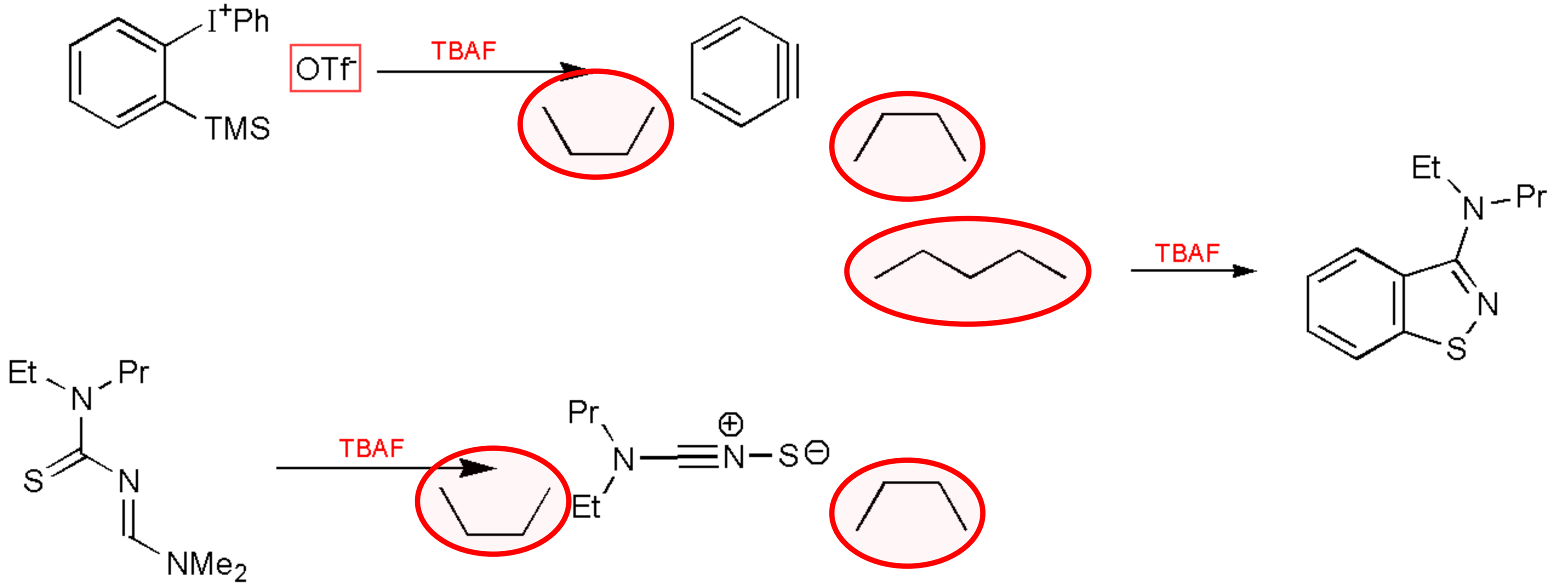
Optical illusions 3

» Variable points of attachment

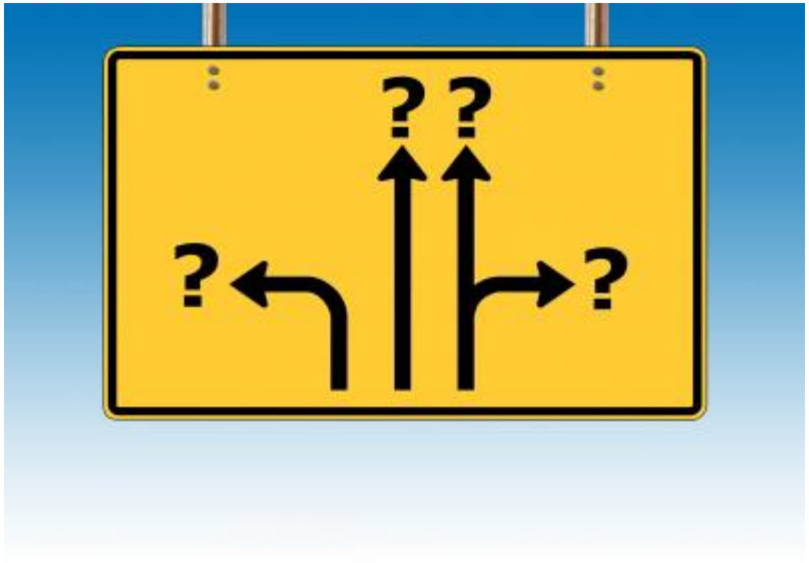


Optical illusions 4

» Reaction arrows / forked arrows / brackets



Approach



© Gerd Altmann / PIXELIO, www.pixelio.de

Approach

» The algorithmic approach:

