Molecular Structure Representation in Chemoinformatics Applications

Christof H. Schwab

Molecular Networks GmbH
Neumeyerstr. 28
90411 Nürnberg, Germany

mn-am.com

Altamira LLC
1455 Candlewood Drive
Columbus, Ohio 43235, USA

Molecular Networks and Altamira MN-AM

Erlangen, Germany
Friedrich-Alexander-Universität
1997

Columbus, Ohio, USA
The Ohio State University
2008

- Chemoinformatics
  - *3D structure generation*
  - *Physicochemical and reaction properties*
  - *Metabolic reaction knowledge*

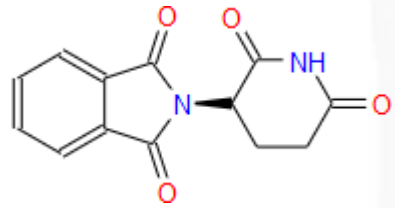
- Computational toxicology and risk assessment
  - *Database and knowledgebase*
  - *Predictive models*
  - *Consulting services*

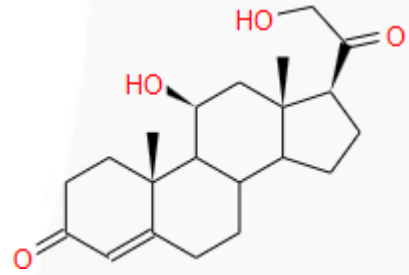
Outline

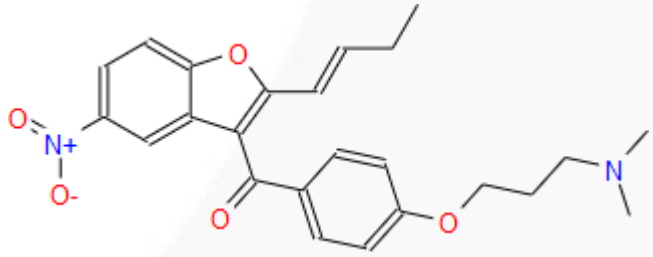
- Representation of chemistry
- Chemoinformatics platform(s)
  - *Some basic concepts*
- Aspects of model buiding
- Examples
  - *Computational toxicology*

Representation of Chemical Structures – General

- Atoms
  - *Atom type (element)*
  - *Charge, radical*
  - *Stereochemistry*
- Bonds
  - *Bond type*
  - *Stereochemistry*
- Electron systems
  - *σ-, π-systems*
  - *Conjugated, aromatic*

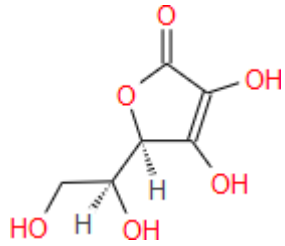






Representation of Chemical Structures – File Formats

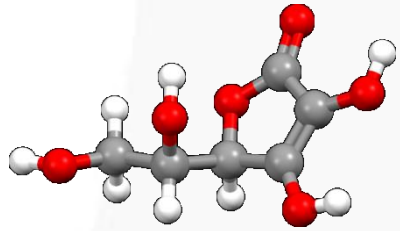
- Linear ("0D") representation
  - *SMILES, InChI, SLN*




```
C1(O)=C(O)C(=O)O[C@@H]1([C@@H](O)(CO)

InChI=1S/C6H8O6/c7-1-2(8)5-3(9)4(10)
6(11)12-5/h2,5,7-10H,1H2/t2-,5+/m0/s1
```

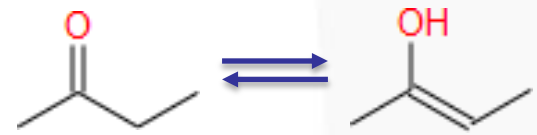
- 2D/3D representation
  - *Molfile, SD V2000/V3000*



- 3D representation
  - *SYBYL MOL/MOL2, MacroModel, Maestro, PDB (Cartesian coordinates)*
  - *CIF, crystallographic file formats (internal coordinates)*

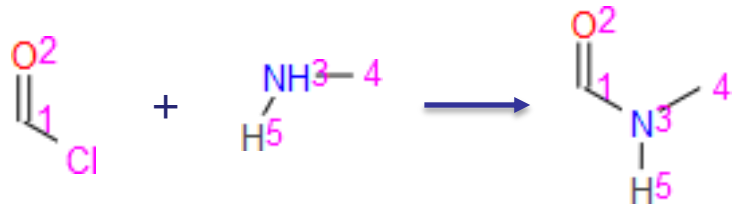
Representation of Chemical Reactions – File Formats

- Linear ("0D") transformations
  - *SMARTS, SMIRKS,…*



CC(=O)CC>>CC(O)=CC

- 2D/3D representations
  - *RD V2000/V3000*




```
$RXN


    Molecular Editor
  2  1
$MOL
[O:2]=[CH:1]Cl.[H:5][NH:3][CH3:4]>>[H:5][N:3]([CH3:4])[CH:1]=[O:2]
JME 2016-11-13 Tue Oct 24 09:02:00 GMT+200 2017

  3  2  0  0  0  0  0  0  0  0999 V2000
    0.0000    0.6984    0.0000 C   0  0  0  0  0  0  0  0  0  1  0  0
    0.0000    2.0951    0.0000 O   0  0  0  0  0  0  0  0  0  2  0  0
    1.2172    0.0000    0.0000 Cl  0  0  0  0  0  0  0  0  0  0  0  0
  1  2  2  0  0  0  0
  1  3  1  0  0  0  0
M  END
$MOL
[O:2]=[CH:1]Cl.[H:5][NH:3][CH3:4]>>[H:5][N:3]([CH3:4])[CH:1]=[O:2]
JME 2016-11-13 Tue Oct 24 09:02:00 GMT+200 2017

  3  2  0  0  0  0  0  0  0  0999 V2000
    0.7005    1.2109    0.0000 N   0  0  0  0  0  0  0  0  0  3  0  0
    2.1016    1.2109    0.0000 C   0  0  0  0  0  0  0  0  0  4  0  0
….
```

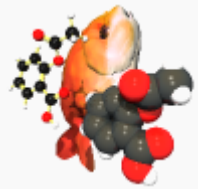
Chemoinformatics Platform – What Is It and What Does It Do?

- Chemistry-aware software tool
- Handling, processing and manipulating chemical information
  - *Chemical structures, reactions and data*
  - *Descriptor calculation*
  - *Reaction generation*
  - *Data mining and analysis*
  - *Database management*
- Visualization
- Interface to the scripting language
  - *Regular tasks and applications*

CDK

RDKit

CORINA/MOSES



OEChem

CACTVS

Many more…, see e.g.,
en.wikipedia.org/wiki/Cheminformatics_toolkits

Chemoinformatics Platform – Requirements

- General
  - *Chemistry awareness for handling a broad range of chemistry and chemical features*
- Robust
  - *Processing of large data sets ("big chemistry/data")*
- Reliable
  - *Deterministic and consistent results*

General – "Big Chemistry"

- Chemistry awareness for handling a broad range of chemistry and chemical features
  - *Different compound classes*
  - *Different areas of application*
  - *Handling of 2D and 3D structures*
  - *Handling of reactions and transformations (e.g., tautomerization)*
  - *Property calculation (e.g., descriptors)*
  - *Storage and retrieval of compounds and reactions*
  - *Datamining capabilities*
  - *Parametrize for entire periodic table*
  - *No restrictions of number of atoms or size of molecules*

Reliability – Consistent Results

- Atom-numbering dependent algorithms
- Unique atom numbering
  - *Canonicalization*

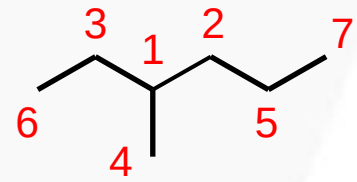
Connectivity

Extended connectivity C_e
$n_{Ce} = 5$

C_e'
$n_{Ce} = 6$

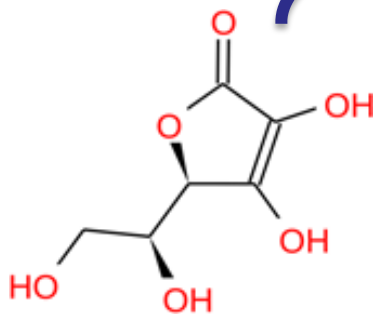
C_e''
$n_{Ce} = 6$

Morgan algorithm (1965)



