11.09.2024

RDKit-powered Chemical Registration

In the Flavors and Fragrances Industry

Givaudan

Human by nature

Agenda

1 Industry Introduction

Givaudan and it's Flavor and Fragrances divisions and what really small molecules are

2 History

Where we are coming from and how that impacted our needs

3 Data Migration

Cleaning, normalizing and converting legacy data

4 Registration Process

Demonstrating the step-by-step process for registering a new chemical compound

5 Core System Capabilities

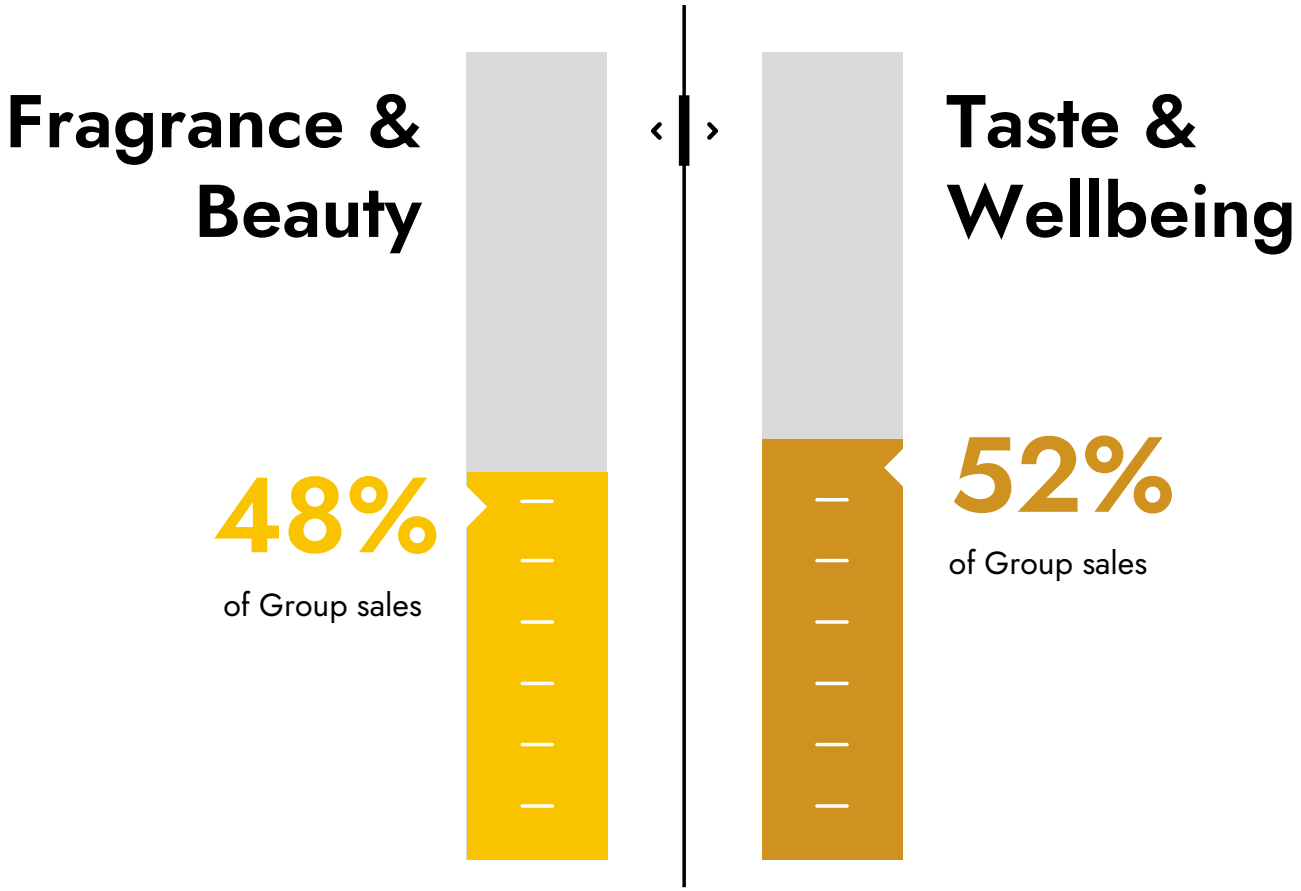
Highlighting key features such as stereo chemistry handling, mixture management, and duplicate checking.