- Application of a set of rules in the software (generic, project unspecific). Software should recognize all cases that might occur!
- project (title-) specific rules (drawing conventions must not change), otherwise further development necessary
- manual post correction required (cost/time intensive)
- problem is infinite, unprecedented issues can not be handled

» The templating approach:

- software is developed to recognize a defined set of problems (PS)
- all content must be manually pre-templated (cost intensive) according to the capabilities of the software

» The hybrid approach:

- depending on the source the focus can be laid on either approach

Templating

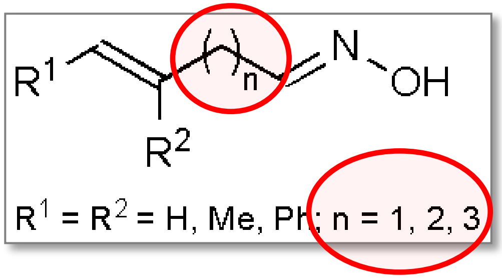
» ***Templating*: Guidelines for authors and typesetters**

- Syntax definitions for tables, R-groups etc.
- Syntax rules for captions
- Reaction arrangement, forked arrows
- Rules for reaction conditions

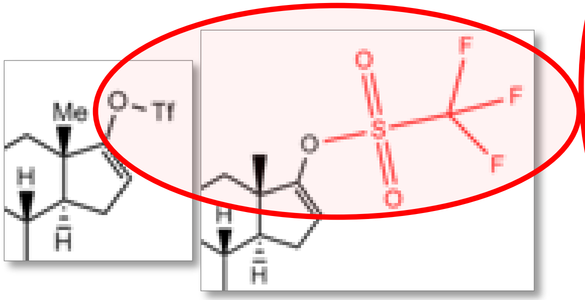
  (reactants, catalysts, solvents, yields, temperature)

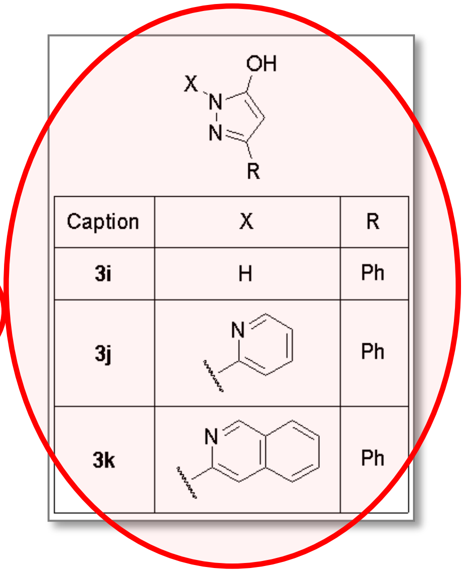
Examples:

- » **Algorithmic detection of features**
  - » Resolution of repeating groups
  - » Enumeration of R-groups
  - » Resolution of aliases/labels
    - source specific alias databases
    - continuously extended
  - » Table Enumeration
    - compound enumeration
    - reaction factual data: Caption/Yield
  - » Variable points of attachment
  - » Forked arrows