Reliability – Different Input Formats

- Representation by different file formats



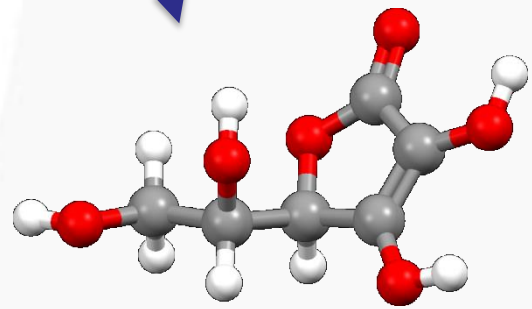

```
Ascorbic acid
                    2D

 12 12  0  0  0  0  0  0  0  0999 V2000
    4.4381    2.4116    0.0000 C   0  0  0
    4.8708    3.7431    0.0000 C   0  0  0
    3.7381    4.5660    0.0000 C   0  0  0
    2.6055    3.7431    0.0000 O   0  0  0
    3.0381    2.4116    0.0000 C   0  0  0
    3.7381    5.9660    0.0000 O   0  0  0
    2.2152    1.2790    0.0000 C   0  0  0
    0.8229    1.4253    0.0000 C   0  0  0
    0.0000    0.2927    0.0000 O   0  0  0
    6.2022    4.1757    0.0000 O   0  0  0
    5.2610    1.2790    0.0000 O   0  0  0
    2.7847    0.0000    0.0000 O   0  0  0
  1  2  2  0  0  0  0
  2  3  1  0  0  0  0
  3  4  1  0  0  0  0
  5  4  1  1  0  0  0
  5  1  1  0  0  0  0
  3  6  2  0  0  0  0
  5  7  1  0  0  0  0
  7  8  1  0  0  0  0
  8  9  1  0  0  0  0
  2 10  1  0  0  0  0
  1 11  1  0  0  0  0
  7 12  1  1  0  0  0
M  END

$$$$
```

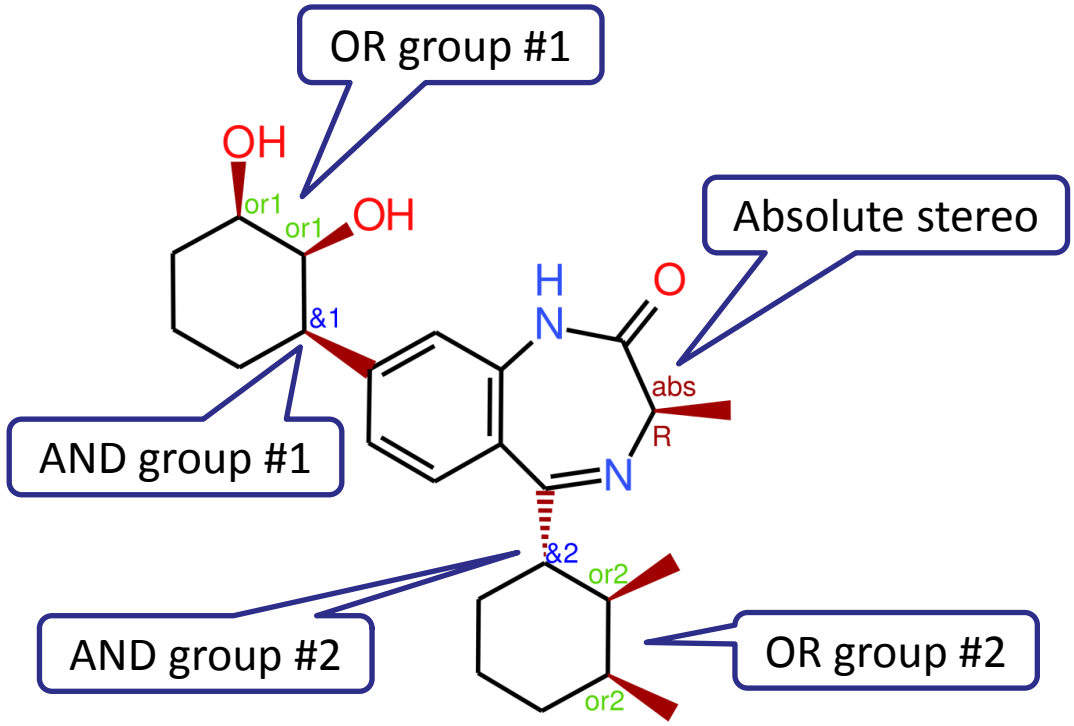
```
InChI=1S/C6H8O6/c7-1-2(8)
5-3(9)4(10)6(11)12-5/h2,
5,7-10H,1H2/t2-,5+/m0/s1
```

```
C1(O)=C(O)C(=O)O[C@@H]1
[C@H](CO)O
```



Reliablity – Handling of Stereochemistry

- Example of stereochemistry coded in SD V3000 format
  - *Total of 7 stereocenters*



Processing options	Number of generated isomers
No restrictions	128 (=2^7)
Preserve all defined isomers	1
Use V3000 stereochemical extension	16 (=2^4)

Robustness – Processing of Large Datasets

- What is big data in chemistry?

- The Cambridge Crystallographic Database (CSD)
  - *Over 875,000 experimentally-determined crystal structures*
- PubChem
  - *Over 90 Million characterized chemical compounds*
- Chemical Abstract Service, CAS Registry
  - *Over 130 Million unique organic and inorganic chemical substances*
- GDB, University of Berne
  - *GDB-13, 1 Billion compounds (13 atoms, C, O, N, S, and Cl)*
  - *GDB-17, 164 Billion compounds (17 atoms, C, O, N, S, and halogens)*

Robustness – Processing of Large Datasets

- 3D structure generation with CORINA Classic

	PubChem	GDB13
Total number of compounds	90,624,841	971,468,301
Total number of converted compounds	90,246,183	961,454,732
Total number of *not* converted compounds	378,658	10,013,569
Conversion rate [%]	99.6	99.0
Computation time [hours]	21	112
Conversion rate [structures per day]	104,000,000	208,000,000

CORINA Classic

www.mn-am.com/online_demos/corina_demo
www.mn-am.com/online_demos/corina_demo_interactive

Modeling Approach

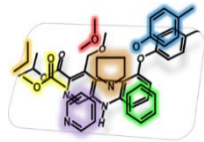
Combination of evidence

Statistical modelling methods

- Partial least squares / logistic regression
- K-nearest neighbors
- Random forest

Descriptors

- Structural features
- Physicochemical properties
- Biological assays

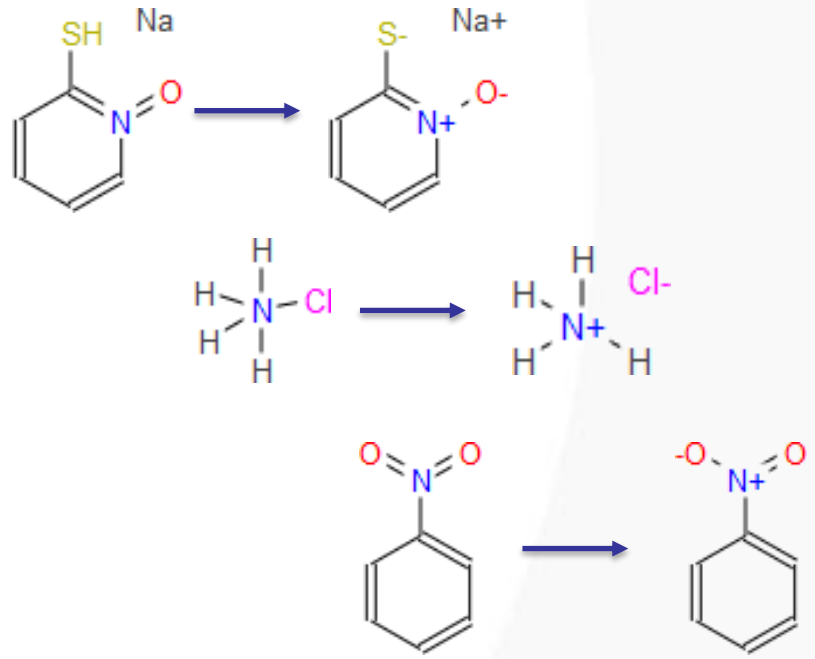


Training set

- Correct structures and endpoint data
- Diversity for wider applicability domain, many large local neighbors for accuracy
- **Mechanistically grouped**

Dataset Preparation – Structure Curation

- Charges and radicals
- Hypervalency
- Bond orders
- Stereochemistry
- Penta-valent nitro
- …

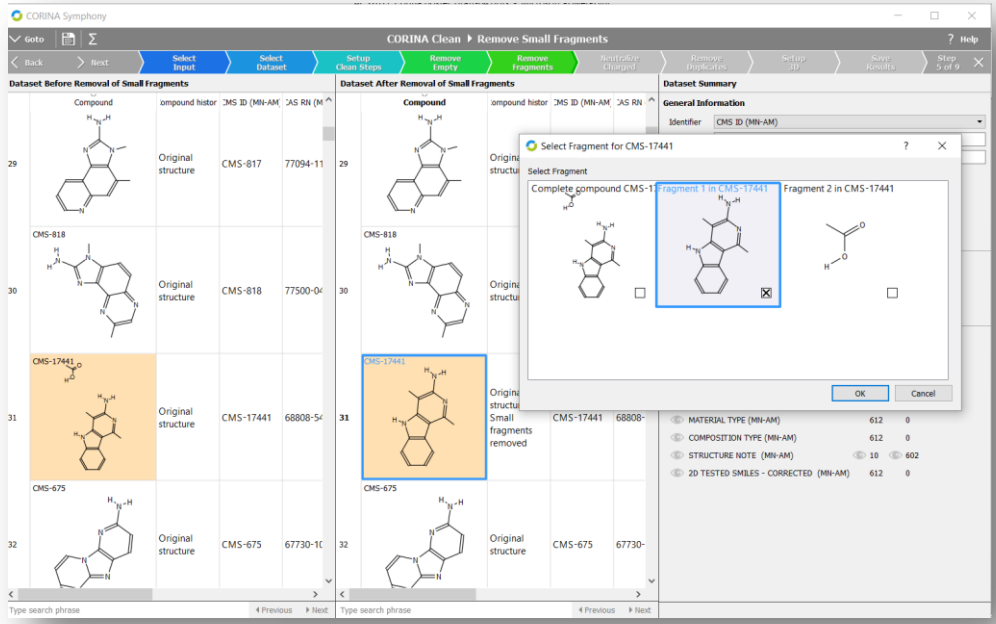


- Check with multiple sources
  - *CAS, PubChem, ChemID Plus, ChemSpider, Wikipedia,…*
- Automatic algorithms, but often manual intervention needed