Industry Introduction

Taste & Wellbeing and Fragrance & Beauty

Givaudan

Fragrances and Flavours drive consumers' product choices
and balanced sales across our two divisions



Givaudan

Always by your side

163

Locations worldwide

64

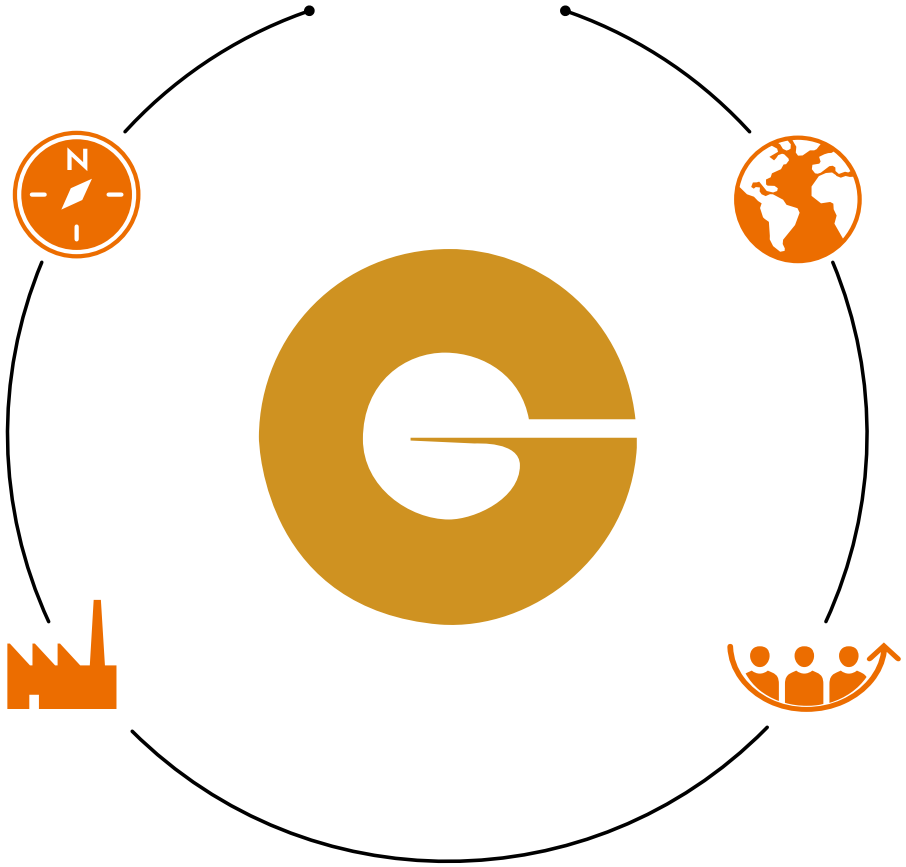
Creation & research centres

78

Production sites

16,260

Full time employees



Givaudan

Taste & Wellbeing

Consumption / Ingestion

- Strongly regulated what compounds can be used in what quantities
- Can't easily introduce new synthetic molecules
- Salts play a role
- Focus on naturals and nature-derived compounds
- Mixtures (simple and complex)
- May need approval before taste evaluation

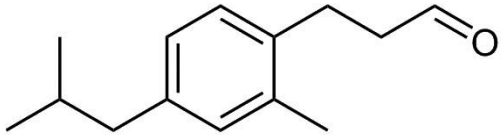
Fragrance & Beauty

Topical Application

- Regulations regarding human and environmental toxicology
- May contain safe synthetic molecules
- No Salts (not volatile)
- Renewable & biodegradable synthetics & naturals
- Mixtures (simple and complex)
- Can be immediately smelled (no assay needed)

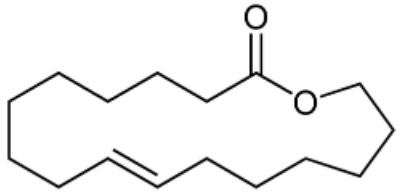
Fragrance & Beauty: Example Molecules

Very small, usually just a single functional group



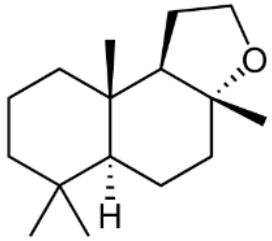
Nympheal™

Diffusive floral cyclamen muguet note with green, watery and linden blossom facets



Ambrettolide™

A macrocyclic musk with an exceptional diffusion and a very fine character



Ambrofix™

A highly powerful, highly substantive and highly stable ambery note.