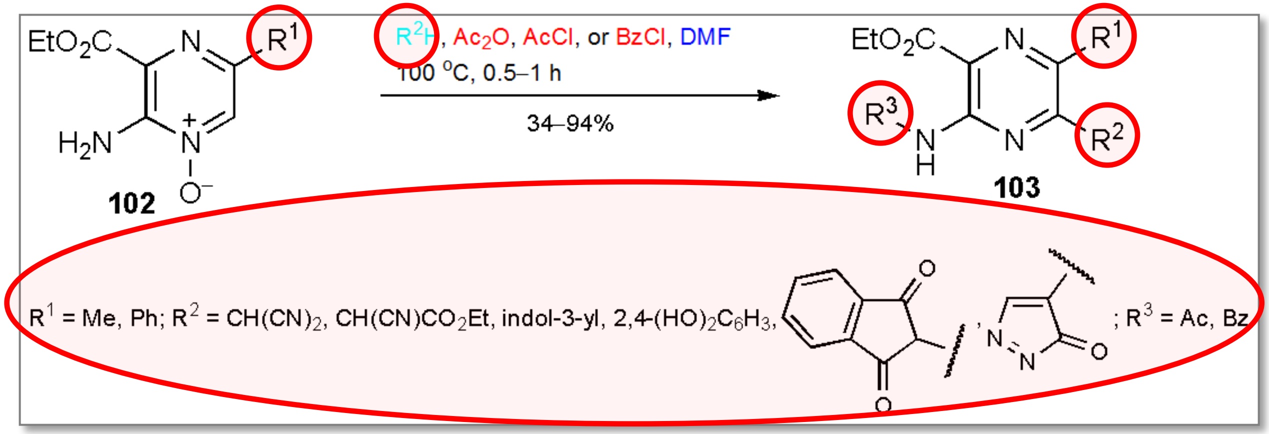


$R^1 = R^2 = H, Me, Ph; n = 1, 2, 3$





Caption	X	R
3i	H	Ph
3j	N	Ph
3k	N	Ph



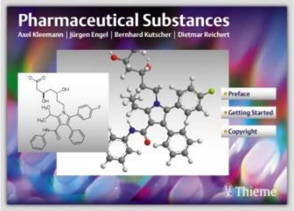
R^1 = Me, Ph; R^2 = $CH(CN)_2$, $CH(CN)CO_2Et$, indol-3-yl, 2,4-$(HO)_2C_6H_3$, , ; R^3 = Ac, Bz