Dataset Preparation – Structure Clean-Up and Standardization

Select Input | Select Dataset | Setup Clean Steps | Remove Empty | Remove Fragments | Neutralize Charged | Remove Duplicates | Setup 3D | Save Results

- CORINA Symphony
  - *Removal of counter ions and small fragments*
  - *Neutralizing formal charges*
  - *Addition of hydrogen atoms*
  - *Generate 3D coordinates*
  - *Placement in preferred 3D orientation*
  - *Detection and removal of duplicates*

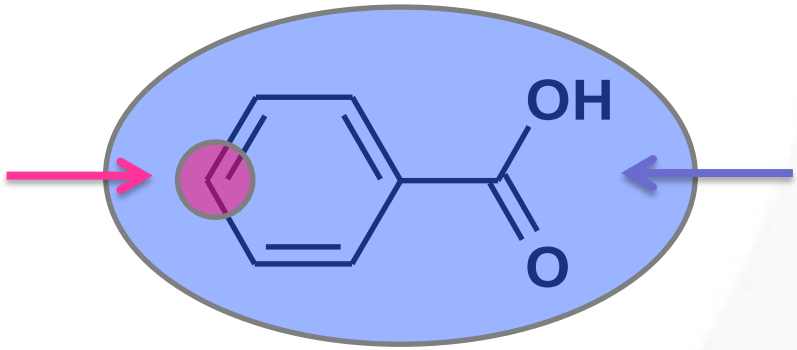


CORINA Symphony

Chemotypes

- Structural fragments with embedded physicochemical properties [1]
  - *Atoms, bonds, electron systems and whole molecule*
- Support by cheminformatics library MOSES
  - *XML-based language to encode chemotypes*

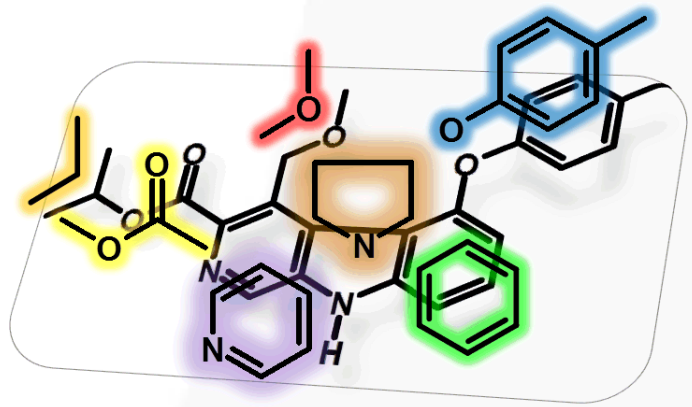
Atom in fragment in certain property range, *e.g.*, partial charge ≤ 0



Molecule that exhibits this fragment in certain property range, *e.g.*, logP > 4

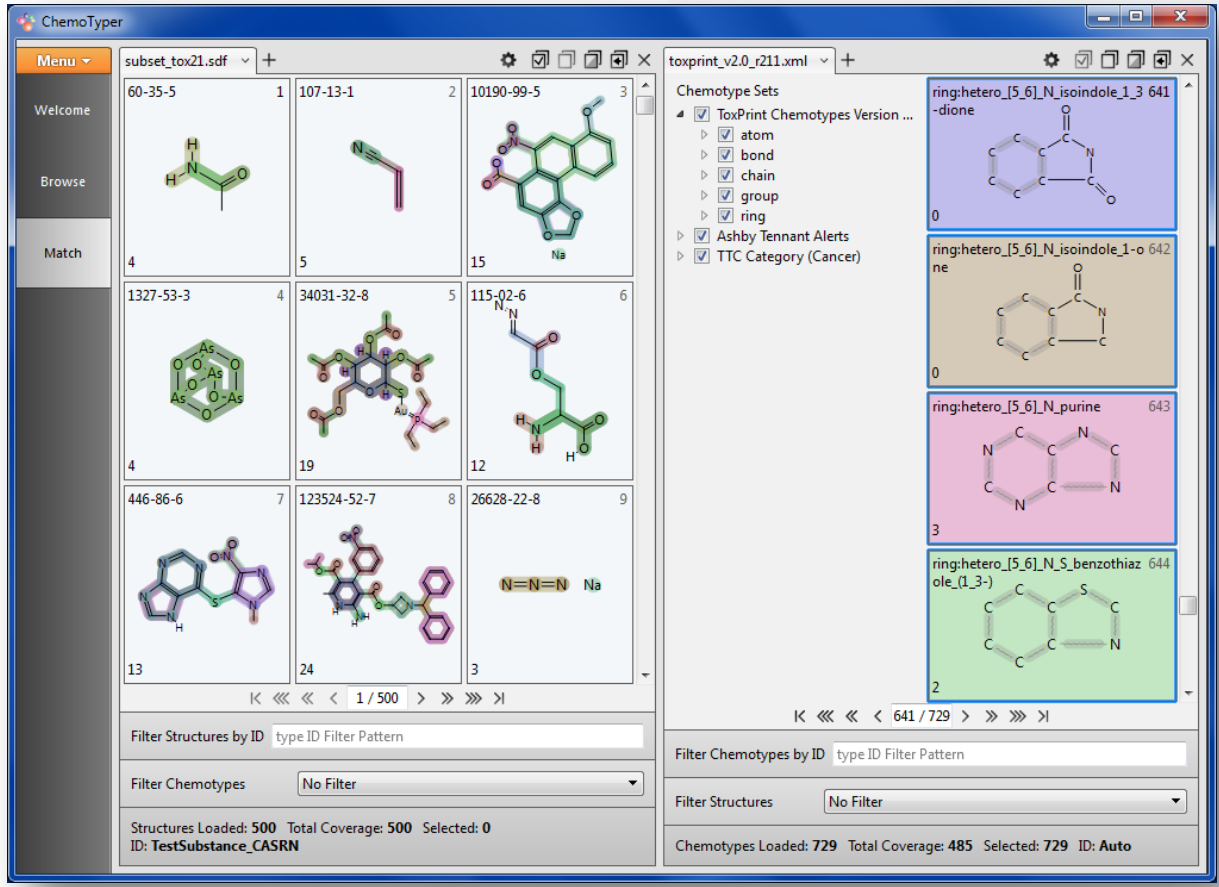
Application of Chemotypes

- Profiling of datasets, inventories, databases
- Endpoint-specific structural alerts
  - *Toxicity*
  - *Reactive toxicity*
  - *Metabolism*
- Risk/safety assessment tool
  - *Building categories*
  - *Read-across*
  - *TTC categories*
  - *Predictive (QSAR) models*