History

Where we came from and how this impacted our needs

History

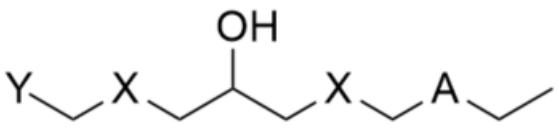
Technology stack of our old Chemical Registration

- ChemBioOffice Enterprise from CambridgeSoft (later PerkinElmer now Revvity)
- Based on ASP (not ASP.Net), a technology from the 90s
- UI is only properly displayed in Internet Explorer 6 compatibility mode in MS Edge
  - Chemical structure display requires ChemDraw ActiveX plugin locally installed
- Oracle 10g Database
  - Outdated and costly

History

Limitations of the old system

- No possibility to validate drawn structures
  - Any valid ChemDraw drawing can be saved
  - No guarantee it can ever be found again with a structure search
  - No control over ChemDraw style (bond length, label size,…)



- Limitations in regards to security and access rights
  - Admins required in many cases for simple updates
  - No tracking of changes

- No API and hence no automation
  - Manual registrations with copy & paste form ELN!!!

- Limited structure conversion to smiles, molfile or inchi
  - Very slow, and often issues in regards to inchi

Data Migration

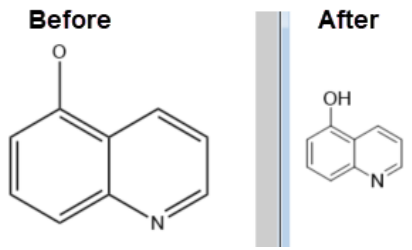
clean, normalize, convert

Data Migration

clean, normalize, convert

Targets of the data migration:

- Normalize all entries to the same ChemDraw Style
  - Original drawing is stored as cdxml (converted from cdx) including for new entries
  - There is no official tool for style normalization -> PyCDXML see my lightning talk from UGM 2022

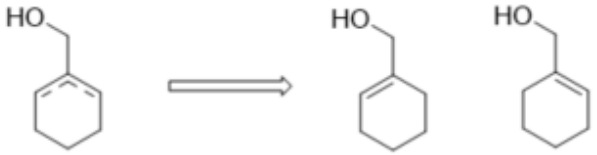


- Format conversions for RDKit cartridge compatibility

- Separation of mixtures (multiple structures) into single structures for searching

- Ensure as many records as possible are valid for RDKit and are therefore structure searchable

- Manual cleaning of problematic drawings
  - Search for likely known issues beforehand
  - Distribute workload over multiple people
  - Examples:
    - Changing single bonds to dative bonds
    - “single or double” bonds changed to multiple explicitly drawn molecules