Projects

ChemProspector

Science of Synthesis

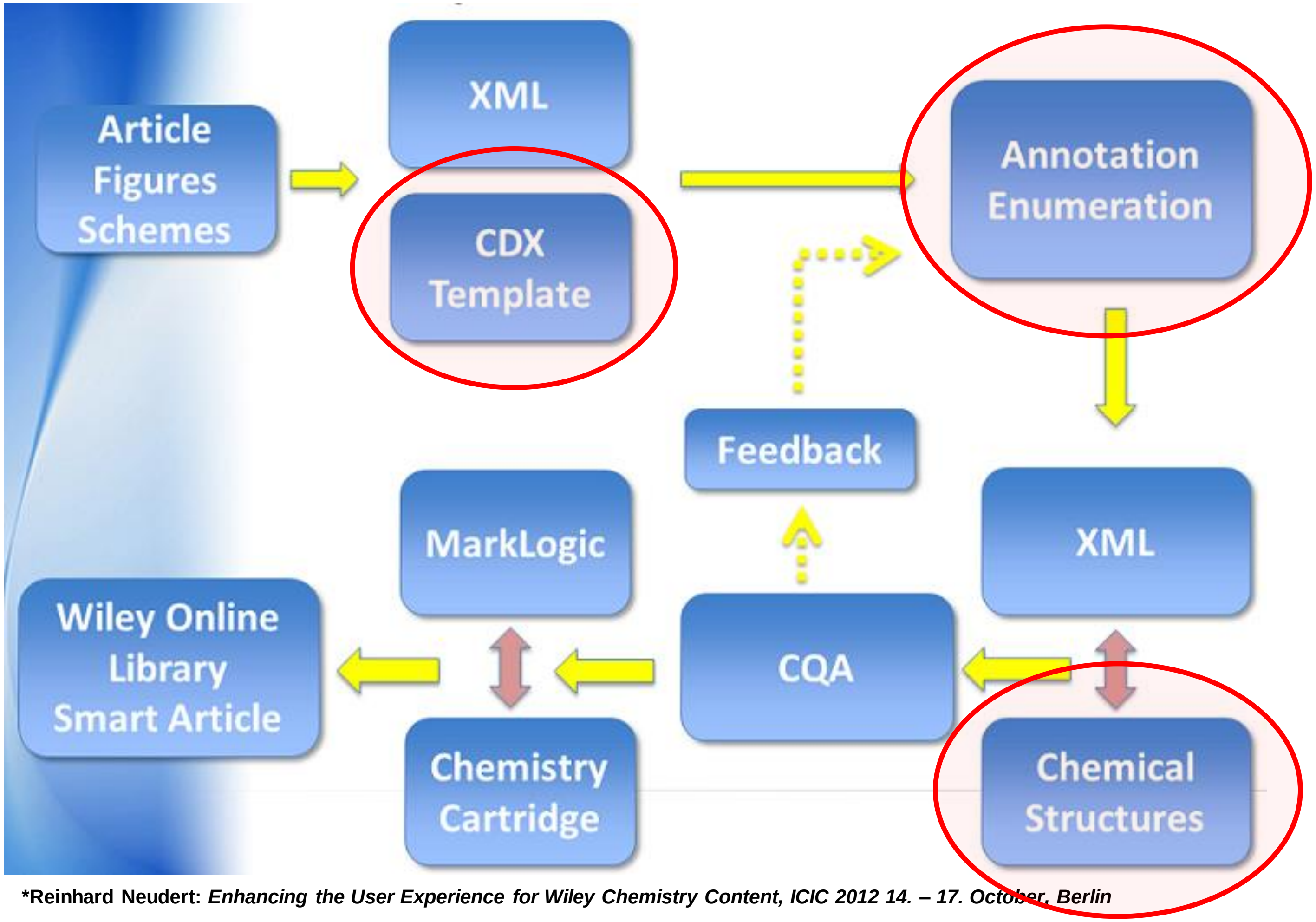
THE SMART ARTICLE
FASTER, SMARTER, CONTENT



Pharmaceutical Substances

Axel Kleemann | Jürgen Engel | Bernhard Kutscher | Dietmar Reichert