ChemoTyper Application

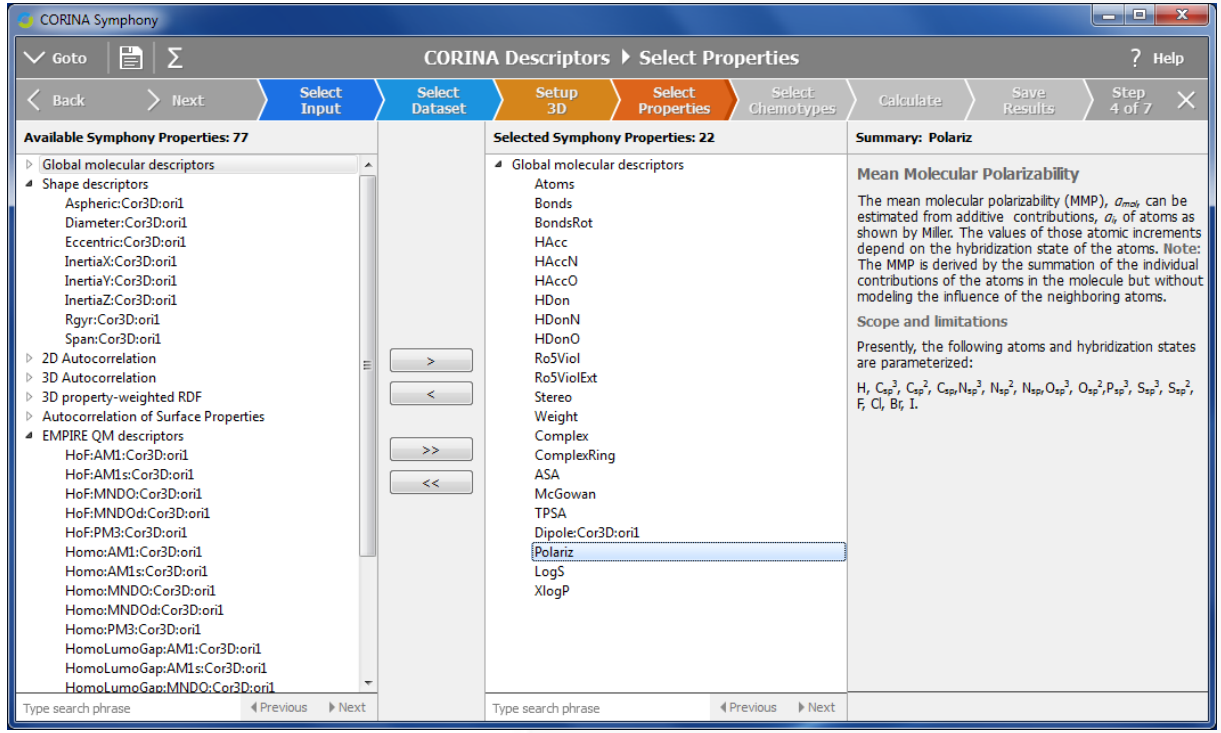
- GUI application
  - *Visual inspection of datasets according to matching rules*
  - *Subset generation*
  - *Grouping of chemicals*
  - *Fingerprinting*
- Contains ToxPrint chemotypes
  - *729 public chemotypes*
  - *Focus on applications in computational toxicity*



chemotyper.org

Descriptors – CORINA Symphony

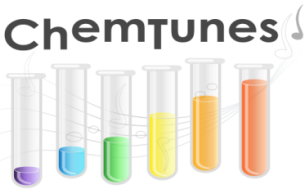
- 2D whole molecule properties
- 3D descriptors
  - *Dipole moment,…*
  - *Shape and size descriptors*
- Quantum mechanical parameters (mechanistic)
  - *HOMO/LUMO energies*
  - ΔH_f

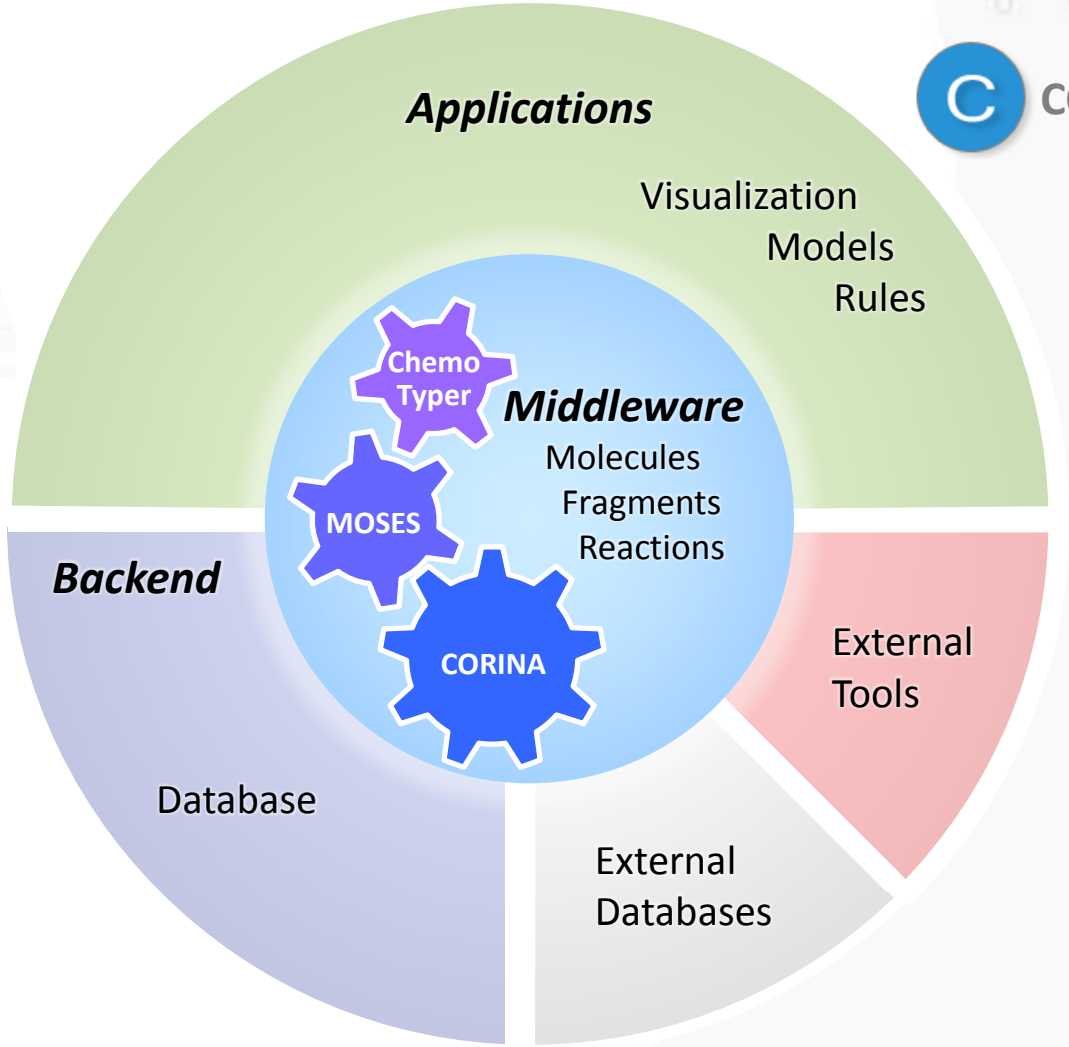


CORINA Symphony

www.mn-am.com/services/corinasymphonydescriptors

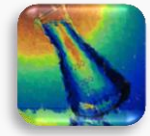
MOSES/CORINA Chemoinformatics Platform at MN-AM



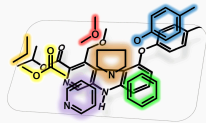


CORINA Classic

CORINA Symphony



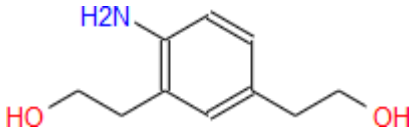
SYLVIA

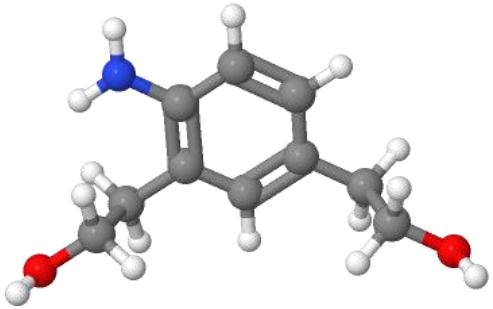


ChemoTyper

Case Study I

- Analyze chemical reactivity of conformations of 2-amino-1,5-benzene-di-ethanol
- Predict skin sensitization potential of conformations