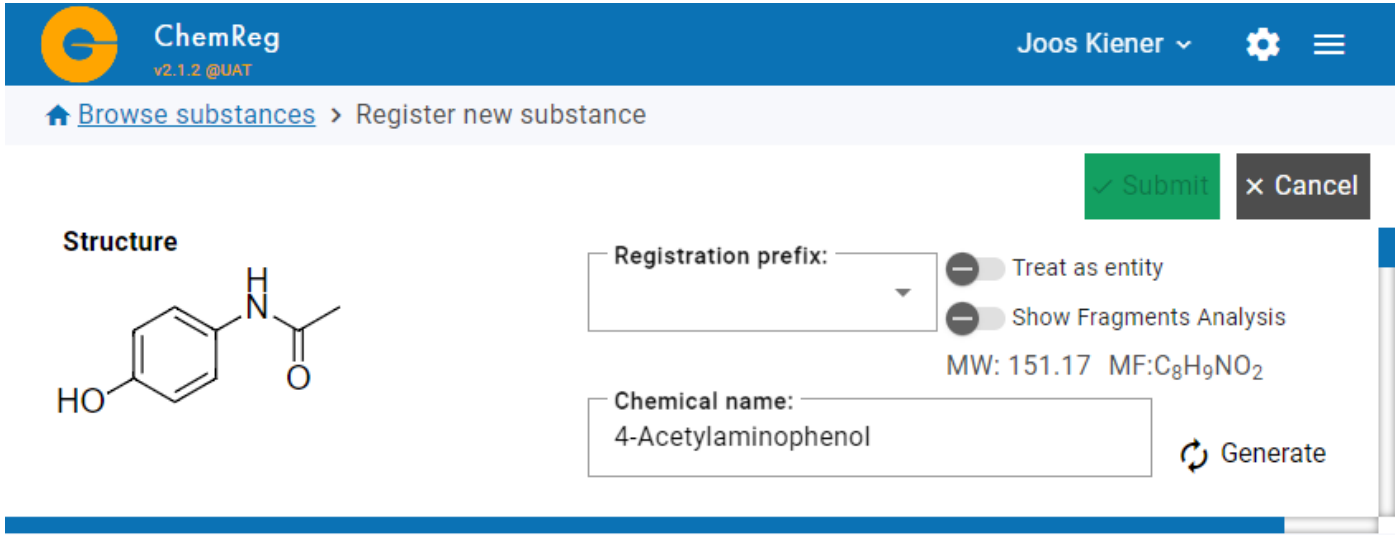
Registration Process

Registration Process

Birds-View

Register a Sample in Signals Notebook:

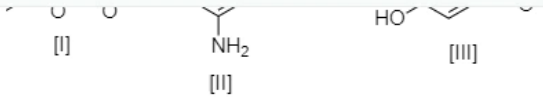
New browser tab opens



Givaudan

Experiment Contents

- Reaction
  - Preparation
  - ChemDraw Document
  - Samples Table
- Analysis



Quick Add

Reactants

Rxn ID	Reactant	GR-Number	Inventory Barcode	CAS Number	MF	MW	Limit?	Eq	Sample Mass	Moles	Molarity	Vol	d	% Wt	Supplier
I	acetic anhydride			108-24-7	$C_4H_6O_3$	102.09	✓	1					1.082 g/ml		
II	4-aminophenol				C_6H_7NO	109.13		1							

Products

Rxn ID	Product	MF	MW	Theo Mass	Actual Mass	Purity	d	Yield	Theo Mol	Actual Mol
III	4-Acetylaminophenol	$C_8H_9NO_2$	151.16							

Solvents

Solvent	Volume	Supplier	CAS

Reaction Conditions

Pressure	Temperature	Reaction Time

Samples Table

ID	GR-Number	Amount	Purity	Physical Form	Color	Comments	Description	Material ID	Chemical Name	MF	MW	Created Date	FM
GLN02516-053-01									4-Acetylaminophenol	$C_8H_9NO_2$	151.16 g/mol	Sep 06, 2024, 08:57 AM	151.16 g/mol