Sucessful Application of CDX Processing: Chemistry Enrichment Workflow*, (Wiley *Smart Article*)



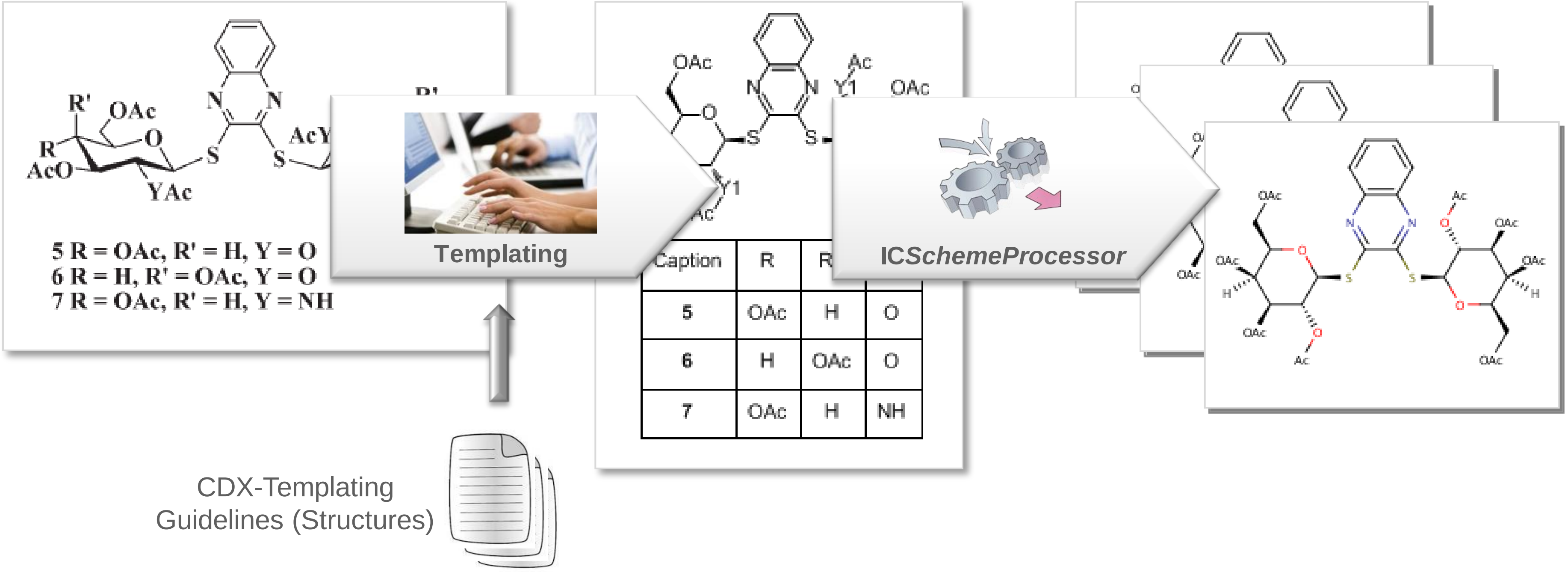
*Reinhard Neudert: *Enhancing the User Experience for Wiley Chemistry Content, ICIC 2012 14. – 17. October, Berlin*