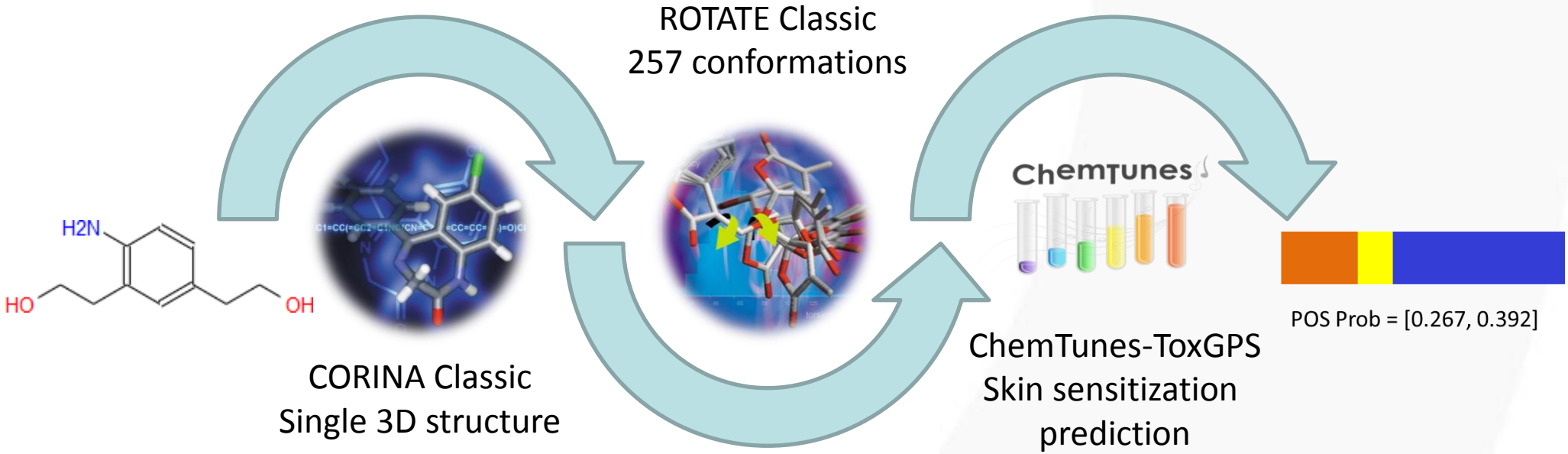




2-Amino-1,5-benzene-di-ethanol

Case Study I – Workflow

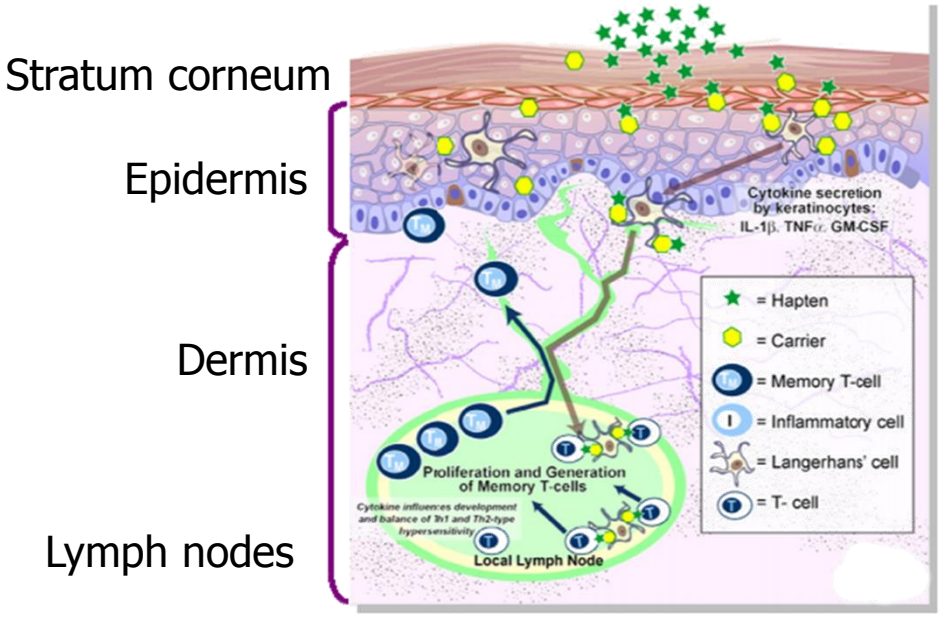
- Analyze chemical reactivity of conformers of 2-amino-1,5-benzene-di-ethanol and predict their skin sensitization potential



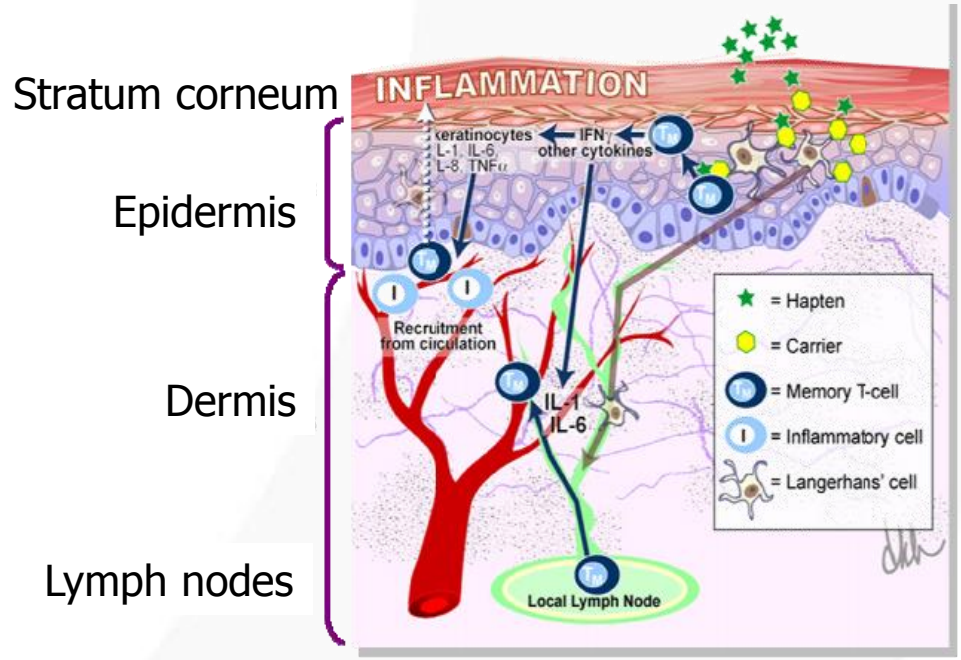
Skin Sensitization – Biology

- Event determining steps in phases of skin sensitization
  - *OECD report ENV/JM/MONO(2012)10/PART1*

Induction

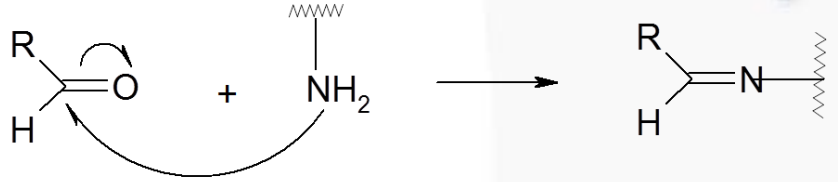


Elicitation

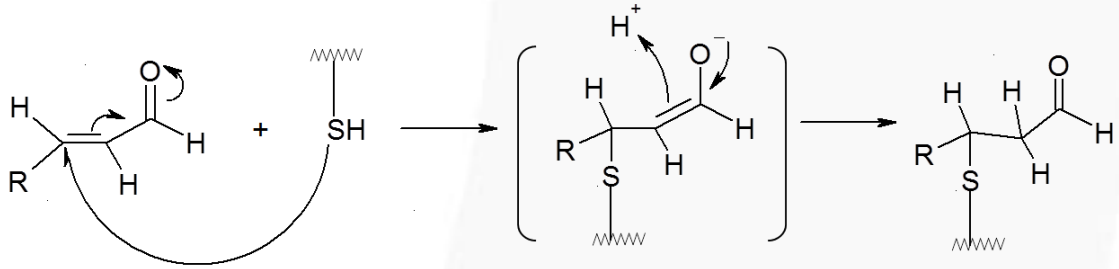


Chemical Reactions in Skin Sensitization – Hapten-Protein

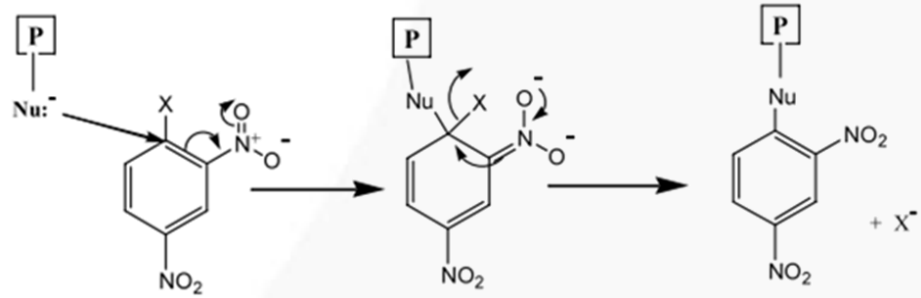
- Schiff base formers
  - *Aldehyde binding to lysine*



- Michael acceptors
  - *α,β-unsaturated carbonyl binding to cystein*

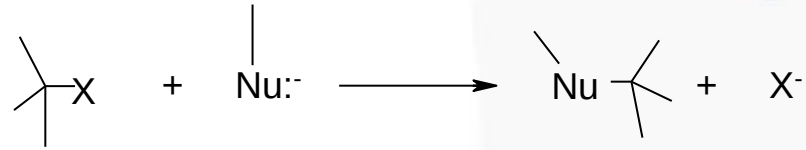


- S_NAr
  - *Aromatic nucleophilic substitution*

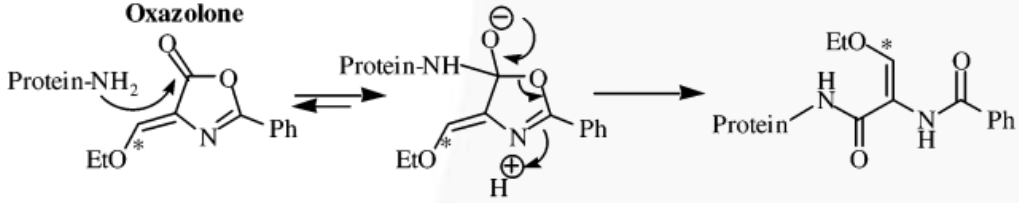


Chemical Reactions in Skin Sensitization – Hapten-Protein

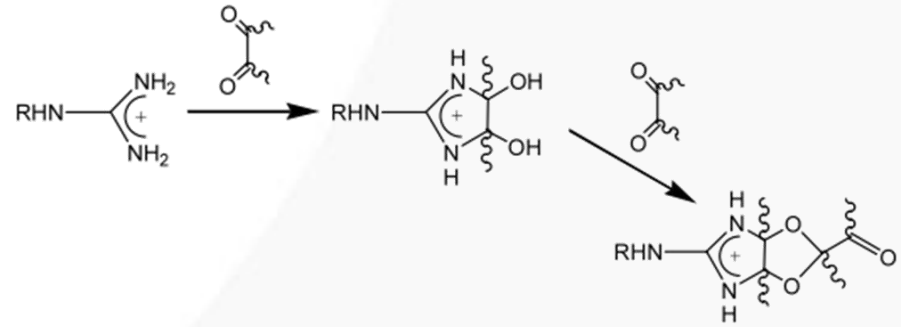
- S_N2
  - *E.g., primary-alkyl halides, etc.*