Core System Capabilities

Highlighting key features of potential general interest

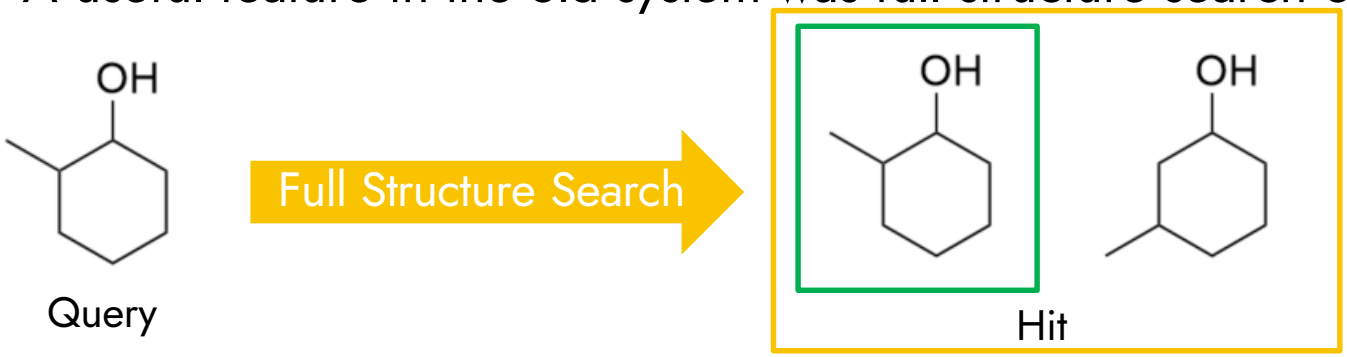
Chemical Structures
Display and formats

- The original drawing made with ChemDraw JS (CDXML) is stored and used for display
- ChemDraw JS web service component converts from CDXML to molfile v3000
- Molfile v3000 is used to create the RDKit molecule for indexing for substructure searching
- Inchi and “flat inchi” are derived from the RDKit molecule (used for look-ups)
  - In case no inchi can be made, canonical SMILES is used (for example dative bonds)
- Molecules that can’t be converted to RDKit can still be registered and displayed
  - But not searched by structure (Example: complex catalysts/organometallics)

Mixture Handling

Chemical Structure Search

- A useful feature in the old system was full structure search on individual molecules of a compound



- This does not work with the RDKit Cartridge, all components inside a molecule must match

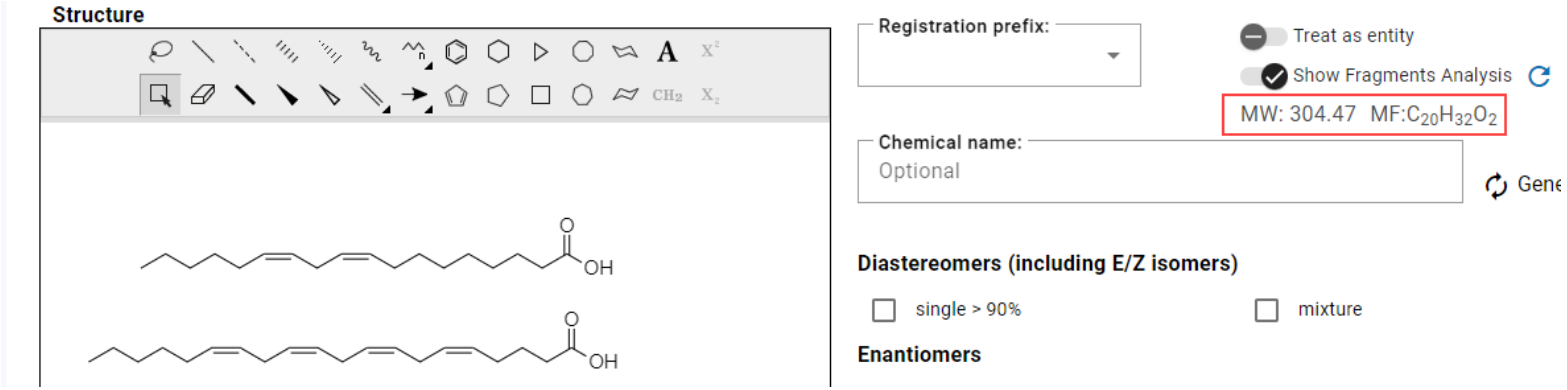
Solution:

- Mixtures are split into their individual molecules (molecules may be reused)
- Structure Search happens on the individual molecules

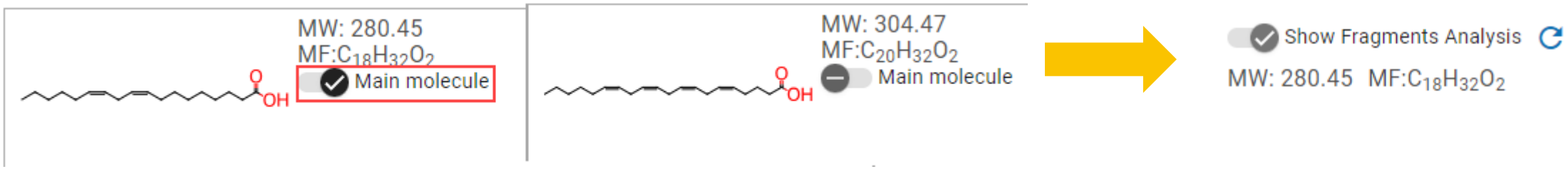
Mixture Handling

Main Component

- If multiple structures are present by default the MW and MF of the heaviest component will be taken