Templating*

Author‘s CDX File

CDX Template

Enumerated structures



5 R = OAc, R' = H, Y = O
6 R = H, R' = OAc, Y = O
7 R = OAc, R' = H, Y = NH

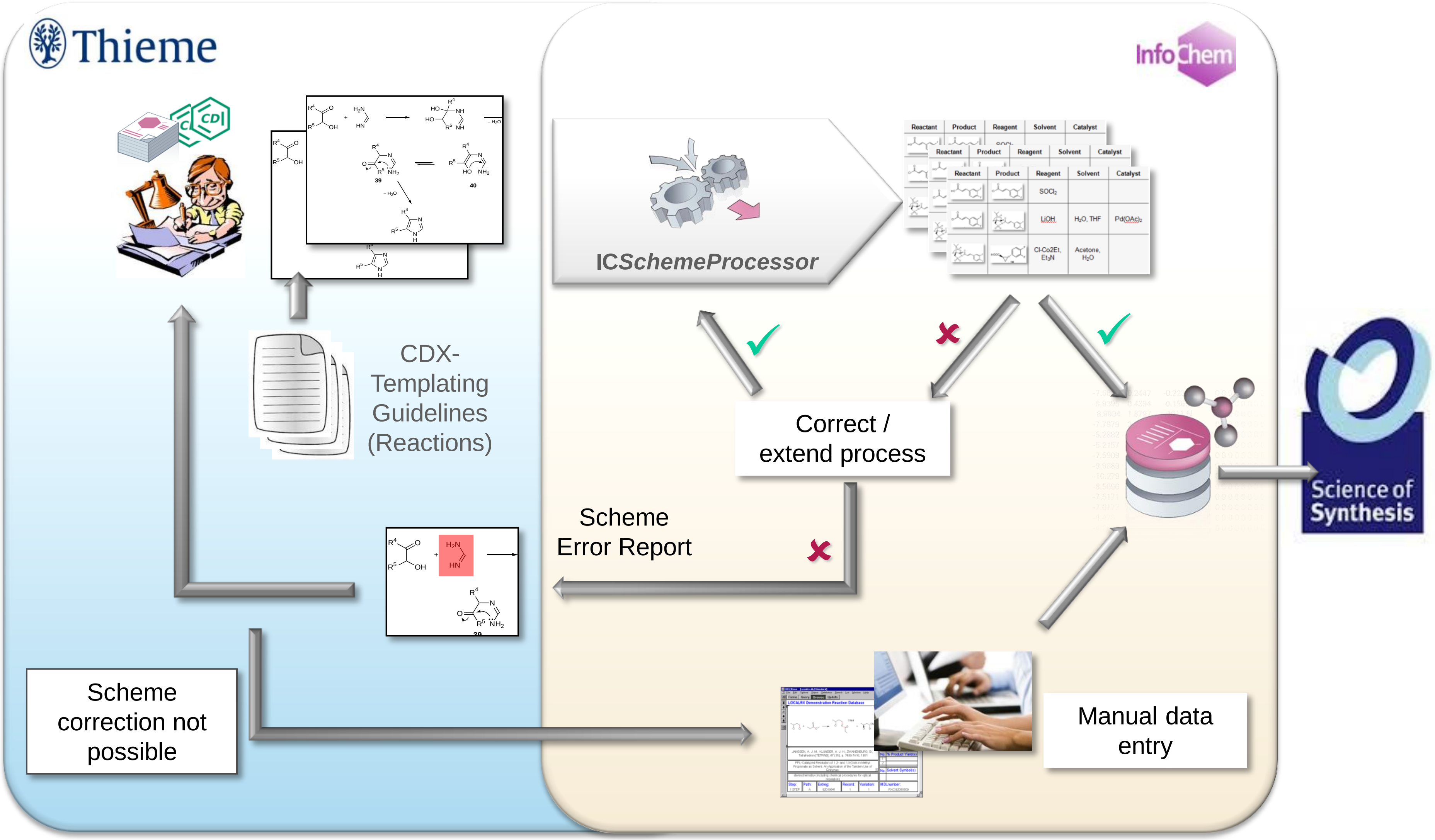
Templating

Caption	R	R	
5	OAc	H	O
6	H	OAc	O
7	OAc	H	NH

CDX-Templating
Guidelines (Structures)

*Reinhard Neudert: *Enhancing the User Experience for Wiley Chemistry Content, ICIC 2012 14. – 17. October, Berlin*

Workflow *Science of Synthesis Update*



CDX-Templating Guidelines (Reactions)