- Acylating agents
  - *Esters of more acidic alcohols such as phenols and carboxylic anhydrides*



- α,β-Diketones
  - *Binding to arginine forming a heterocyclic adduct*



Chemical Reactions in Skin Sensitization – Hapten-Protein

- Pro-electrophiles

  Alkyl amino amines

- Pro-Michael acceptors
  - *Oxidation*

  Phenol (dihydroxy) → quinone →
  Aromatic amines (diamino) → di-imine →

- Metabolic activation
  - *Oxidation*

  E.g., phenols (eugenol)

- Literature [2], [3]

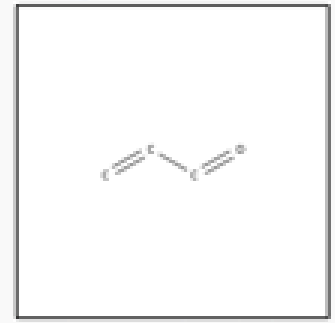
Skin Sensitization – Modeling

- Dataset
  - *Local lymph node assay*
    - Hazard model: 602 unique structures, 390 sensitizers / 212 non-sensitizers
    - Potency model: 390 sensitizers, 282 GHS 1B / 108 GHS 1A

- Descriptors
  - *CORINA Symphony descriptors*
    - Global molecular, size and shape descriptors
  - *Semi-empirical quantum mechanical descriptors calculated by EMPIRE*
    - ΔH_f, HOMO, LUMO, HOMO/LUMO gap
  - *Toxprint chemotypes*

Skin Sensitization – Modeling

- Structural alerts
  - *Chemotype with positive correlation to endpoint*
  - *Trained for respective endpoint*
  - *Odds ratio as quantitative severity score*



- Mechanistically grouping according to mode of action (MoA)
  - *Global and local models*
    - Alcohols
    - Aldehyde and ketones
    - Amines
    - Michael acceptors
    - Phenols and quinones

Skin Sensitization – Modeling

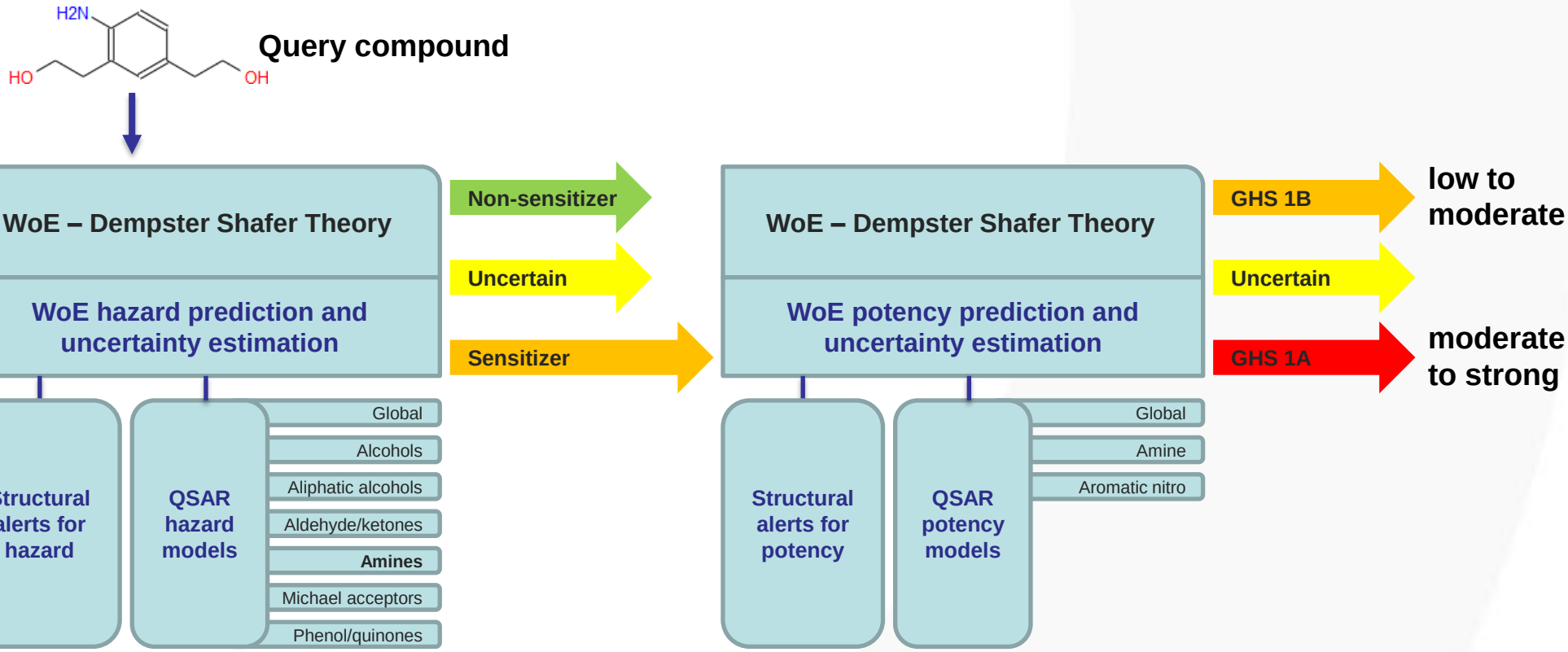
- Modeling technique
  - *Bivariate analysis for descriptor selection*
  - *Hybrid method of partial least squares (PLS) and ordinal logistic regression*
    - Composite set of orthogonal latent variables

- Weight of evidence approach
  - *Dempster-Shafer theory (decision theory)*
    - Probabilistic approach providing quantitative estimation of uncertainties of predictions
  - *Overall outcome combines QSAR models and structural alerts*

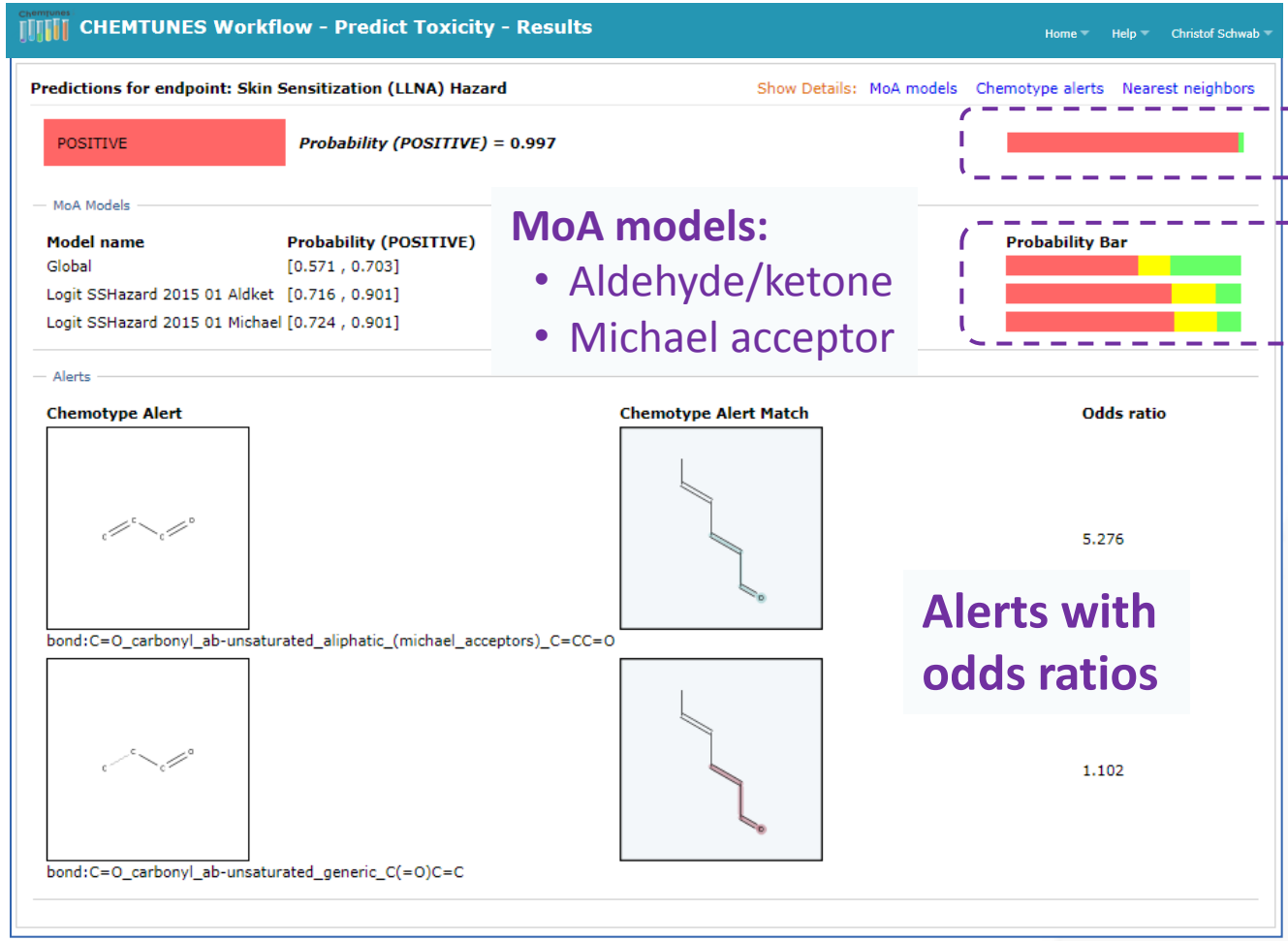
POS Prob = [0.267, 0.392]