- In the "Fragment Analysis" this can be overridden, a Main component can be selected



- "Treat as entity" will use combined MW and MF of all drawn structures

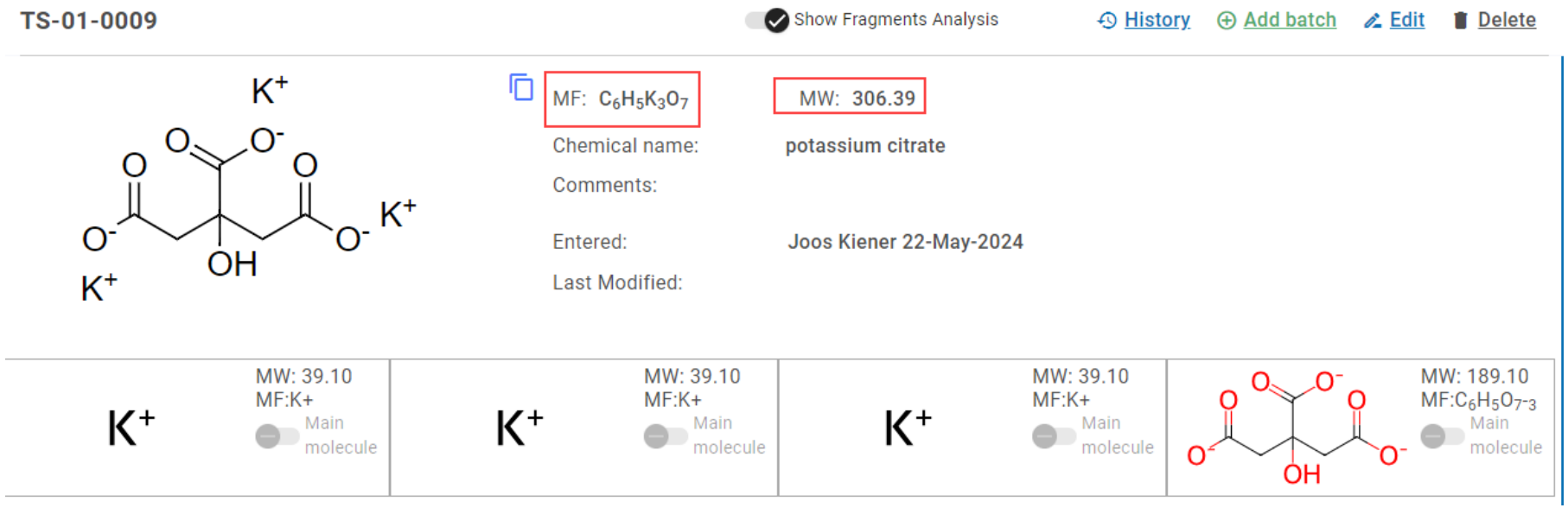
Salts

For Flavors only

- No Salt Splitting!!!
- By default treated as entity -> combined MF and MW

Why?

- Both ions matter for the taste
- Therefore they need a different registration number



Stereochemistry
Background

- Stereochemistry may greatly matter for activity

- But: Regulations allow the selling of mixtures

➡ Limited importance to have a stereo chemically "clean" product at time of registration

- At time of registration there will be limited information available in regards to stereochemistry
  - Mostly from raw materials and the reaction

- This impacts how we record stereochemistry, the duplicate checking and default search behavior