ICSchemeProcessor

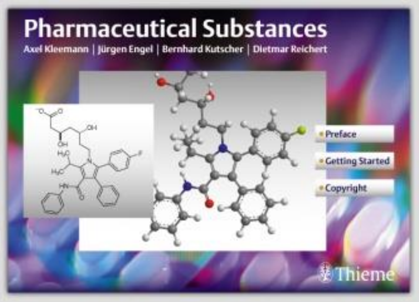
Correct / extend process

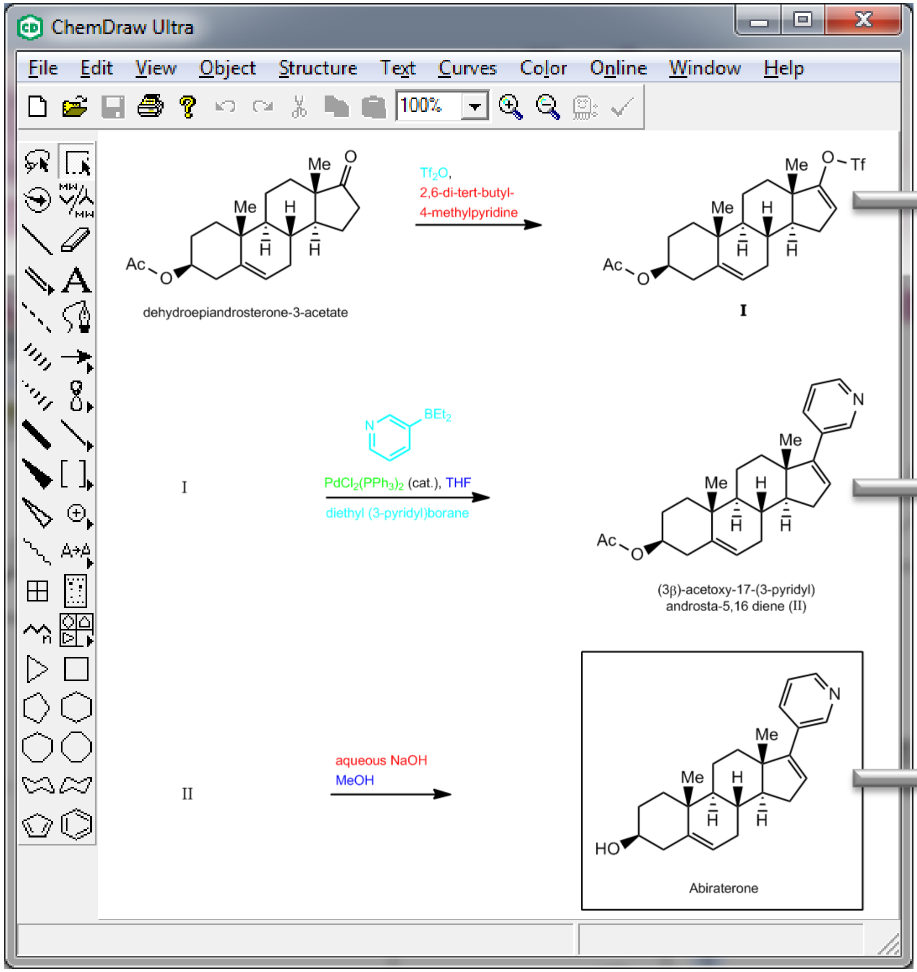
Scheme Error Report

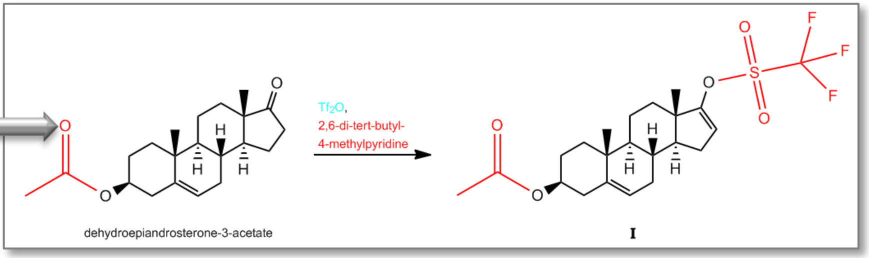
Scheme correction not possible

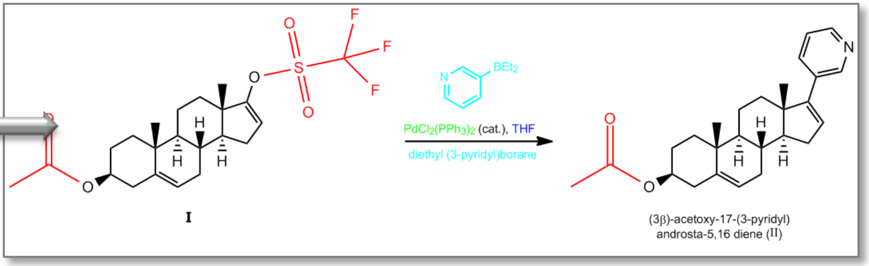
Manual data entry

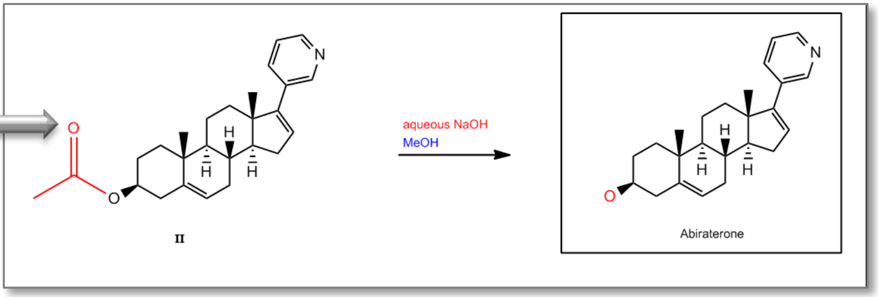
Sample Pharmaceutical Substances *Update*


Pharmaceutical Substances
Axel Kleemann | Jürgen Engel | Bernhard Kutscher | Dietmar Reichert
Preface
Getting Started
Copyright
Thieme









Source: Thieme *Pharmaceutical Substances*, Abiraterone

Conclusion

- » **As much as possible algorithmic processing desirable**
  - generic: can be applied to other contents as well
  - cheaper (humans cost!)
- » **100% conversion (without human interaction) never possible**
- » **Solutions are project / source specific**
- » **Relevance of automatic extraction will continuously increase**
- » **Authors / Publishers play an essential role in a successful conversion**

Acknowledgements

- » **Wiley**
  - Michael Forster
  - Reinhard Neudert
- » **Thieme**
  - Guido Herrmann
  - Rolf Hoppe
  - Klaus Köberlein
- » **InfoChem**
  - Hans Kraut, Sascha Hausberg, Thomas Menke, Manuela Rauh

    Fanny Irlinger, Huyen Ngyen, Dagmar Kunzmann

© Thomas Link / Flickr

Thank you!

Questions?