ToxGPS Skin Sensitization Model

- Hazard and potency model



Skin Sensitization – (LLNA) Hazard Prediction



ChemTunes ID: CMS-4336
Name: Trans,Trans-2,4-Hexadienal
Registry number(s): 142-83-6; 4488-48-6
studies in ChemTunes: 26

ToxGPS Skin Sensitization Model

- External validation of skin sensitization hazard model (138 compounds)

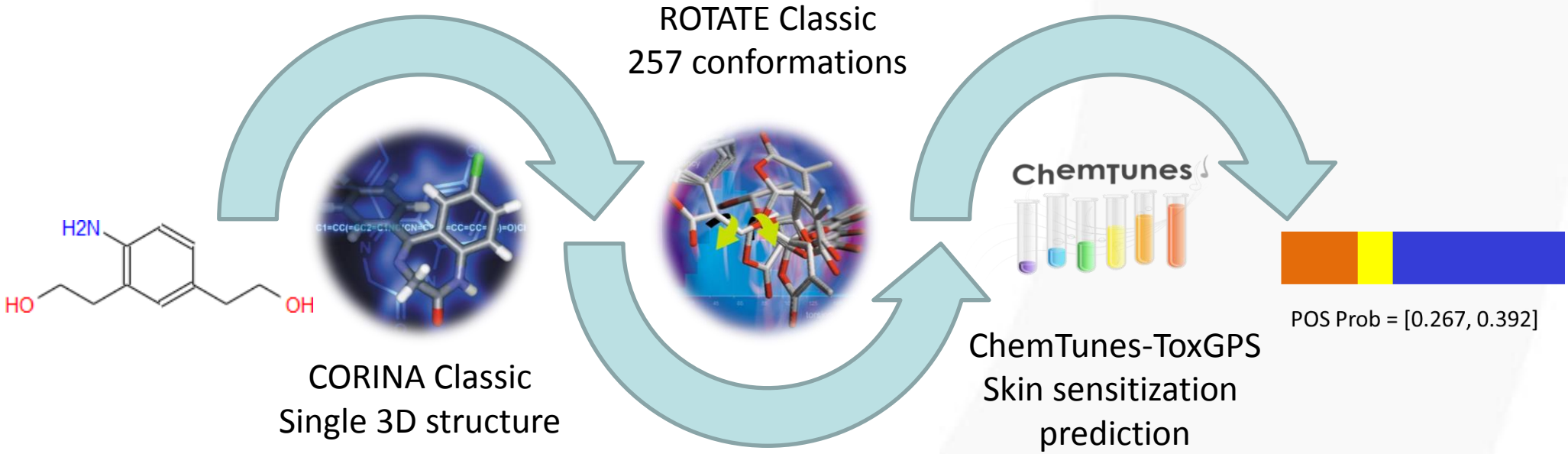
Statistical parameter	QSAR models	QSAR models and structural alerts
Sensitivity (skin sensitizers)	87%	88%
Specificity (non-sensitizers)	82%	82%
Equivocal predictions	5%	5%

- External validation of skin sensitization potency model (99 compounds)

Statistical parameter	QSAR models	QSAR models and structural alerts
Sensitivity (GHS 1A)	83%	85%
Specificity (GHS 1B)	83%	81%
Equivocal predictions	12%	10%

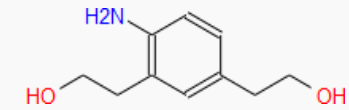
Case Study I – Workflow

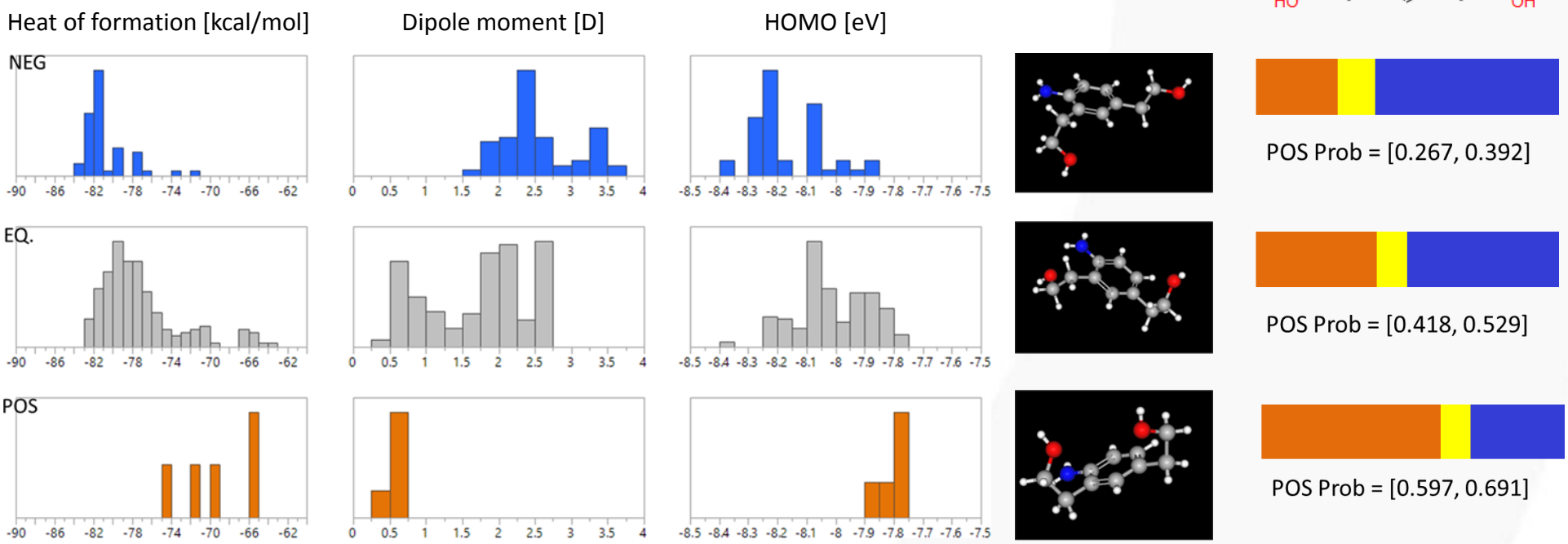
- Analyze chemical reactivity of conformers of 2-amino-1,5-benzene-di-ethanol and predict their skin sensitization potential



ToxGPS Skin Sensitization Model

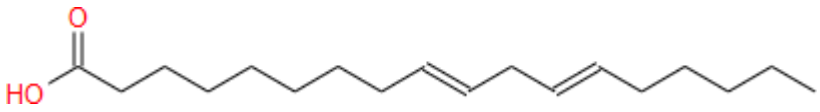
- Local amines model for skin sensitization hazard QSAR model
  - *Distribution of selected molecular descriptors across conformers predicted as negative, equivocal, and positive*





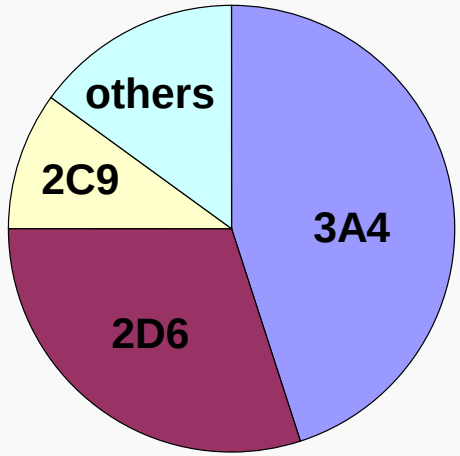
Case Study II

- Analyze conformers of linoleic acid and predict predominant isoform of human P450



isoCYP Model

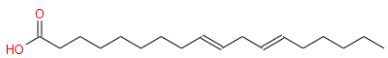
- Prediction of the predominant human P450 isoform
  - *3A4, 2D6, 2C9*
- Classifier based on decision tree (isoCYP)
  - *Dataset of about 380 compounds with known predominant isoform*
  - *Five 2D- and 3D-based descriptors*