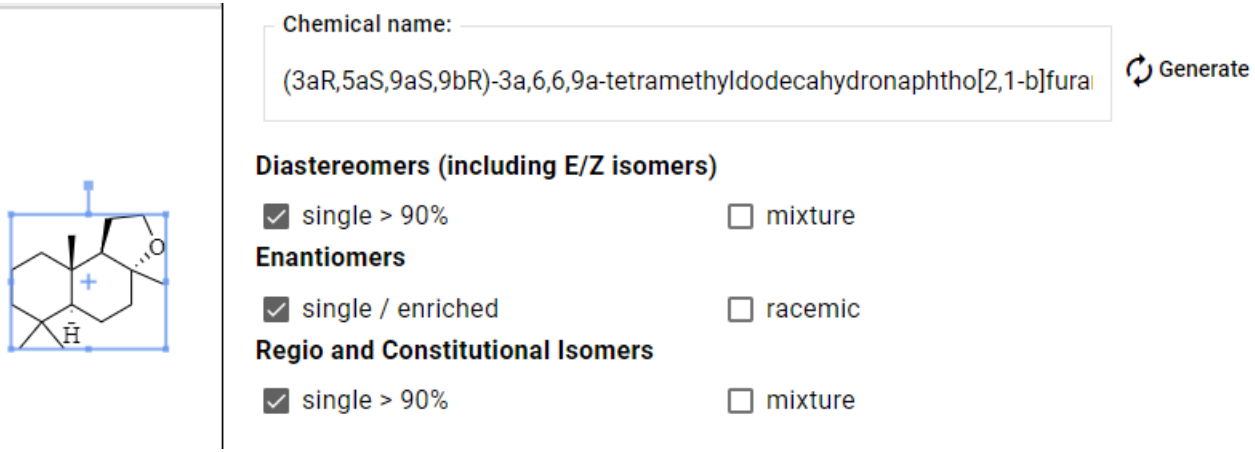
Stereochemistry

How we track the information

- Enhanced stereochemistry was deemed as “too complex”

Ways stereochemistry is tracked:

- By the drawing (drawing rules)
- By the chemical name
- By a set of check boxes
- Free-text comment



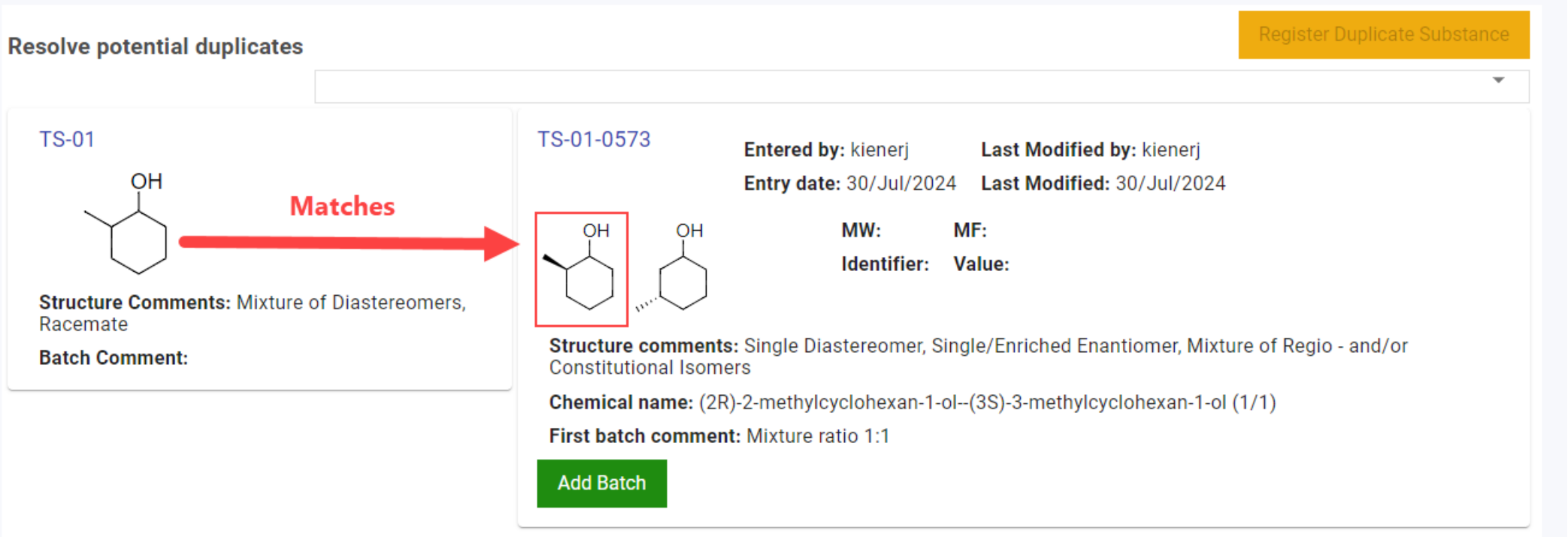
- Comment can also be used to define ratios between isomers

Batch comment:
mixture of 4 isomers (56:15:18:11)

Duplicate checking

Fuzzy

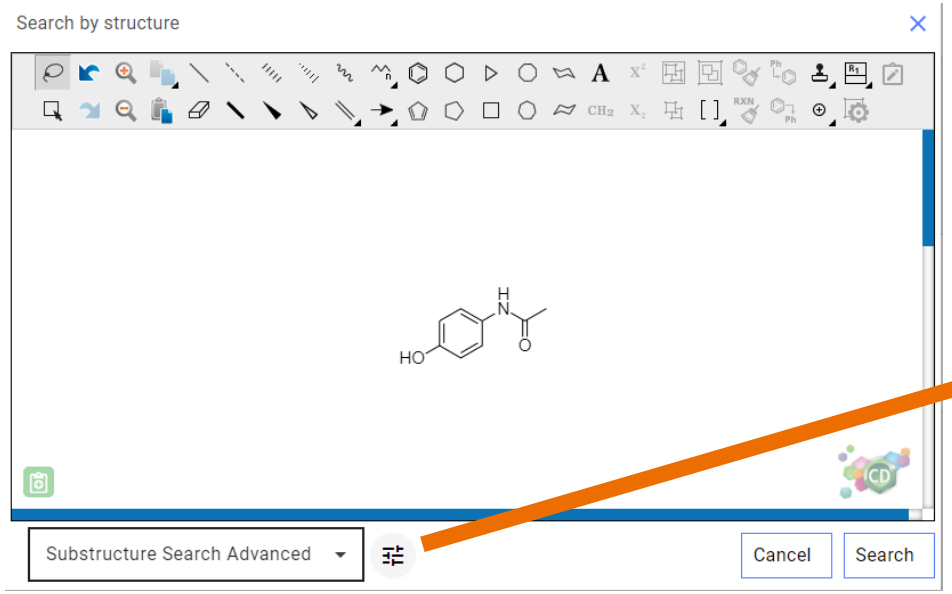
- Duplicate checking is fuzzy
- Avoid duplicates as much as possible, let the expert choose vs application trying to being too smart
- If any component matches "flat" (ignoring stereochemistry), then it is a potential duplicate

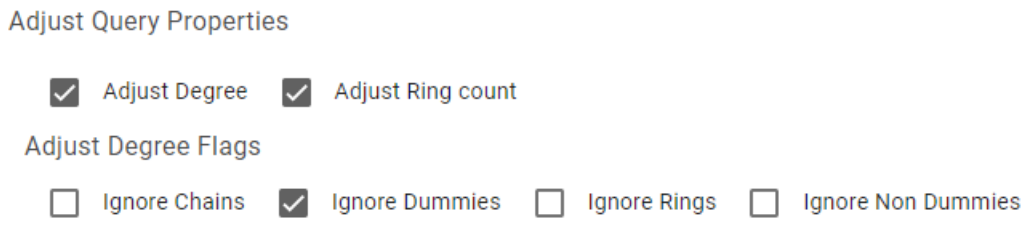


Structure Searching

Four different methods

Search Type	Method	Comment
Substructure	RDKit Cartridge	Configured to best mimic old system
Full Structure	InChI -SNon	Matches any component ignoring stereochemistry
Exact Structure	InChi	Matches any component exactly
Advanced Substructure	RDKit Cartridge	Exposes “AdjustQueryProperties” options





Givaudan

Take-away & Conclusions

- You can use RDKit to build a chemical registration
  - How you do it will be heavily impacted by the specify industry's needs and regulations
- Don't underestimate the complexity and time needed for the data migration
  - Multiple test iterations are needed to work out the kinks! Automate it!
- How to handle stereochemistry, duplicate checking and so forth is not only a technical problem
  - You will have as many opinions as chemists you ask
  - At some point a decision needs to be made and that decision will likely impact the data migration
- Users are very happy
  - Hyperlinks between ChemReg and ELN
  - Can make corrections without need to contact administrators
  - No more copy & paste

Follow us on social media @givaudan

Givaudan

Human by nature