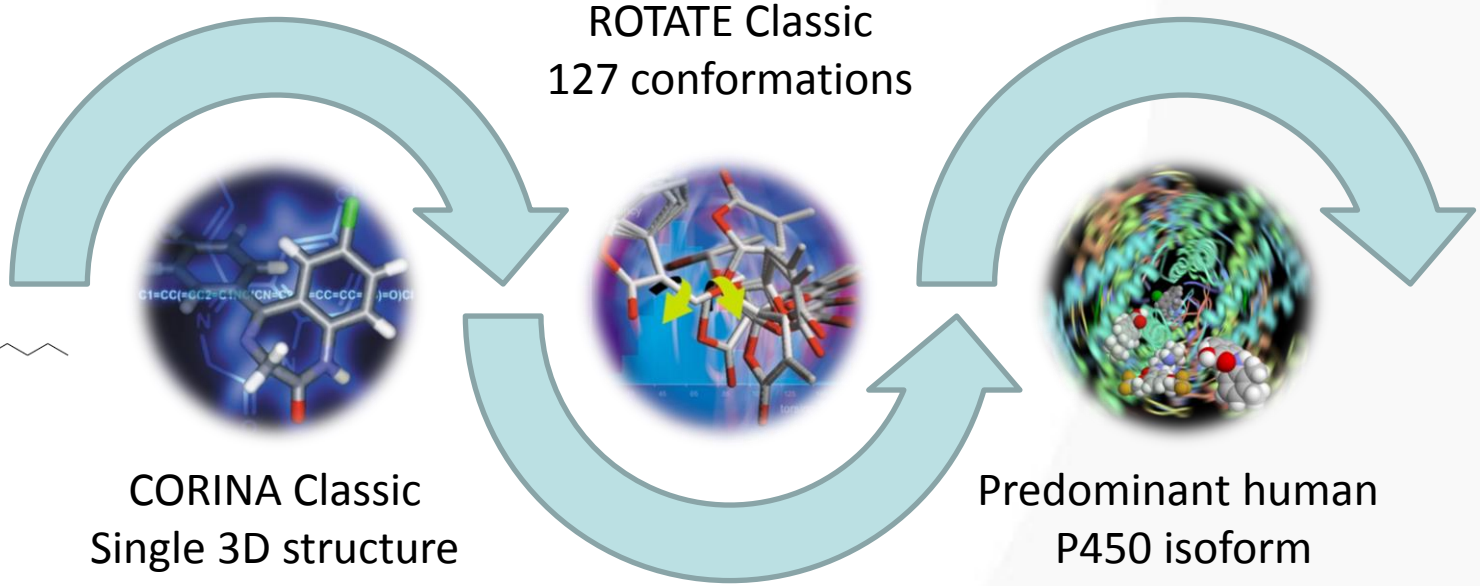


Drug metabolism

Case Study II – Workflow

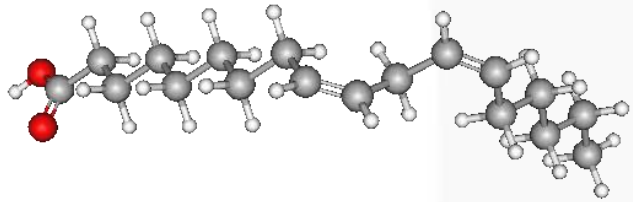
- Analyze conformers of linoleic acid and predict predominant isoforms of human P450





Case Study II

- Linoleic acid
  - *14 rotatable bonds*

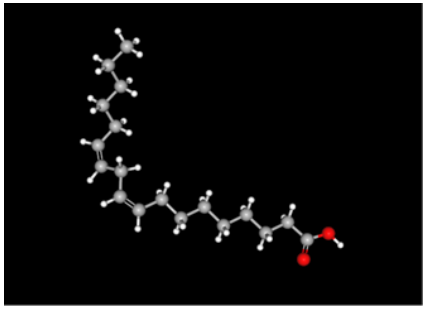


- ROTATE Classic parameters
  - *Process all rotatable bonds*
  - *Use the 2 most preferred torsion angle values for each rotatable bond*
  - *Keep only conformers having at least an RMS deviation of 15 degree in torsion angle space to each other*
  - *Save conformers with ROTATE energy score of up to 30*

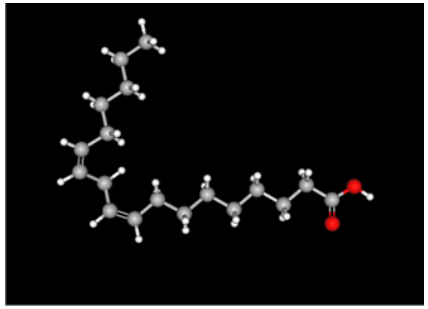
- 127 conformations

Case Study II

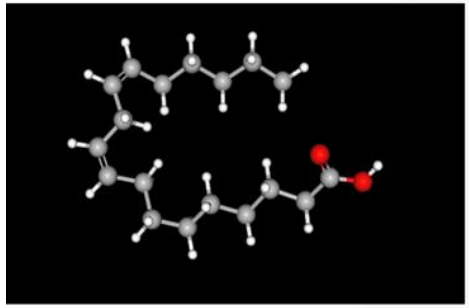
- Different conformation predicted for different predominant isoforms
  - *Different mechanisms reported in literature*



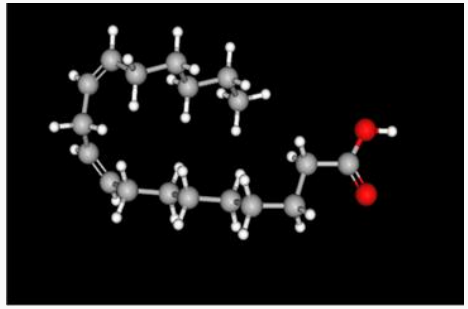
ROTATE energy score: 1.7
AM1 HoF: -524.4 kJ/mol



ROTATE energy score: 5.3
AM1 HoF: -503.4 kJ/mol



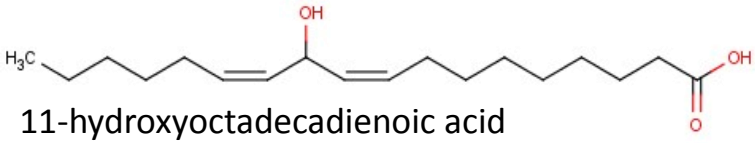
ROTATE energy score: 27.9
AM1 HoF: -470.7 kJ/mol



ROTATE energy score: 29.4
AM1 HoF: -459.9 kJ/mol

CYP 3A4

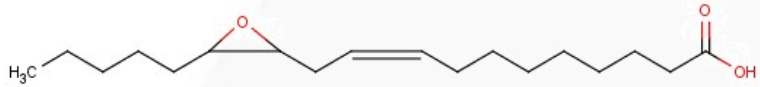
Conversion to bisallylic hydroxy fatty acid [4,5]



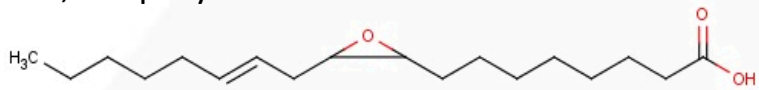
11-hydroxyoctadecadienoic acid

CYP 2C9

Linoleate epoxygenase in human liver microsomes [5,6]



12,13-epoxy-9-octadecenoate



9,10-epoxy-12-octadecenoate

Summary

- Basic concepts of cheminformatics platforms
  - *General, robust and reliable*
  - *MOSES/CORINA at MN-AM*
- Different steps in model building
  - *Data*
  - *Mechanistic knowledge*
  - *Descriptor selection*
  - *Uncertainties in predictions*
- Influence of 3D structure/conformation

Acknowledgements

- BigChem project
  - *Nadja Saendig, Guilio Rastelli*
  - *Barbara Gasset, Igor Tetko*

- My colleagues at MN-AM
  - *Aleksandra Mostrag-Szlichtyng, Chihae Yang, Jie Liu, Vessela Vitcheva, Aleksey Tarkhov, Bruno Bienfait, James F. Rathman, Jörg Marucszyk, Jongjin Park, Michael Mulcahy, Oliver Sacher, Thomas Kleinöder, Tomasz Magdziarz*

- Thank you for your attention!

Literature

[1] Yang C *et al.* 2015. J. Chem. Inf. Model. 55(3): 510-528 DOI: 10.1021/ci500667v

[2] Smith Pease, C.K. 2003. Toxicology, 192: 1-22

[3] Aptula, *et al.* 2006. Chem. Res. Toxicol. 19: 1097

[4] Bylund J *et al.* 1998. Analytical Biochemistry 265: 55-68

[5] Bylund J *et al.* 1998. Journal of Pharmacology and Experimental Therapeutics 284(1): 51-60

[6] Draper AJ, Hammock BD. 2000. Archives of Biochemistry and Biophysics 376(1): 199-205